



Linear Models and Řurità Kràlovství

Using the Kingdom for Greater Insight

Version 0.704442η

LINEAR MODELS AND ŘURITÀ KRÀLOVSTVÌ

USING THE KINGDOM FOR GREATER INSIGHT

Version: $0.704442\eta(\alpha)$

OLE J. FORSBERG

*Department of Mathematics
Knox College
Galesburg, IL, USA*

December 2024

ALL RIGHTS RESERVED. No part of this work covered by copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including, but not limited to photocopying, recording, scanning, digitizing, taping, Internet distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 US Copyright Act, without the prior written permission of the author.

The current draft version of this document is free (without cost). This document is distributed in the hope that it will be useful, but without any warranty; without even the implied warranty of merchantability or fitness for a particular use or for a particular purpose.

This document was typeset using the $\text{\LaTeX} 2_{\epsilon}$ environment. All graphics were created by the author using the \R statistical environment (and referenced packages). All data is either public domain or specifically generated for this text. The following are the photograph credits.

Page	Site	Source/Holder	CC Type
Cover	Own AI Artwork	Ole J. Forsberg	CC0
1	Boomer Lake in Stillwater, OK, USA	Ole J. Forsberg	BY-NC-SA
2	King Harald V of Norway	Government of Norway	CC0
6	Ambassador Mark Brzezinski	US Embassy to Poland	CC0
12	King Olav V of Norway	Government of Norway	CC0
15	Street View, Karlstad, Sweden	Ole J. Forsberg	BY-NC-SA
43	A Steadfast Tin Soldier statue, Odense, Denmark	Ole J. Forsberg	BY-NC-SA
75	A bridge in Karlstad, Sweden	Ole J. Forsberg	BY-NC-SA
105	A canal in The Hague, Netherlands	Ole J. Forsberg	BY-NC-SA
151	The Peace Palace in The Hague, Netherlands	Ole J. Forsberg	BY-NC-SA
189	Sunne Church in Sunne, Sweden	Ole J. Forsberg	BY-NC-SA
227	The Vigeland Park in Oslo, Norway	Ole J. Forsberg	BY-NC-SA
243	A map of San Marino	Wikipedia User Aotearoa	BY-SA
261	Bergen, Norway	Ole J. Forsberg	BY-NC-SA
281	Bergenshus Fortress in Bergen, Norway	Ole J. Forsberg	BY-NC-SA
307	St. Canute's Cathedral, Odense, Denmark	Ole J. Forsberg	BY-NC-SA
331	Street scene in Malmö, Sweden	Ole J. Forsberg	BY-NC-SA
375	The North Shore of Våtsjön, near Villingsberg, Sweden	Ole J. Forsberg	BY-NC-SA
407	Windmill scene in Malmö, Sweden	Ole J. Forsberg	BY-NC-SA
439	Fall colors near Boomer Lake in Stillwater, OK	Ole J. Forsberg	BY-NC-SA
475	Street scene in Karlstad, Sweden	Ole J. Forsberg	BY-NC-SA
507	A Cherry Blossom Tree in Galesburg, IL, USA	Ole J. Forsberg	BY-NC-SA
533	Royal Guard in Oslo, Norway	Ole J. Forsberg	BY-NC-SA
571	William Sealy Gosset, 1908	Wikipedia	CC0
571	Ronald Fisher, 1913	Wikipedia	CC0
573	Augustin-Louis Cauchy, 1901	Wikipedia	CC0

Contents

List of Figures	xi
List of Tables	xv
Preface	xvii
1 An Introduction to Ruritania	1
1.1 Background of Ruritania	3
1.2 Economics	5
1.3 US–Ruritanian Relations	6
1.4 Illustrations of Analyses	7
1.5 Conclusion	12
I Least Squares	13
2 Introduction to Linear Regression	15
2.1 Scalar Representation	20
2.2 Results	31
2.3 First Assumptions	34
2.4 The PRE Measures	38
2.5 Conclusion	41
2.6 End-of-Chapter Materials	42
3 Matrices and Linear Regression	43
3.1 Matrix Representation	46
3.2 Predictions and the Hat Matrix	57
3.3 The PRE Measures	63
3.4 Multicollinearity and Categorical Independent Variables	64
3.5 Conclusion	72

3.6	End-of-Chapter Materials	73
4	Improved! Now with Probabilities	75
4.1	Probability Distributions	79
4.2	Test Statistics and Hypothesis Testing	91
4.3	Confidence Intervals	94
4.4	The Working-Hotelling Bands	100
4.5	Conclusion	101
4.6	End-of-Chapter Materials	102
5	Dood! Check the Requirements	105
5.1	Normality	108
5.2	Constant Expected Value	122
5.3	Constant Variance	129
5.4	Multicollinearity	138
5.5	Conclusion	145
5.6	End-of-Chapter Materials	146
6	A Time for Some Examples	151
6.1	Full Example: Violent Crime	153
6.2	Full Example: Violent Crime, Wealth, Region	158
6.3	Full Example: Cows in the City of Děčín	166
6.4	Conclusion	184
6.5	End-of-Chapter Materials	185
7	Fixing the Violations	189
7.1	The Issue of Boundedness	192
7.2	Full Example: The South Sudanese Referendum	209
7.3	Heteroskedastic Adjustments	216
7.4	Conclusion	218
7.5	End-of-Chapter Materials	219
II	Beyond the Ordinary	225
8	Other Least Squares	227
8.1	Ordinary Least Squares	229
8.2	Weighted Least Squares	230
8.3	Generalized Least Squares	240
8.4	Full Example: May the (Strong) Force be with You	246
8.5	Full Example: Elections in Ruritania	250

8.6	Conclusion	255
8.7	End-of-Chapter Materials	256
9	Quantile Regression	261
9.1	Parameter Estimation	263
9.2	Quantile Regression	269
9.3	Conclusion	275
9.4	End-of-Chapter Materials	276
10	Maximizing the Likelihood	281
10.1	The Likelihood	283
10.2	The MLE and the CLM	294
10.3	Conclusion	302
10.4	End-of-Chapter Materials	303
III	Beyond the Classical Model	305
11	Generalized Linear Models	307
11.1	The CLM and the GLM	309
11.2	The Requirements for GLMs	310
11.3	Assumptions of GLMs.	315
11.4	The Gaussian Distribution	316
11.5	Generalized Linear Models in \mathbb{R}	318
11.6	Conclusion	325
11.7	End-of-Chapter Materials	326
12	Binary Dependent Variables	331
12.1	Binary Dependent Variables	333
12.2	Latent Variable Modeling	336
12.3	The Mathematics	339
12.4	Modeling with the Logit	347
12.5	Prediction Accuracy	351
12.6	Modeling with Other Links.	359
12.7	Model Selection	362
12.8	Conclusion	368
12.9	End-of-Chapter Materials	369
13	Binomial Dependent Variables	375
13.1	Binomial Distribution	377
13.2	The Mathematics.	379

13.3	Full Example: O Canada!	381
13.4	Full Example: Sri Lanka in 2010.	392
13.5	Beta-Binomial Regression*	397
13.6	Conclusion	400
13.7	End-of-Chapter Materials	401
14	Count Dependent Variables	407
14.1	Linear or Poisson Regression?	410
14.2	The Mathematics	413
14.3	Overdispersion	420
14.4	Full Example: Body Counts	426
14.5	The Bias-Variance Trade-Off	433
14.6	Conclusion	434
14.7	End of Chapter Materials	435
15	Nominal and Ordinal Dependent Variables	439
15.1	Nominal Dependent Variable	442
15.2	Ordinal Dependent Variable	454
15.3	Extended Example: Cattle Feed	458
15.4	Extended Example: The State University of Ruritania	463
15.5	Conclusion	467
15.6	End-of-Chapter Materials	468
IV	The Appendices	473
	Appendix M The Appendix of Matrices	475
M.1	Matrix Basics	476
M.2	Addition	479
M.3	Multiplication	481
M.4	Other Matrix Terms and Operations	493
M.5	Consequences	499
M.6	Statistics in Matrices	502
M.7	End-of-Appendix Materials.	506
	Appendix R Experimenting with R	507
R.1	Installing R	509
R.2	R Packages	510
R.3	R Functions	512
R.4	Programming Practice.	517

Appendix S	The Appendix of Statistics	533
S.1	Importantly Confusing Points.	535
S.2	Sample Statistics	537
S.3	Population Parameters	545
S.4	Probability Distributions	556
S.5	Distributions of Sample Statistics	576
S.6	Other Topics	580
S.7	End-of-Appendix Materials.	603
Index		607

List of Figures

1.1	His Majesty Rudolph II, King of Ruritania.	2
1.2	The Flag of Ruritania, 1954	3
1.3	Economic Output	5
1.4	Mark Brzezinski	6
1.5	Rudolph II (1952)	12
2.1	Sample data	17
2.2	Illustrating a line of best fit	21
2.3	Example an OLS line of best fit	27
3.1	Sample data	45
3.2	Comparison of \mathbf{Y} and $\hat{\mathbf{Y}}$ spaces	59
4.1	Illustrating residuals	77
4.2	Comparing confidence interval and prediction interval widths	88
4.3	A visual of a confidence interval	94
4.4	Central (or symmetric) confidence interval	96
4.5	Minimum-width confidence interval	97
5.1	Example quantile-quantile plot	109
5.2	Default residuals histograms	110
5.3	Enhanced residuals histograms	111
5.4	Residuals plot of misspecified model	123
5.5	Residuals plot of properly-specified model	124
5.6	Illustration of homo- and heteroskedasticity	130
5.7	Diagram of multicollinearity	143
6.1	Plot of the violent crime rate in 2000 against that in 1990	156
6.2	Plot of the 2000 violent crime rate against the 1990 GSP per capita	164
6.3	Correlation plots between the three independent variables.	169
6.4	Prediction graphs of our <code>cows</code> model.	180

6.5	Plot of the predicted vote outcomes from the MC experiment. . .	183
7.1	Schematic of the variable transformation procedure.	193
7.2	Graphic of the logistic function. The logit function is the inverse of the logistic. Note that the graph is symmetric about the point (0,0.5).	194
7.3	Schematic of the transformation procedure for Example 7.1.1. . .	197
7.4	Histogram of the MC experiment for Děčín.	198
7.5	Histogram of the MC experiment for Ruritania.	204
7.6	A scatterplot of the results of the South Sudan referendum. . . .	210
7.7	The results of the South Sudan referendum with predictions. . .	214
8.1	A map of Ruritania	243
8.2	Invalidation plot for the Parliamentary election	253
9.1	Comparing OLS and medreg estimators	268
9.2	Quantile effects of crime on crime	271
9.3	Quantile effects of wealth on crime	273
10.1	Graph of MLE of a Binomial	286
10.2	Graph of MLE of a Poisson	287
10.3	Graph of MLE of an Exponential	293
11.1	Graphs comparing CLM and GLM transformed results.	321
11.2	Comparison of CLM and GLM predictions.	324
12.1	Residual scatterplot for example 12.1.	334
12.2	Schematic of logistic regression.	338
12.3	Three symmetric link functions.	345
12.4	Two asymmetric link functions.	346
12.5	Plot of predicted coin weightings.	350
12.6	The model accuracy against various thresholds.	354
12.7	The ROC curve for the coin flipping model.	356
12.8	The ROC curve using the <code>ROC</code> command.	357
12.9	Plot of logit and complementary log-log functions.	359
12.10	Plot of logit and log-log functions.	362
13.1	A map of the world and Canada	381
13.2	Reasonable dispersions for O Canada	384
13.3	The 'O Canada' data and estimates	390
13.4	A map of the world and Sri Lanka	392

13.5 Sri Lankan Presidential returns from 2010, without postal	395
13.6 Sri Lankan Presidential returns from 2010	396
13.7 Beta-Binomial regression figure	399
14.1 Poisson data plot with regression curves.	411
14.2 Plot of initiative use against state population.	419
14.3 Plot of the number of deaths due to terrorism.	432
15.1 Explanatory schematic for thresholding.	455
15.2 Feed type probabilities for RUR ranch.	462
15.3 Class level probabilities.	466
M.1 A schematic of commensurate matrices for multiplication	482
R.1 The probability density function (pdf) of a Cauchy(2,1.5) distribution.	518
R.2 The estimated probability density function (pdf) of the volume of a cylinder.	519
R.3 The estimated probability mass function (pmf) of the number of customers arriving in eight hours. Note that it closely follows the $\mathcal{P}(\lambda = 24)$ distribution. Using the techniques of probability theory, one can prove this relationship.	524
R.4 The estimated probability density function (pdf) of the lifetime of a flashlight.	526
R.5 The estimated probability density function (pdf) of the sample means from the unknown population. The mean of the sample means is denoted by the triangle on the axis; the 95% confidence intervals, thick bar on the axis.	529
S.1 William Sealy Gosset, 1908	571
S.2 Ronald Fisher, 1913	571
S.3 Augustin-Louis Cauchy, 1901	573
S.4 Abraham Wald	587
S.5 Distribution of the number of runs	589

List of Tables

1	The Greek alphabet	xxii
3.1	Raw data for Example 3.4.	64
3.2	Structured data for Example 3.4.	65
4.1	P-value calculations	93
6.1	Statistics on the <code>cows</code> data set.	167
6.2	Correlations between the variables in the <code>cows</code> data	168
6.3	The grammar of formulas	171
6.4	Results table for the <code>cows</code> example.	177
7.1	Results table for the <code>cows</code> vote model (logit).	196
7.2	Results table for the GDP per capita model.	201
7.3	Results table for the South Sudan referendum model.	211
7.4	Regression table using White standard errors	217
8.1	Data for the Strong Force example	246
9.1	Example of median regression	264
11.1	GLM distributions and links.	311
11.2	Results table for the <code>cows</code> example.	320
11.3	Results table for the <code>cows</code> vote model (logit).	321
11.4	Results table from fitting the GDP data with GLMs.	322
12.1	Insurance data to accompany Example 12.1.	334
12.2	Binary dependent variable link functions.	343
12.3	Results table for logistic regression on the coin flip data.	349
12.4	Results table for clog-log regression on the coin flip data.	361
14.1	Initiative model, with adjusted standard errors.	421

14.2 Initiative model, fitted using MQLE.	422
14.3 Results table for initiatives using the Negative Binomial.	424
14.4 Three terrorism models, days as independent variable.	428
14.5 Three terrorism models, using days as offset.	429
14.6 Two terrorism models with higher degrees.	431
15.1 Correlation matrix.	448
15.2 Results table for the logit model.	449
15.3 Results table for the multinomial regression model.	451
15.4 Result of ordinal regression in \mathbb{R}	456

Preface

I sat there, as I was wont to do, in the Spillane Reading Room, drinking coffee in the early morning hour, trying to find my wakefulness. The air was peaceful, with two first-years discussing and sharing insights about international politics and world events with each other and with me. It was life as I expected it to be in academia.

Frequently, more so as the term wore on, “Simon” would rush into the room and explode, “I don’t know what this *thing* is telling me!” Then, he would throw down five pages of printout from a well-known statistical package and throw up his hands, as if beseeching the gods of statistics to send down an answer to him.

After allowing the situation to calm, and the first-years to start breathing again, I would ask the question “What did you do to get the printout?” (Show your work.) For some reason, I always expected the answer to differ from all the times previous; I expected him to tell me what he intended to do, what specific actions he performed, what information those commands were supposed to give him, and why he needed that information.

“I clicked on some menu things and this came out.” As expected.

He and I would then sit down and go over the fifteen pages of printout, examining what each of the tables and numbers meant in relation to his research. Eventually, after dealing with several “Why is it giving me *that* information?!” questions, Simon would be vaguely satisfied with the printout and could select several statistics from the printout that would provide the information he sought.

However, there were many more questions I wanted to ask him. Most centered on questions about the validity of the tests performed. I knew, however, that such a line of questioning would be moot with the statistical package he (and his class) used. Either the company that owned the package had the test available, or it did not. There was no (easy) way to add tests and procedures. Thus, Bartlett’s test of equal variances was not an option, even when the data was such that it should be analyzed using it. Furthermore, the number of available tests was quite limited.

In addition to the extensibility issue, there was the issue of clicking your way to an analysis. If Simon needed to repeat the exact analysis, except

for a tweak or two, he would have to start from scratch and repeat it all, clicking all the same menu items, hoping that he did not make a mistake along the way. This repeat analysis happens quite frequently in life.



One of the many pleasures I have as a statistician is my exposure to many different disciplines. I consulted for a woman doing research in science education. Her specific problem was to determine if a specific Science Teaching Unit (STU) had the students like science more. She performed her experiment on a class of fifth- and sixth-graders in rural North Dakota (hardly representative). She gave them a 50-question pre-test, taught the Unit, then gave the same students a 50-question post-test (what about reliability?!?!).

She then contacted me and sent the data.

On the data she sent to me, I spent about three hours analyzing the data, coming to some interesting (and counter-intuitive) conclusions. In my experience, researchers have a sufficient feel for their discipline that surprising results are frequently a result of analysis error. Thus, I checked my analysis for errors.

I was actually able to check the analysis because I wrote a script — a series of commands — detailing every bit of the analysis I performed. Mouse-clicking my way through the analysis would make it all but impossible to check my work (something my math teachers in grade school always emphasized). Thus, I was confident that the analysis I returned to her were correct ...

... conditional on the data being correct.

The next day, she emailed me and let me know that she found several serious errors in her data. Relying on mouse-clicks would have made those original three hours a waste of time. However, once she sent the corrected version of the data, the analysis took 90 seconds. Clicking my way to re-analysis would have taken the same amount of time as the first analysis — time we did not have (we were facing a deadline). Re-running the script only took processing time.

While I *am* a fan of mouse-clicks under many circumstances, I am not a fan when it comes to serious statistical analysis. Scripting provides at least three definite advantages over mouse clicks: You get what you request, you can check your work easily, and you can re-run the analysis with little effort.



One thing I love about being an academic is that I get to travel the world sharing my research and its results. One June found me standing in front of a formidable group in Odense, Denmark, giving such a presentation. After I had discussed the current literature on the causes of terrorism, my statistical model, the results, and my conclusions, I opened it up for questions. After about five minutes of questions on the validity of my statistical model, one professor asked how my analysis would change were I to add an offset to the model.

After a brief panic, I decided to actually run the regression with the offset in front of the audience. I apologized for not knowing the answer to his question, but I would be happy to hypothesize the effect of the variable offset while running the analysis. The professor smiled as I tried to open my analysis script to modify and run it. It turned out that R was not installed on the presentation computer. No worries. I opened the R folder on my USB drive, double-clicked on the R program, and proceeded with the altered analysis — all the while discussing the theoretical effects of using such an offset under these circumstances. Before I could finish hypothesizing, R gave me the answer (which, thankfully, agreed with my hypotheses).

Now that I had my model laid bare before all, many more questions arose about different alterations I could (or should) make to the model. All of which I was able to perform in front of the now hyper-interested crowd.



These three vignettes illustrate many strengths of a statistical environment like R. First, it encourages one to write out the analysis and “show the work.” This makes it easier to see the entire scope and sequence of the analysis. It also makes it easier to check for errors. Second, it is extensible. If there is a cutting-edge test or procedure you wish to run, there is probably a package that contains it. If not, you are quite free to write it yourself. Finally, one can carry R around on a USB drive, allowing anyone to perform analyses whenever there is a computer, like in Denmark.

Oh yeah, that R is free is also a nice feature, especially as statistical packages can run from \$600 to \$6000 and up, *and* can have licenses requiring annual payments. As budgets get tighter, an ability to work successfully with a free (and powerful) statistical environment is invaluable.

Prerequisites

For any book (or course), there is a necessary assumption made about the background of the reader (or student). For the material in this book, I assume that you have had experiences with elementary statistics, matrices, and calculus.

In **differential calculus**, you will need to understand how to optimize (minimize or maximize) functions. In **integral calculus**, you should understand how to calculate areas under a curve (probabilities related to density functions). Beyond that, there is little calculus needed.

The **matrix topics** you need consist of being able to perform algebra on matrices. Beyond that, anything you remember from a typical linear algebra course will help things make a bit more sense. Matrix rank, invertibility, idempotency, projection matrices, orthogonality, etc., are all important in ordinary least squares. So, if you remember those topics, you will be ahead of the curve. If you do not remember them, then you will need to (re-)learn them in this course. Appendix M will help with remembering and learning the important matrix topics.

Finally, I wrote this book to be a second course in statistics, one that started where a typical **introductory course** ended. Because of this, I also assume you remember many topics from such a course. These topics definitely include the meanings of confidence intervals and p-values. They also include probability distributions, t-tests, issues with multiple testing, and the Central Limit Theorem (CLT). To help refresh your memory, work through Appendix S. Note that Appendix S also introduces you to some (optional) advanced items. These topics were included at the request of past students who wanted to actually see a proof of the Central Limit Theorem. Rest assured that understanding the CLT is more important than being able to prove it. Furthermore, the proof offers little in the way of a deeper understanding.

A Note on Notation

Sadly, notation varies across the discipline. This is a result of the history of statistics: Many of the methods came from disciplines that *used* statistics, rather than from statistics itself. Different disciplines use different notation for the same idea. Thus, any discussion of Survival Analysis needs to include Event History Analysis and Reliability Analysis, as they all study the same phenomena but from different disciplines (medicine, social sciences, and engineering, respectively).¹

Even within a discipline, there is often a variety of notation used to indicate the same ideas. For instance, probability functions are often parametrized in different ways. The parameter of the Exponential distribution can be the rate λ or the expected value θ ; the second parameter of the Normal distribution (Gaussian distribution, Gauss-Laplace distribution) may be the variance σ^2 , the standard deviation σ , or the precision as $\tau = 1/\sigma$ or as $\tau = 1/\sigma^2$. The symbol for the average rate in the Poisson distribution can be μ or λ . In this volume, I will (try to) keep consistent with notation, and I will explain the notation before I use it.

To that end, population parameters will be signified using Greek minuscules. Sets from which the population parameters can belong (parameter spaces) will be Greek majuscules. Both are included in Table 1. All random variables are Roman majuscules. All realized random variables (data) are Roman minuscules. Violations of these rules will exist, but should be kept to a minimum.

Thus, if we theorize that our measurements come from a population that is Normally distributed, with mean μ and standard deviation σ , we would specify that $\mu \in M$ and that $\sigma \in \Sigma$, where $M = \mathbb{R}$ and $\Sigma = (0, \infty)$.²

Now, if we know that the mean is 15 and the standard deviation is 10, I would write this as

$$X \sim \mathcal{N}(\mu = 15; \sigma = 10), \quad (0.1)$$

where the mean of the population is denoted by μ , the standard deviation by σ , and the Normal distribution by \mathcal{N} .

¹Furthermore, the term “reliability analysis” means different things in different areas. It could mean studies of how long until a part or a machine breaks. It could also mean how robust conclusions are to changes in model assumptions.

²Be aware of the difference between Σ , \sum , and \mathfrak{Y} . The first is the set of possible values of σ . The second is the symbol indicating the sum of what follows. The third is a generic covariance matrix. Frequently, the difference will be obvious from context. If it is not, ask me.

Minuscule Letter	Majuscule Letter	Name
α	A	alpha
β	B	beta
γ	Γ	gamma
δ	Δ	delta
ϵ, ε	E	epsilon
ζ	Z	zeta
η	H	eta
θ	Θ	theta
ι	I	iota
κ	K	kappa
λ	L	lambda
μ	M	mu
ν	N	nu
ξ	Ξ	xi
\omicron	O	omicron
π	Π	pi
ρ	R	rho
σ	Σ	sigma
τ	T	tau
υ	Y	upsilon
ϕ	Φ	phi
χ	X	chi
ψ	Ψ	psi
ω	Ω	omega

Table 1: The usual Greek alphabet in the canonical order. Being familiar with the letters will make it easier to recognize the implied meaning behind the letter.

Once we take those measurements, we would call the variable x . The difference between random variables and *realizations* of those random variables is that the random variable has a probability distribution associated with it; the realized data are just numbers.

By the way, if we wish to specify “parameter” in general, we use θ . As a result, the symbol for the generic parameter space is Θ .

Matrices (and vectors) will be indicated with **bold-faced** letters. Thus, \mathbf{x} is the data matrix (observed values) and \mathbf{X} is the data matrix (theoretical values).

Between these two, it *does* make sense to say something like

$$\mathbf{X} \sim \mathcal{N}(\mu = 15; \sigma = 10)$$

It does NOT make sense to say something similar about \mathbf{x} . The observed data do *not* have a theoretical distribution.

Conclusion

And so, with all of this said, turn the page and begin your trek through linear models. The first chapter introduces you to the topics of both linear models and the Kingdom of Ruritania. The former is the purpose of the book. The latter is a common theme and source of examples. Since the Kingdom of Ruritania does not exist, think of it as a generic country with no real information about it beyond what is given.

Had I used a real country, it would be perfectly defensible for the student to bring in real information about that country. This may cloud the intended statistical lesson.

Also, using Ruritania allows me the ability to be creative in my storytelling.

I hope you enjoy the journey.

~ Ole J. Forsberg
December 2024
Knox College of Illinois

Vyčkej Času
Adolph Heyduk

Nespěchej k štěstí, tiše jen,
vše nemůž' najednou býti;
nejprv se travou zjeví len,
pozděj jak modravé kvítí.

Poupě, jež touží růží být,
obalu hledí se zbavit;
dříve než můžeš v nebi žít,
musíš se v očistci stavit.

Translation:

Bide Your Time
Adolph Heyduk

Don't rush to fortune, ease your tracks,
all can't at once be present;
You would for grass mistake the flax
ere blue-flower'd iridescent.

A bud, so keen to be a rose,
is for its calyx sorry;
But ere in heaven you'll repose,
you'll bide in purgatory.

Source: <http://www.vzjp.cz/basne.htm#Heyduk>

The background of the page is a photograph of a large body of water, likely a lake or reservoir, under an overcast sky. The water's surface is shimmering with light. In the distance, a line of trees separates the water from a blue building with a white roof. The foreground is filled with the branches and leaves of trees in autumn, with some leaves showing yellow and brown hues. The top right corner of the page is a solid dark blue rectangle.

CHAPTER 1:

AN INTRODUCTION TO RURUITANIA

OVERVIEW:

Behind any good textbook is a good writer. Unfortunately, behind this textbook is me.

Clearly, the mathematics, probability, and statistics are correct. However, a good book creates a good story about the material. I sought to do this in every page by making the Kingdom of Ruritania the setting for the “story of the book.”

This first chapter provides background information about Ruritania and gives you some foreshadowing about what we will be doing in this book. I hope you enjoy it.

Chapter Contents

1	An Introduction to Ruruitania	1
1.1	Background of Ruruitania	3
1.2	Economics	5
1.3	US–Ruritanian Relations	6
1.4	Illustrations of Analyses	7
1.5	Conclusion	12



You are about to undertake an educational journey. This journey will take you to new and exciting places. Here, you will turn your mathematical knowledge into statistical knowledge, giving you skills in detecting relationships between variables in life.

This introduction has two primary purposes. The first is to introduce you to the Kingdom of Ruruitania (Řurità Kràlovství). This fictional country is used as the backdrop for many of the examples. Why use Ruruitania? It offers no conflicting information, thus allowing us to focus on the research questions at hand.

Second, this chapter provides many instances of analyses discussed in this text. Think of this chapter as a peek into the future, as a foreshadowing of the great things to come!



Figure 1.1: *His Majesty Rudolph II, King of Ruruitania.*

1.1: Background of Ruritania

The Kingdom of Ruritania (officially: *Řurità Kràlovství*) is a small kingdom (62 mi²; 161 km²) surrounded by Germany and the Czech Republic. Its inhabitants are Slavic-speaking Roman Catholics currently under the absolute monarchy of Rudolph II (Figure 1.1).

Under Rudolph II, ascension talks between Ruritania and the European Union have stalled. The two issues are the deficiency of democratic structures and an excess of controls on the press. Regardless of not being a member of the European Union, or of its monetary union (euro area), Rudolf pegged the Ruritanian Crown (*Koruna Řurità*) to the Euro at 2.00Kř to €1.00. This offers greater stability for the koruna in the world currency market. Three decades ago, economic reform allowed the koruna to be entirely convertible on the world market. It also completely eliminated the black market in Ruritania.

1.1.1 THE GEOGRAPHY Ruritania is land-locked. It is located between southern Germany (Saxony) and western Czech Republic (Bohemia). The land is a beautiful combination of high mountains in the west (the Ruritanian Alps) and rolling farmland in the east (the Ruritanian Veldt).

1.1.2 THE GOVERNMENT Note that Ruritania is an autocratic kingdom, not a constitutional monarchy; the king rules by fiat. While the king is advised by a council of ministers, they are people that he selects. Furthermore, these ministers need *not* be citizens of Ruritania. The current council is composed of five Ruritanians and one Oregonian: the current Minister of Economics, who is Knox-educated and from Portland, Oregon. These councilors also perform the tasks of a 'Committee King' when the king is unable to perform his duties as Chief of State and Head of Government.



Figure 1.2: *The Vlajka, the current flag of the Kingdom of Ruritania adopted in 1954. Blue represents the sky; white, the snow-capped peaks; green, the lush farmland; and red, the passion of the people.*

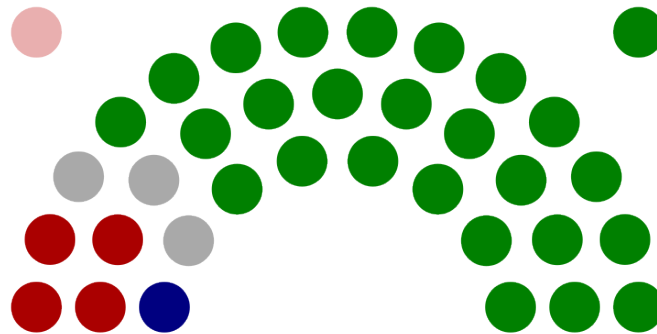
Strešlau, which serves as the official capital city, lies on the rail line between Dresden and Prague and has a population of 24,312. The royal capital, Sčwânstein, has a population of 7478 and lies on a spur (branch) line. According to the 2020 census, the total population of Ruritania is 43,670, with about 90% of the population living in its three main cities: Strešlau, Sčwânstein, and Děčín.

Ruritania is divided into seven states (*státy*). Each *stát* is named after its largest city. Thus, the seven *státy* are Děčín, Hora, Reka, Sčwânstein, Strešlau,

Venkovský, and Zámek. Each *státy* is further divided into five counties (*kraj*). Officially, the name of each *kraj* is the *stát* name followed by a letter between A and E.

In addition to the Council of Ministers, there is a parliament (*Národní Shromáždění*; National Assembly) that can also advise the King — when he wishes. Members of parliament are elected by all adults when the King calls for elections. Since World War II, elections have happened every four years or so. The 2024 election saw two parties field candidates in every *kraj*. In that election, the *Král a Země* (King and Country) party, led by Vasilij Vasiljevič Kuzněcov, won 26 of the 35 seats in the parliament. The *Republikánská Strana* (Republican Party), led by Saša Ondřej Ivanović, won the second-most number of seats, four.

The following illustrates the current structure of the parliament.



The green dots represent seats won by a *Král a Země* candidate; grey, independent; red, *Republikánská Strana*; blue, *Liberálně-Demokratická Strana* (Liberal-Democratic Party); and pink, *Komunistický* (Communist).

Note that the *Republikánská Strana*, *Liberálně-Demokratická Strana*, and pink, *Komunistický* parties for the *Republikán Bloc* — the parties that seek to eliminate the monarchy. The *Král a Země* and independents support the monarchy, as such, they are the *Monarčista Bloc*.

1.2: Economics

While Ruritania's size limits its ability to diversify its economy, Rudolf has been an able administrator and businessman. As such, he was able to lift many Ruritarians out of poverty. When he took power in 1940, the GDP per capita (PPP) was approximately USD50, with a poverty rate of 99%. Today, it is approximately USD55,000 (the eighth highest in the world), with a poverty rate of just 50%.

The primary source of revenue for Ruritania is its banking industry — the source of 75% of its GDP (see Figure 1.3). The remainder comes from tourism (15%) and agriculture (10%). Ruritania's tourism industry primarily bases itself on winter recreation in the west. Because of favorable visa requirements, low costs for hotels, and fantastic skiing in its alps, Ruritania is the destination of choice for vacationers from (in descending order of visitors) Czechia, Germany, Turkey, Oman, Morocco, United States, and Palau.

The primary crops are corn (55%), summer wheat (25%), and filbert nuts (15%), with soybeans and hops being secondary. What corn is not eaten is exported to Germany (75%) and Czechia (25%). Similar export patterns hold for the excess wheat. Filbert exports go to Czechia (40%), Germany (35%), and Switzerland (25%), where they are turned into delicious confections. Imports to Ruritania come from Russia, Turkey, and Oman (petroleum), and Czechia and Germany (manufactured goods and foodstuffs).

Because of the strength of the monarchy, Ruritania is neither a production point nor a transshipping point for drugs. Illicit drug use is the lowest in Europe, with approximately 2% of the population using marijuana, and none using harder drugs.



Figure 1.3: A tree diagram of the economic output of Ruritania.

1.3: US–Ruritanian Relations

The United States and Ruritania share full diplomatic recognition. However, the United States does not have an ambassador to Ruritania. US interests in Ruritania are handled by the US Ambassador to Poland, Mark Brzezinski (since February 22, 2022). This is not an unusual circumstance, as embassies are quite expensive to build and operate. Unfortunately, this reduces the amount of reliable information coming out of Ruritania.

Rudolf was a staunch ally of the United States during the Cold War. However, with the geographic position of his country in the world (entirely surrounded by Soviet satellite states East Germany and Czechoslovakia), he was able to offer the United States little more than intermittent vocal and moral support in the United Nations. For fear of losing sovereignty, Rudolf often kept quiet and followed the lead of the Soviet Union in all but domestic economic matters.

When the Soviet Union fell, Rudolf increased his support of the United States and its efforts to bring peace and prosperity to the world. As Rudolf often pointed out to various US presidents, Ruritania has never seen one of its sons die in battle. With his declining health and advancing age, Rudolf became much more vehement in his support of the United States, especially with respect to the Global War on Terror.



Figure 1.4: *US Ambassador to Poland, Mark Brzezinski.*

1.4: Illustrations of Analyses

That concludes the background to Ruritania. The following examples show you what you will be able to do by the time you finish this book. The illustrations provide neither the data nor the code. All they provide are examples of how linear models are helpful to Ruritania...and to the rest of the world. Read through them and become excited about what this term will bring to you!

1.4.1 ILLUSTRATION: THE MISSING KRAJ Every 10 years of a king's reign, Ruritania holds a census. Their last was in 2020, marking King Rudolph's 80th year as monarch. After compiling all information, they discovered that information was missing for one *kraj* (state). This is unfortunate, because King Rudolph needs the information to evaluate his latest five-year plan and determine what he should do to make it better.

To fill in the missing information, we can regress all other variables on the GKP per capita, then predict the GKP per capita based on the known values for the missing kraj.

With this data and model, we just predict the GKP per capita in the kraj. The most-likely value is \$2400, with a 95% prediction interval from \$2100 to \$2650.

From this information, His Majesty concludes that the plan helped the entire country, but did a better job with the rural areas than the urban. As a result of this analysis, he asks his ministers to generate a plan that does a better job of spreading the prosperity to more of the Kingdom.



This use of regression has a very important use: estimating values for missing data in a data set (imputation). Frequently, the amount of missing data will be significant with respect to the amount of complete data. In such a case, the researcher may use multiple imputation to create multiple data sets, estimate the parameters of interest on each, and report them and their standard errors.

impute

1.4.2 ILLUSTRATION: RURITANIAN CROPS As usual, His Majesty Rudolph II would like some input from his Council of Ministers on his next five-year plan. Currently, the primary crop in Ruritania is corn. To help optimize the profits made by farmers, Rudolph wants to know if that crop should be changed to summer wheat or to soybeans.

To help him, let us model the relationship between farmer profit and crop in Ruritania. The dependent variable is the profit per acre, and the independent variable is the crop. Using linear models, we see that it is more profitable at this point to grow wheat. The average profit per acre is \$845.75. This is about \$150 greater than that of corn and about \$300 greater than soybeans.



Linear models can also be applied to cases where the independent variable is categorical, as here. This method is actually termed “analysis of variance” (ANOVA). The difference between ANOVA and regression is only conceptual, in that each level of the categorical variable is treated as a separate variable.

Using statistics to inform policy decisions is an important use. However, we professionals need to be aware of the limitations of our research — and let our clients know them as well. Here, we only looked at the current average profit per acre for each crop. Future prices may fluctuate enough to make soybeans more valuable. Furthermore, shifting all Ruritanian crops to wheat puts the entire economy at risk of a drop in wheat prices. Diversification may be the better strategy, with some of the profit shared among all farmers.

1.4.3 ILLUSTRATION: COWS IN DĚČÍN The voters of DĚČÍN are being sent to the polls to vote on a city referendum that proposes to limit the number of cows that can be kept inside the city limits. The wording on this referendum is quite similar to ones proposed over the past several years.

Given the information from previous votes, the Director of the Independent Electoral Commission (NVK) estimates that the probability the referendum has of passing is 40%, with an estimated 48% of the voters supporting it.



Note that this research question deals with the *probability* of winning, not just the best estimate of the vote in favor. This requires estimating the entire probability distribution of the dependent variable.

While there are some rather sophisticated methods, we will be able to answer a similar question using Monte Carlo simulation. Such simulation consists of drawing large samples from each of the parameter-estimate distributions, calculating a predicted outcome for each of those sets of estimates, and examining the distribution of these predictions.

Monte Carlo estimation is a powerful technique that allows you to estimate results when the assumptions of the mathematical model are not fully met by the data.

1.4.4 ILLUSTRATION: WEALTH IN RURITANIA The gross domestic product (GDP) per capita is one of many measures of average wealth in countries. If extant theory is correct, then the wealth in the country is directly affected by the level of corruption in the government — countries with higher levels of corruption should be poorer (on average) than those with low levels of corruption. Furthermore, if theory is correct, the level of democracy in a country should *also* influence the country’s level of wealth — countries with higher levels of democracy should be wealthier than countries with lower levels of democracy.

His Majesty is curious to see how Ruritania fits in this model. If the actual GDP per capita is greater than what is expected from modeling the rest of the world, then Rudolph is doing a great job as king. Otherwise, he needs to improve the lot of his people.

And so, to help Rudolph, we predict that the GDP per capita for Ruritania, according to the model, is \$26,795.64, with a 95% prediction interval from \$5232 to \$48,360. Since the actual GDP per capita is \$55,000, King Rudolph is happy that he is better than average at guiding Ruritania forward towards prosperity.



Frequently, we can use our models in novel ways. Usually, we would model the data and calculate predictions and confidence intervals.

However, if we have confidence in our model, we can use it to determine which units are under- or over-performing expectations (the line of best fit). In this case, Ruritania’s GDP per capita is significantly higher than what the model predicts.

This means either the model needs to take more covariates into consideration or that Ruritania is much more prosperous than one would expect... or both.

1.4.5 ILLUSTRATION: ELECTIONS IN RURITANIA Even though it is an absolute monarchy, national elections are held in Ruritania to elect members of the Ruritanian parliament, the *Národní Shromáždění* (National Assembly).

After the most recent election, Ruritanian exiles in Denmark claimed that the ballot boxes were stuffed. That is, the ballot boxes had votes for the government party in them even before voting began. Because guarantees of the “secret ballot” are built into the Ruritanian Constitution, the ballot boxes are opaque. As such, there is no direct evidence of stuffing.

Ordinary least squares regression is not well-suited for this type of data. There is inherent heteroskedasticity in the proposed model. However, we can use Binomial regression to test the election for evidence of ballot box stuffing.

As a result of our analysis, we were able to detect some evidence for stuffing (p-value = 0.0441). However, because the p-value is so high with respect to the usual $\alpha = 0.05$ value, do we really have the necessary support in the data to claim there was unfairness in the election?

To claim something so heinous, we should really contemplate the real meaning of the p-value and selecting an appropriate value for α .



Here, we had a good understanding of the data-generating process. This allowed us to use that understanding to create a stronger model. Heteroskedasticity can be “adjusted for” in ordinary least squares. It should be a full part of the model, however.

1.4.6 ILLUSTRATION: INSURANCE IN RURITANIA The decision to buy life insurance is related to several variables, including age and income. We would like to explore this relationship in Ruritania.

Since the dependent variable is dichotomous (life insurance purchased *or* life insurance not purchased), we need a new type of regression to ensure that our predictions make sense. One option is called logistic regression.

Using this regression, we find significant positive relationships between the person’s age and the likelihood to buy life insurance, as well as between the person’s income and the likelihood to buy life insurance.

Additionally, we predict that the Knox graduate on the Council of Ministers, a 65-year-old making \$125,000 annually, has a 74% chance of having life insurance.



In this example, we had to use a different type of regression because the dependent variable was dichotomous, could only take on two values. This type of regression is known collectively as logistic regression, even though the link function can be almost any that map the real line to the interval (0,1). Such functions include the venerable logit function (inverse of the logistic function). It also includes the probit (used in a lot of medical studies) and the cauchit (used in some financial studies to allow for highly variable events).

It was this class of problems that forced Nelder and associates to formulate an over-arching framework for regression. He called it the “generalized linear model” (GLM). While he rues the name to this day, he created it to signify that this class of regression problems is actually just a generalization of the class that can be solved using ordinary least squares regression. In other words, OLS regression is a special case of GLM regression.

1.4.7 ILLUSTRATION: WARMTH FOR THE KING Finally, let us help the King be more beloved of his people — if that is possible. We took a poll and asked Ruritarians their ‘warmth of feeling’ for Rudolph and his political agenda. In addition to this one variable, we also asked several demographic questions, allowing us to provide suggestions to the King. The demographic information includes gender, race, age, and number of years of education. The response variable has four ordered levels: Strongly Disagree, Disagree, Agree, and Strongly Agree.

With this information, we are able to let the King know that he is widely loved, but that women tend to agree with his policies more than do men. Furthermore, the better-educated also tend to support him more. The younger members of Ruritania also feel warmer towards him and his agenda. Finally, there was no relationship between race and support; all races seem to support him equally.

With this information from the poll, what can Rudolph do to help his people? Such is the question royals have asked for generations.



Noting that the dependent variable is an ordinal variable, we could not use ordinary least squares regression. We had to use something called “ordinal regression.” The concepts behind ordinal regression are quite similar to those behind other types of regression covered in this book. The mathematics are a bit more difficult, however.

Thankfully, statistical programs make using ordinal regression almost as easy as using other types. We just have to know how to get the data in the right form for the program *and* how to test the assumptions made by the technique.

In fact, this seems to be the lesson we need to learn throughout this course. The concepts are quite similar. The mathematics are different. And, to make usage easier, those who write statistical environments try to make the functions as similar as they can.

1.5: Conclusion

And that brings us to the end of this introductory chapter. In this chapter, you were introduced to the incredible Kingdom of Ruritania. Ruritania offers us a rich source of examples, which will be exploited throughout the text.

This chapter also offered several examples that foreshadow what you will learn in this course. While you may not be able to do — or understand the underlying theory for — any of them at this point, you will by the end of the course.

So, take a deep breath and turn the page to the book's first part: Ordinary Least Squares (OLS). In this part of the book, we start by asking what we mean by “summarizing the relationship with a line of best fit.” From that point, we leverage that definition to inform our mathematics, thus allowing us to create formulas for estimating the population parameters.

After that chapter, we use the mathematics and elementary probability theory to create test statistics and confidence intervals for testing hypotheses of interest.

And after that... the sky is the limit! It is a great journey, and King Rudolph II thanks you for starting it.



Figure 1.5: *Rudolph II, King of Ruritania, in 1952.*

Part I

Least Squares

2	Introduction to Linear Regression	15
3	Matrices and Linear Regression	43
4	Improved! Now with Probabilities	75
5	Dood! Check the Requirements	105
6	A Time for Some Examples	151
7	Fixing the Violations	189

CHAPTER 2:

INTRODUCTION TO LINEAR REGRESSION

OVERVIEW:

Regression is a set of methods that seek to learn the specific *relationship* between one or more influenced (dependent, response) variables and one or more influencing (independent, predictor) variables. There are many existing regression methods, each focusing on different ways of determining how best to quantify that relationship.

As is tradition, this chapter starts with our first definition of “best fit” and derives many results from that definition. This chapter is entirely mathematical in that probability distributions are not considered (until Chapter 4).

Chapter Contents

2	Introduction to Linear Regression	15
2.1	Scalar Representation	20
2.2	Results	31
2.3	First Assumptions	34
2.4	The PRE Measures	38
2.5	Conclusion	41
2.6	End-of-Chapter Materials	42



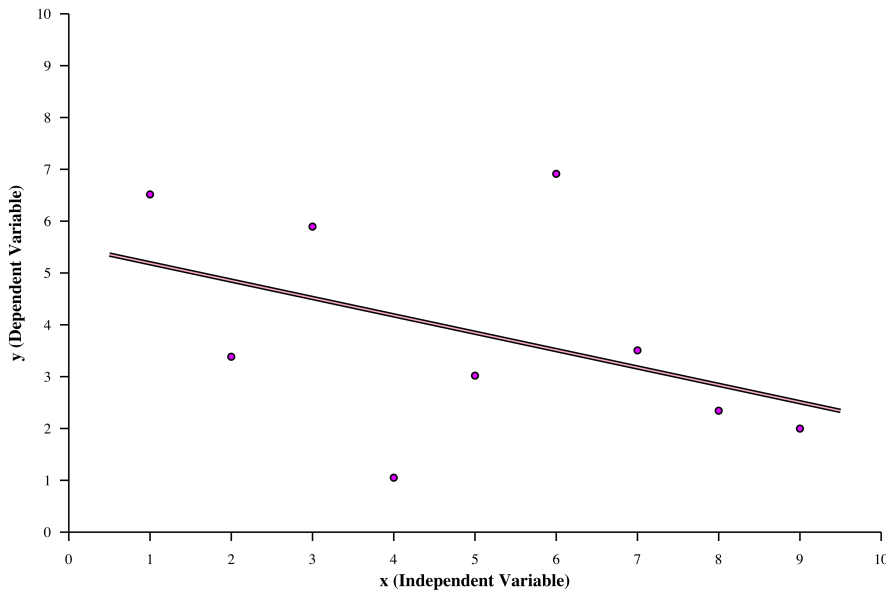


Figure 2.1: Sample data and a line of best fit for that data. Note that the slope of the line is negative. This indicates that increasing values of x **tend to** correspond to lower values of y . Regression detects for such trends.

Let x and y be numeric variables. The linear relationship between x and y can be summarized by a line that “best” fits the observed data. That is, we can summarize the relationship between x and y using a linear equation:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (2.1)$$

Here, parameter β_1 represents the **slope** and parameter β_0 represents the **y-intercept** (the value of y on the line when $x = 0$). The slope is usually the only thing in the equation that is interesting; it is the effect of x on y . The ε represents the vertical distance between the observation and the population line of best fit. It contains all of the things that affect y that are not included in x .

We said that the line given in equation 2.1 “best” fits the observed data. What we mean by “best” determines where we go from here. In thinking about “best,” it may help to see some sample data and the “line of best fit” for it (Figure 2.1).

A good statistician will ask:

What makes *this* line the “best”?

Here is the answer:

It depends.

best

Note that there are *at least* three definitions of “best” that we can use:

1. Maximize the likelihood that the data were generated
2. Minimize the sum of the absolute value of the residuals
3. Minimize the sum of the square of the residuals

estimator

All three definitions are entirely legitimate — as are many other definitions. However, each leads to different estimation methods and estimators.

Note: While different models will usually give different estimates, the substantive conclusions will rarely differ significantly in a well-formed model.

model

Note that the result will be a line represented by

$$\hat{y} = b_0 + b_1x \quad (2.2)$$

Using Latin characters indicates that these are based on your particular sample; they are sample estimates. Contrast this with using Greek characters to indicate population parameters. The “hat” on the y indicates that this is an estimate. All together, this is our **model equation**. It is the equation of the line of best fit based on the data you collected.

MLE

The first definition leads to “**maximum likelihood estimation**,” which will be covered in Chapter 10. It is an excellent technique that can be generalized to many more settings than can ordinary least squares. Its greatest strength is that it makes use of the researcher’s greater understanding of the data-generating process (Chapters 10 to 15). Its greatest weakness is the mathematics involved.

robust

The second definition leads to a type of robust regression frequently termed “**median regression**.” This method is helpful for times when there are outliers in the data that you cannot (or *should* not) remove. The drawback to this method is that estimating the two parameters (β_0 and β_1) does not provide a closed-form solution. In other words, it requires a repetitive sequence of steps and can only approximate those estimates. Furthermore, the approximation process is computationally intensive. Because of this, median regression was little used until recently. Because of this, the statistical theory behind it is not as well explored as other types. We will see this in Chapter 9.

The most popular definition of “best,” and the one that starts our journey, is the final definition. It leads to an estimation method called **ordinary least squares**

(OLS). It is rather straight-forward to minimize a sum of squared values using differential calculus. One strength is that an equation results from this process — a closed-form solution with no need for iteration. This means that the process returns mathematically exact values. The drawback is that it is limited in the types of processes that can be modeled.

We start exploring ordinary least squares immediately.

2.1: Scalar Representation

This section and the next show how our definition of “best” mathematically leads to specific results. That leading can be done by representing the regression problem in scalar or in matrix form. At one level, there is no difference in the two representations. At another level, one representation may make proofs — and understandings — easier and much more manifest. And so, let us begin with the scalar representation of the regression problem. From experience, it seems to make more sense than starting with the matrix representation (Chapter 3).

Ordinary least squares estimation defines “best” as “having the lowest sum of squared errors.” These errors (or residuals) are the *vertical* distance between the observed point and the corresponding point on the line (see Figure 2.2). With this explanation of what we mean by “error” in this context, we can use our third definition of “best” to obtain the OLS estimators for β_0 and β_1 .

Remember that to optimize (maximize or minimize) a function using calculus, one takes its derivative(s) with respect to the parameter(s) of interest, sets the resulting equations equal to 0, then solves the system of equations.¹

And so, the first step is to form the objective function that we want to minimize. Since we seek to minimize the sum of squared errors, that Q function is the sum of squared errors:

$$Q = \sum_{i=1}^n \varepsilon_i^2 \quad (2.3)$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.4)$$

$$= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (2.5)$$

$$= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.6)$$

Now that we have the objective function, we take its derivative with respect to each parameter, set it equal to 0, and solve for that parameter.

¹Also, one should perform the second derivative test to determine the type of optimization point found: minimum, maximum, and saddle point (neither).

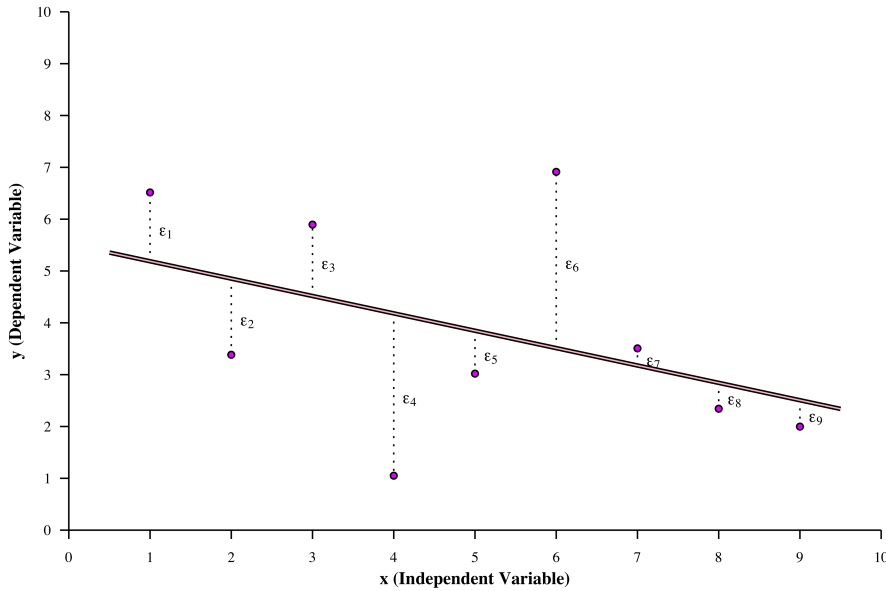


Figure 2.2: Sample data and a line of best fit for that data. Also marked are the residuals, the difference between what was observed (dots) and what is predicted by the model (line). This particular line of best fit minimizes the sum of squared residuals.

Let us start with β_0 :

$$\frac{\partial}{\partial \beta_0} Q = \sum_{i=1}^n -2(y_i - \beta_0 - \beta_1 x_i) \quad (2.7)$$

$$0 \stackrel{\text{set}}{=} \sum_{i=1}^n -2(y_i - b_0 - b_1 x_i) \quad (2.8)$$

$$= \sum_{i=1}^n y_i - \sum_{i=1}^n b_0 - b_1 \sum_{i=1}^n x_i \quad (2.9)$$

$$= n\bar{y} - nb_0 - nb_1 \bar{x} \quad (2.10)$$

This immediately leads to

$$b_0 = \bar{y} - b_1 \bar{x} \quad (2.11)$$

This is called the OLS estimator of β_0 . Note that this formula needs \bar{x} and \bar{y} . These are both easily calculated from the data. This formula also need b_1 , which is the OLS estimate of β_1 . Thus, we will need to determine the value of b_1 to use this formula.

And so, to obtain a formula for b_1 , we take the derivative of Q with respect to the second parameter, β_1 :

$$\frac{\partial}{\partial \beta_1} Q = \sum_{i=1}^n -2x_i (y_i - \beta_0 - \beta_1 x_i) \quad (2.12)$$

$$0 \stackrel{\text{set}}{=} \sum_{i=1}^n -2x_i (y_i - b_0 - b_1 x_i) \quad (2.13)$$

$$= \sum_{i=1}^n x_i y_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 \quad (2.14)$$

$$= \sum_{i=1}^n x_i y_i - n b_0 \bar{x} - b_1 \sum_{i=1}^n x_i^2 \quad (2.15)$$

Substituting our estimator b_0 , we have

$$0 = \sum_{i=1}^n x_i y_i - (\bar{y} - b_1 \bar{x}) n \bar{x} - b_1 \sum_{i=1}^n x_i^2 \quad (2.16)$$

$$= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} + b_1 n \bar{x}^2 - b_1 \sum_{i=1}^n x_i^2 \quad (2.17)$$

$$b_1 \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \quad (2.18)$$

Finally, we have

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad (2.19)$$

Thus, the two OLS estimators of β_0 and β_1 are

$$\begin{cases} b_0 = \bar{y} - b_1 \bar{x} \\ b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \end{cases} \quad (2.20)$$

requirement

Note that this mathematical process had but one requirement:

$$\sum_{i=1}^n x_i^2 - n\bar{x}^2 \neq 0 \quad (2.21)$$

If that requirement is not met by the data, then the divisor of b_1 is zero in equation 2.19, which leads to dividing by zero, armageddon, and a *really* bad hair day. However, note that

$$\sum_{i=1}^n x_i^2 - n\bar{x}^2 = (n-1) s_x^2 \quad (2.22)$$

As such, this requirement is met when the variance of x is non-zero. In other words, we require that the independent variable varies.

exercise

Note: For a mathematician, this is an important observation. For a statistician, it gives insight into how to “break” OLS: Measure all of the observations using the same value of the independent variable. In other words, to a statistician, the steps have meaning beyond the mathematics.

Also note that some sources will give the formula for b_1 as:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.23)$$

exercise

I leave it as an exercise to show that $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ is equivalent to $\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$ and that $\sum_{i=1}^n (x_i - \bar{x})^2$ is equivalent to $\sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Definition 2.1. S_{xx}

We will come across the denominator of equation 2.23 in many settings. Thus, to save ink, we will symbolize it as S_{xx} and define it as:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.24)$$

Thus, our OLS line of best fit is the line defined by the set of points (x, \hat{y}) , where

$$\hat{y} = b_0 + b_1 x \quad (2.25)$$

Note that \hat{y} is the expected value of Y (dependent variable), given that value of x (independent variable). In other words,

$$\hat{y}_i = \mathbb{E}[Y | x_i] \quad (2.26)$$

It is the **conditional mean** of Y given x_i ; the expected value of Y , given this value of x_i ; the mean of Y when the independent variable has value x_i .

Note: There is a difference between an “expected” and a “predicted” value. The **expected value** is the mean: If you were to rerun the universe a gazillion times, collected the same amount of data, and estimated \hat{y} each time, the expected value is the average of all of those \hat{y} s. The **predicted value** is the value of an additional observation measured at the same value of x .

That the two are both on the line is happy happenstance — but happenstance nonetheless. The main difference, as you will discover, is in the intervals that surround that value. Intervals on predictions (**prediction intervals**) are wider than intervals on the mean/expected value (**confidence**

intervals). This is because we are more uncertain about a future value than we are about an average.

And this is as far as we can go without making additional assumptions. As such, it marks a great place for a toy example.²

Example 1

Let us measure two variables on four subjects. Those two variables are x and y . For the first subject, the value of x is -2 and the value of y is 3 . For the second subject the x and y values are 0 and 0 . For the third subject, the values are 0 and 2 . For the fourth subject, they are 2 and -1 .

Given this information, let us calculate the ordinary least squares estimators of β_0 and β_1 .

²Such examples are called “toy” examples because they are simple to work through, not because they deal with toys.

Solution: First, the formulas for b_0 and b_1 require we calculate \bar{x} and \bar{y} . To do this, we need the data. Here they are

x	y
-2	3
0	0
0	2
2	-1

The means are 0 and 1, respectively. And, with that, we can use the formula for b_1 (Equation 2.20b):

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad (2.27)$$

$$= \frac{(-2(3) + 0(0) + 0(2) + 2(-1)) - 4(0)1}{((-2)^2 + (0)^2 + (0)^2 + (2)^2) - 4(0)^2} \quad (2.28)$$

$$= -1 \quad (2.29)$$

For the OLS estimator of the intercept, β_0 , we have (Equation 2.20a):

$$b_0 = \bar{y} - b_1 \bar{x} \quad (2.30)$$

$$= 1 - (-1)0 \quad (2.31)$$

$$= 1 \quad (2.32)$$

Thus, the OLS line of best fit is the line defined by the set of points (x, \hat{y}) , where

$$\hat{y} = 1 - 1x \quad (2.33)$$

Figure 2.3 shows the points and the OLS line of best fit.

So, what does the equation *mean*? It means that the expected value of Y when $x = 0$ is 1, the y-intercept. It also means that for every one increase in the value of x , the expected value of Y increases by -1 (decreases by 1), which is the value of the slope.

To go beyond this rote interpretation, we need to know what the numbers represent. That information was lacking in this toy example. ♦

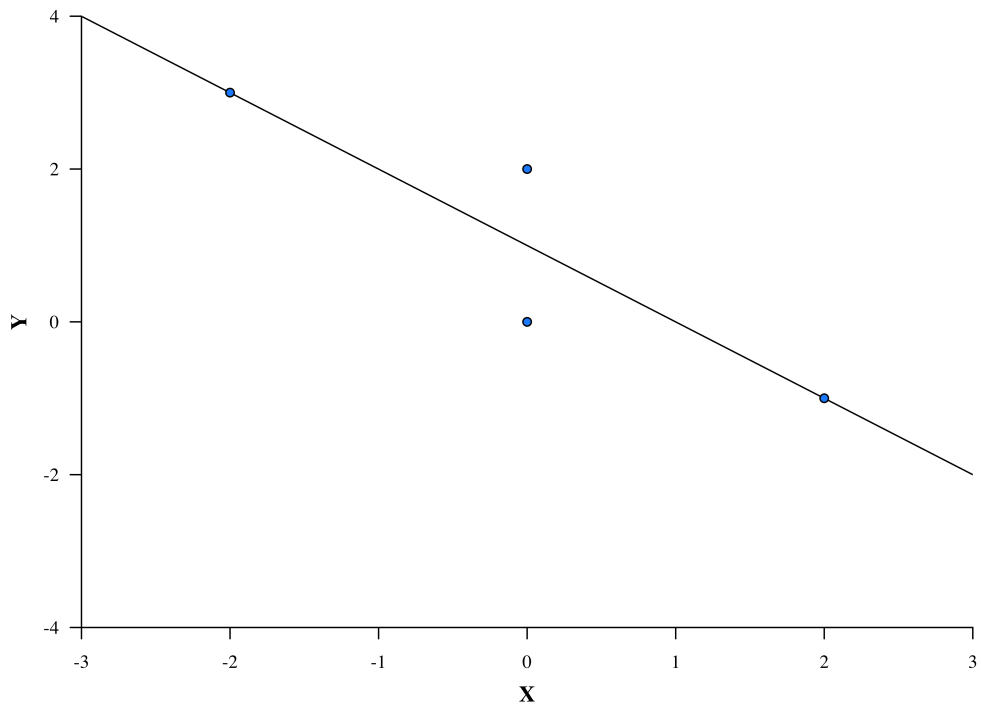


Figure 2.3: Graphic of the data and the OLS line of best fit for the toy data of Example 2.1.

Note: From a scientific standpoint, it is dangerous to interpret the y-intercept when $x = 0$ is outside the observed range of the data (x-values). Models are best when you are trying to understand the relationship within the observed ranges of the independent variable(s). This is **interpolation** — “inter” from “within.”

interpolation

Trying to use the model to understand relationships outside the observed values of the independent variables is called “**extrapolation**,” where ‘extra’ means ‘outside.’ Extrapolation is dangerous: all curves look linear at a small enough scale (remember Newton’s Method from Calculus). Thus, fitting the data with a line may be a good approximation in one scale, it may not make sense at a wider range, where the non-linearity of the relationship may become more pronounced.

Example 2

Let us measure two variables on four Ruritanian subjects. Those two variables are *handedness* and *time* to cut a sheet of paper. Sounds familiar? The difference here is that the independent variable is dichotomous.

Given the information in the table, let us calculate — and interpret — the ordinary least squares estimators of β_0 and β_1 .

Solution: Here are the data:

Handedness (x)	Time (y)
Left	3
Right	1
Right	2
Left	2

The first thing to do is change our independent variable into a numeric variable. When the variable is dichotomous (has only two possible values), this is easy. Set one value to 0 and the other to 1. So, without loss of generality, let us follow the alphabet and replace `Left` with 0 and `Right` with 1. With this transformation, the values for handedness are now $\{0, 1, 1, 0\}$ and we can use the same procedure as we used in Example 2.1.

dichotomous

First, the formulas for b_0 and b_1 require we calculate \bar{x} and \bar{y} . They are 0.5 and 2, respectively. With that, we can use the formula for b_1 (Equation 2.20b):

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad (2.34)$$

$$= \frac{(0(3) + 1(1) + 1(2) + 0(2)) - 4(0.5)2}{((0)^2 + (1)^2 + (1)^2 + (0)^2) - 4(0.5)^2} \quad (2.35)$$

$$= \frac{3 - 4}{2 - 1} \quad (2.36)$$

$$= -1 \quad (2.37)$$

For the OLS estimator of the intercept, β_0 , we have (Equation 2.20a):

$$b_0 = \bar{y} - b_1 \bar{x} \quad (2.38)$$

$$= 2 - (-1)0.5 \quad (2.39)$$

$$= 2.5 \quad (2.40)$$

Thus, the OLS line of best fit is the line defined by the set of points (x, \hat{y}) , where

$$\hat{y} = 2.5 - 1 x \quad (2.41)$$

Now that we have some context, what does *this* equation mean?

Remember that an x -value of 0 indicates we are discussing left-handed people. Thus, the expected value of Y for the lefties is $1.5 + (1)0 = 1.5$. The expected value of Y for right-handed people is $1.5 + (1)1 = 2.5$.

Thus, the y -intercept is the predicted value for **base level** (lefties). The “slope” is the “effect of handedness” (moving from left- to right-handed) on that y -intercept. ◆

base level

You have seen an analysis of this type in your past introductory statistics course. This is just the **two-sample t-procedure** under the guise of linear models.

Note: Since we can compare the means of two group in the regression realm (Example 2.1), can we compare the means of more than two groups? In other

words, can we extend linear models to ANOVA? The answer is Yes! In fact, **ANOVA** is built on a base of linear models, as we will see in the future (Example 3.4).

Please do not forget that all statistical procedures have requirements that have to be met. So far, the only requirement is that there is variation in the independent variable.

To draw stronger conclusions, perhaps calculate confidence intervals and test hypotheses, we will need to make stronger requirements. We will do this in the future. For now, let's just *require* that there is variation in the independent variable.

2.2: Results

Now that we have formulas for our estimators, we have several important mathematical results. The first result is that the OLS line of best fit passes through the center of gravity.

Theorem 2.2.1

The point (\bar{x}, \bar{y}) , the center of gravity, is on the OLS line of best fit.

Proof. To see this just substitute \bar{x} for x in the prediction equation and show that $\hat{y} = \bar{y}$.

From equation 2.2,

$$\hat{y} = b_0 + b_1 x \quad (2.42)$$

Substituting \bar{x} for x gives

$$= b_0 + b_1 \bar{x} \quad (2.43)$$

Substituting the value of b_0 gives

$$= (\bar{y} - b_1 \bar{x}) + b_1 \bar{x} \quad (2.44)$$

Finally, simplification gives our result:

$$= \bar{y} \quad (2.45)$$

Thus, we have shown that $\hat{y} = \bar{y}$ when $x = \bar{x}$. In other words, we showed that the OLS line of best fit passes through the center of gravity. \square

Note: You should also be able to prove that *any* line passing through the center of gravity has the sum of the residuals being zero.

exercise

Example 3

Illustrate this result using the previous example. In other words, show that the point $(\bar{x}, \bar{y}) = (0.5, 2)$ is on the line.

Solution: We have already shown that the line of best fit is $\hat{y} = 1.5 + x$. Substituting $\bar{x} = 0.5$ gives $\hat{y} = 1.5 + 0.5 = 2$. Note that 2 is also the value of \bar{y} .

Thus, we have illustrated the result of Theorem 2.2. ♦

second

A second result is that the slope estimator b_1 is the ratio of the covariance between x and y to the variance of x .

Theorem 2.2.2

An equivalent formula is

$$b_1 = \frac{\text{Cov}[x, y]}{\text{V}[X]} = \frac{s_{xy}}{s_x^2} \quad (2.46)$$

Proof. To see this we substitute the formulas for the covariance and variance into this equation and quickly simplify:

$$b_1 = \frac{s_{xy}}{s_x^2} \quad (2.47)$$

$$= \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.48)$$

$$\begin{aligned} & \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \tag{2.49}$$

□

A third result is that the slope estimator can also be represented as

third

$$b_1 = r_{xy} \frac{s_y}{s_x} \tag{2.50}$$

That is, the slope estimator is the correlation between the two variables times the ratio of their standard deviations. I leave this as an exercise for you to prove.

exercise

A fourth result is that the slope estimator is zero if the y-values do not vary. I leave this as an exercise, as well. By the way, this is a “three-line proof.”

exercise

2.3: First Assumptions

This was fun. We were able to determine the correct formula for the line of best fit — given our particular definition of “best.” Those equations lead to other equations. These are the *mathematical* results for our given sample.

Cool mule.

Until this point, we have only required variation in the independent variable. If we make three additional assumptions, we have additional results.³

Recall that the data model is

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.51}$$

With this, here are the three assumptions we will make, all about the residuals:

- The first assumption is that the residuals are realizations of a random variable (ε has a distribution).
- The second is that the expected value of the residuals is zero, $\mathbb{E}[\varepsilon] = 0$ (the measurements are not systematically biased).
- The third is that the residuals are independent and have a finite and constant variance, $\mathbb{V}[\varepsilon] = \sigma^2 < \infty$ (the residuals are homoskedastic).

The above simple assumptions lead to several additional interesting results. Some are proven here, some are left as exercises.

³This is how mathematical statistics progresses. Assumptions are made, then we play with the equations to learn about the consequences of those assumptions. Then, when we have exhausted our efforts, we make additional assumptions... *ad infinitum*.

Theorem 2.3.1

The OLS estimator for β_1 is unbiased. That is,

$$\mathbb{E}[b_1] = \beta_1 \quad (2.52)$$

Proof. To prove this, we will start with the formula for b_1 and simplify until we obtain the results.

$$\mathbb{E}[b_1] = \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad (2.53)$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad (2.54)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.55)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + x_i\beta_1 + \varepsilon)}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.56)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})\beta_0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})x_i\beta_1}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.57)$$

$$= \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\varepsilon \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.58)$$

$$= \beta_0 \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \varepsilon \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.59)$$

$$= \beta_1 \quad (2.60)$$

Thus, the OLS estimator of β_1 is unbiased. This is a nice property. It means $\mathbb{E}[b_1] = \beta_1$. \square

Note: Did you notice where we used these three results?

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i, \quad (2.61)$$

$$\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})x_i, \text{ and} \quad (2.62)$$

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad (2.63)$$

All three are just simple algebra.

Here are a few other results for you to prove:

Theorem 2.2. $\mathbb{E}[b_0] = \beta_0$

Theorem 2.3. $\mathbb{V}[b_1] = \sigma^2/S_{xx}$

Theorem 2.4. $\mathbb{V}[b_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right)$

Theorem 2.5. $\text{Cov}[b_0, b_1] = -\sigma^2 \frac{\bar{x}}{S_{xx}}$

Finally, let us define the mean square error (MSE) in the case of simple linear regression, SLR (one dependent and one independent variable).

MSE

Definition 2.6. *Mean Squared Error*

$$\text{MSE} = \frac{1}{n-2} \sum_{i=1}^n e_i^2 \quad (2.64)$$

This definition leads to the following theorem.

Theorem 2.3.2

$$\mathbb{E}[\text{MSE}] = \sigma^2$$

unbiased

In other words, definition 2.6 provides an *unbiased* estimator of the variance of the

residuals. This is why we define it in this manner.

Note: Be aware that definition 2.6 only holds in the case of simple linear regression (SLR); that is, it holds when there is just one dependent and one independent variable.

I leave this space to give you the opportunity to **prove this theorem**.

2.4: The PRE Measures

PRE

Now that we have reality (y_i) and our errors (e_i), we can create a measure of how well the model summarizes (fits) the data. In fact, we will create two of them! Both are “proportional reduction in error” measures; that is, they both are measures of how well the model reduces the unexplained variation in the dependent variable. The first is the venerable R^2 (“R-squared”) measure. The second is the \bar{R}^2 (“adjusted R-squared”) measure.

Both measure how much the model reduces the variation in the dependent variable. They differ in how that variation is measured. The R^2 measure uses the sum-of-squares; the \bar{R}^2 , the variance.

2.4.1 R^2 MEASURE The formula for the R^2 measure is

$$R^2 = 1 - \frac{SSE}{TSS} \quad (2.65)$$

null model

Here, SSE is the sum of the squared errors using the model, and TSS is the sum of squares without using the model (or with using the **null model**).

$$SSE = \sum (y_i - \hat{y})^2 \quad (2.66)$$

$$TSS = \sum (y_i - \bar{y})^2 \quad (2.67)$$

In this formula, \hat{y} is the predicted value of y for each value of x_i in the data according to the model; \bar{y} is the predicted value of y for each value of x_i in the data in the absence of the model (the mean of the dependent variable).

Thus, the SSE is a measure of how much variation remains in the model — the residual (unexplained) variation after applying the model. The TSS is a measure of the variation in the original data. It is called the residual variation after applying the “null model.”⁴

⁴The “null model” always refers to the model with no independent variable. Thus, it is the model with only the y-intercept (here). The concept of the “null model” is extremely important in statistics, because it allows us to determine how much the model is an improvement over the “lack of” model.

Note: The R^2 measure only tells us how much of the variation in the dependent variable is described by the model (as compared to how much was there originally). It tells us nothing beyond that.

For instance, an R^2 value of 0.04 tells us that the model explains only 4% of the variation in the dependent variable. There is a lot of variation left unexplained by the model. It does not mean that the model is “poor.” An R^2 value of 0.98 tells us that the model explains 98% of the variation in the dependent variable. It does not tell us that the model is “good.”

Note that the formula for R^2 (2.65) is equivalent to

$$R^2 = 1 - \frac{\frac{1}{n-1}SSE}{\frac{1}{n-1}TSS} \quad (2.68)$$

Note that the numerator of the fraction is an estimator for the unexplained *variance* — but using the wrong number of degrees of freedom ($n - 1$ instead of $n - p$). Thus, the R^2 measure is how much the model reduces the unexplained variance, when the variance is estimated using a *biased* estimator.

biased

2.4.2 \bar{R}^2 MEASURE Where R^2 measures how much the model reduces the unexplained variance, when that variance is estimated using a biased estimate of the total variance, \bar{R}^2 measures how much the model reduces the unexplained variance, when that variance is estimated using an *unbiased* estimate of the total variance. In other words, the adjusted R^2 measure uses the correct degrees of freedom to describe the proportional reduction in error.

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p}SSE}{\frac{1}{n-1}TSS} = 1 - \frac{(n-1)SSE}{(n-p)TSS} \quad (2.69)$$

Here, p is the number of parameters estimated in the model. For simple linear regression (one independent variable), $p = 2$, because we are calculating b_0 and b_1 from the data (to estimate β_0 and β_1). For the null model, $p = 1$, because we are only calculating \bar{y} from the data (to estimate β_0).

fight the power!!!

Usually, one reports the R^2 value and uses the \bar{R}^2 to help with model selection (select the model with the larger \bar{R}^2). This, I believe, needs to change. It is the *adjusted* R-squared value that better estimates the proportion reduction in error. This is because the adjusted R-squared measure uses unbiased estimators of the variances.

One strength of the R^2 measure is that it ranges from 0 to 1. The \bar{R}^2 measure *can* be less than 0. However, it is only less than 0 when the model describes very little of the variance.

2.4.3 OTHER PREs These are the two most frequently used PRE measures in linear regression. They are not the only ones, however. There is an entire class of PRE measures called “pseudo- R^2 ” measures. These are all genuine measures of how well the model helps reduce unexplained variation in the dependent variable. Their formulas tend to follow the structure of

$$\text{PRE} = 1 - \frac{\text{variation in the dependent variable with the model}}{\text{variation in the dependent variable without the model}}$$

In the future, you will be introduced to “pseudo- R^2 measures” for several different modeling schemes. Because these follow the same scheme as above, they have the same interpretation. They are measures of how well the model reduces the uncertainty in the dependent variable. The differences are in how that uncertainty is measured.

2.5: Conclusion

This chapter started with defining what we can mean by “best.” Because we decided to define “best” as “minimizing the sum of squared residuals,” we were able to obtain closed-form solutions for the estimates. From those equations, we were able to learn even more about our estimators — things not obvious from the definition.

That was the entire purpose of this chapter: to see that our results arise from applying mathematics to our selected definition of “best.” Had we chosen a different meaning, we may have arrived at different results.

In the next chapter, we will change how we represent the data. Using matrices allows us to generalize what we did in this chapter to more than one independent variable. It also allows us to draw some interesting results. For instance, the estimators are only independent in simple linear regression if $\bar{x} = 0$. Showing this requires us to find the covariance between b_0 and b_1 . This is much easier when working with matrices.

2.6: End-of-Chapter Materials

Here are the expected materials to supplement the chapter.

2.6.1 EXERCISES I left many things as exercises for you. Here they are. You should be able to prove any and all of them using your prior knowledge of mathematics (matrices and calculus).

1. Perform the second derivative test on b_0 and b_1 to show that these estimators are really minima.
2. Show that $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ is equivalent to $\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}$ and that $\sum_{i=1}^n (x_i - \bar{x})^2$ is equivalent to $\sum_{i=1}^n x_i^2 - n \bar{x}^2$.
3. Prove $b_1 = r_{xy} \frac{s_y}{s_x}$.
4. Prove that the slope estimator b_1 is zero if the y-values do not vary.
5. Using the scalar form, show that $\text{Cov}[b_0, b_1] = -\sigma^2 \frac{\bar{x}}{S_{xx}}$.

CHAPTER 3:

MATRICES AND LINEAR REGRESSION

OVERVIEW:

In the previous chapter, we were introduced to the classical linear model and estimating the parameters using ordinary linear regression. All of the work was done using a scalar representation of the data. When moving beyond simple linear regression, the estimators are more difficult to determine. The calculus remains almost as simple, but solving the system of equations becomes prohibitive.

The usual solution to solving complicated system of equations is to use a matrix representation of the problem. That is what this chapter does. Along the way, we discover more about linear models than we expected.



Chapter Contents

3	Matrices and Linear Regression	43
3.1	Matrix Representation	46
3.2	Predictions and the Hat Matrix.	57
3.3	The PRE Measures	63
3.4	Multicollinearity and Categorical Independent Variables	64
3.5	Conclusion	72
3.6	End-of-Chapter Materials	73



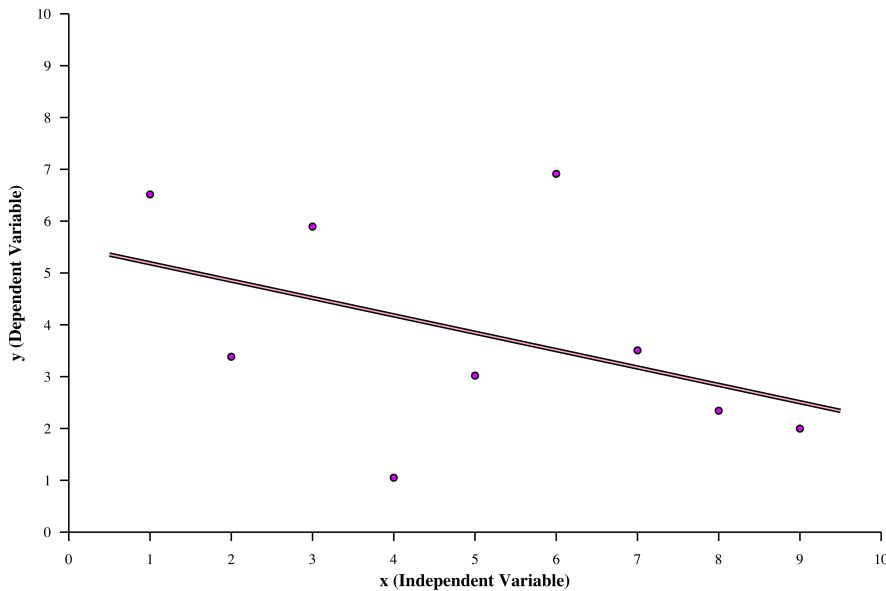


Figure 3.1: Sample data and a line of best fit for that data. Note that the slope of the line is negative. This indicates that increasing values of x **tend to** correspond to lower values of y . Regression searches for such trends.

As in the previous chapter, let x and y be numeric variables. The linear relationship between x and y can be summarized by a line that “best” fits the observed data. That is, we can (and will) summarize the relationship between x and y using a linear equation:

$$y = \beta_0 + \beta_1 x \quad (3.1)$$

The above holds in the case of simple linear regression (SLR). So, what do we do when there are more independent variables? Here is that representation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad (3.2)$$

Here, β_0 is the y -intercept (still). And, β_i is the effect of variable x_i on the dependent variable, assuming all the other variables remain constant.¹

¹This is called the *ceteris paribus* assumption. If the independent variables are independent of each other, then this requirement is met. However, there is frequently some correlation among the independent variables. Read on to see what to do about this.

Note: We say that the “line” given by equation 3.2 best fits the observed data. However, when dealing with two independent variables, it is not a line but a plane; with three, a space; with four, a hyperplane; etc. Clearly, meaningfully representing an entire four-variable model is quite difficult.

3.1: Matrix Representation

We learned a lot about our solution by exploring the scalar representation of the system of equations in the previous chapter. We may be able to gain some additional insights by exploring its matrix representation.



Warning: *It may be helpful to re-familiarize yourself with Appendix M at this point.*

And so, let us begin with our matrix model.

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (3.3)$$

In this model, \mathbf{Y} represents the response variable; \mathbf{X} , the predictor variable(s) prepended with a column of 1s; \mathbf{B} , the coefficient vector; and \mathbf{E} , the residuals. The dimensions are $n \times 1$ for \mathbf{Y} , $n \times p$ for \mathbf{X} , $p \times 1$ for \mathbf{B} , and $n \times 1$ for \mathbf{E} . Thus, in the case of simple linear regression, the matrices are

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathbf{E} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In this formulation, n is the sample size and $p = 2$ is the number of parameters that need to be estimated. Usually, this is one more than the number of independent variables, $k = 1$.

Note that \mathbf{X} is “the predictor variable(s) prepended with a column of 1s.” What does this mean? Also: Why are those 1s needed?

Let our independent variable be the same as in Example 2.1, $\{-2, 0, 0, 2\}$. The corresponding \mathbf{X} matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \quad (3.4)$$

Again, we want to minimize the **sum of squared errors**. Again, we will create the objective function Q , take its derivative with respect to the parameter vector, \mathbf{B} , and solve:

$$Q = \mathbf{E}'\mathbf{E} \quad (3.5)$$

$$= (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \quad (3.6)$$

$$= \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB} + \mathbf{B}'\mathbf{X}'\mathbf{XB} \quad (3.7)$$

Note that each of these terms is a 1×1 matrix, thus each is equal to its transpose. Using that on the third term and gathering the two like terms together gives our objective function.

$$Q = \mathbf{Y}'\mathbf{Y} - 2\mathbf{B}'\mathbf{X}'\mathbf{Y} + \mathbf{B}'\mathbf{X}'\mathbf{XB} \quad (3.8)$$

Now, taking the derivative with respect to \mathbf{B} gives

$$\frac{d}{d\mathbf{B}}Q = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{XB} \quad (3.9)$$

$$\mathbf{0} \stackrel{\text{set}}{=} -\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{Xb} \quad (3.10)$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Xb} \quad (3.11)$$

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \mathbf{b} \quad (3.12)$$

This formula is so important that I will repeat it here:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.13)$$

Note the switch between \mathbf{B} and \mathbf{b} . The former concerns the population. It is a population parameter that we are trying to estimate.

$$\mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad (3.14)$$

estimator

The latter concerns the sample. It is the estimator we are using to estimate the population parameter.

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \quad (3.15)$$

With this, the equation for our OLS regression line (plane, space, hyperplane, etc.) is

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} \quad (3.16)$$

3.1.1 REQUIREMENT In performing these calculations, we made one assumption: $(\mathbf{X}'\mathbf{X})^{-1}$ exists. If it does not exist, then the last step in the process cannot be done. So, the first question to ask is:

Question

When does $(\mathbf{X}'\mathbf{X})^{-1}$ not exist?

Ans. It does not exist when $\det \mathbf{X}'\mathbf{X} = 0$.

Question

So, when does $\det \mathbf{X}'\mathbf{X} = 0$?

From linear algebra (and Appendix M) we know that this determinant is zero when the \mathbf{X} matrix is not of full (column) rank; that is, when $\text{rank} \mathbf{X} \neq p$. This happens when one column of the \mathbf{X} matrix is a linear combination of the other columns. A statistician would say that there is redundant information in \mathbf{X} ; one variable can be determined by the others.

full rank

When in the realm of multiple regression (more than one independent variable), this happens when one variable is a linear combination of the others. This condition is called “multicollinearity” or “super multicollinearity.”

multicollinearity

When in the realm of simple linear regression (SLR), this happens when there is no variation in the x variable (it is a constant multiple of the columns of 1s).

3.1.2 ASSUMPTION/S Before we continue, as before, let us make the three assumptions about our residuals. These are just the same non-parametric assumptions we made back in Section 2.3, but in matrix form. The first is that they are realizations of a random variable (\mathbf{E} has a distribution). The second is that the expected value of the residuals is zero, $\mathbb{E}[\mathbf{E}] = \mathbf{0}$ (the measurements are not systematically biased). The third is that the residuals are independent and have a finite constant variance, $\mathbb{V}[\mathbf{E}] = \sigma^2 \mathbf{I}$, with $\sigma^2 < \infty$.

In other words, let us make this assumption:

$$\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (3.17)$$

Here, the variance is finite.

3.1.3 RESULTS Again, we have several results from this simple assumption.

Theorem 3.1.1

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\mathbf{B}$$

Proof. The proof of this proceeds from algebra.

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{X}\mathbf{B} + \mathbf{E}] \quad (3.18)$$

$$= \mathbb{E}[\mathbf{X}\mathbf{B}] + \mathbb{E}[\mathbf{E}] \quad (3.19)$$

One **pervasive requirement** is that the values of \mathbf{X} are not random variables. That is, the *researcher* selected those particular x values. Since this is true,

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X} \mathbb{E}[\mathbf{B}] + \mathbb{E}[\mathbf{E}] \quad (3.20)$$

Also, the values in the \mathbf{B} matrix are population parameters. They, too, are not random variables. In fact, the only random variable on the right-hand side of that matrix equation is the zero-mean \mathbf{E} matrix. Thus, we have

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\mathbf{B} + \mathbb{E}[\mathbf{E}] \quad (3.21)$$

$$= \mathbf{X}\mathbf{B} \quad (3.22)$$

□

Note: The requirement that the independent variables are not random allows us to easily calculate expected values, variances, and covariances. When designing experiments, this assumption is not problematic.

When working with observational data, this becomes troublesome in terms of the mathematics. It also becomes troublesome in terms of the variances of \mathbf{Y} ... and the \mathbf{b} . The confidence intervals for the estimates are wider than estimated here. Also, if the variability in the \mathbf{X} are not independent, even more difficulties arise.

If any of this interests you, please look into errors-in-variables models (among other topics).

Similarly, it is quite easy to prove $\mathbb{V}[\mathbf{Y} | \mathbf{X}\mathbf{B}] = \sigma^2\mathbf{I}$. I leave that to you as an exercise.

exercise



Another result is that the two estimators are unbiased (i.e., their expected values equal the population parameter):

Theorem 3.1.2

The OLS estimator \mathbf{b} is unbiased for \mathbf{B} .

Proof. An estimator is unbiased for the parameter if its expected value equals the parameter. Thus, we need only show $\mathbb{E}[\mathbf{b}] = \mathbf{B}$.

$$\mathbb{E}[\mathbf{b}] = \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}] \tag{3.23}$$

$$= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbb{E}[\mathbf{Y}] \tag{3.24}$$

$$= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\mathbf{B} \tag{3.25}$$

$$= \mathbf{B} \tag{3.26}$$

□

A third result is that the two estimators are not necessarily independent.

Theorem 3.1.3

The OLS estimators b_0 and b_1 are not necessarily independent.

Proof. To see this, we calculate the covariance matrix of \mathbf{b} :

$$\mathbb{V}[\mathbf{b}] = \mathbb{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \quad (3.27)$$

$$= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbb{V}[\mathbf{Y}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (3.28)$$

$$= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (3.29)$$

$$= \sigma^2((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (3.30)$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (3.31)$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (3.32)$$

If this matrix is diagonal, then the estimators are independent.

To see that the two estimators are linearly correlated (not independent), we just need to calculate the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. In general, this is rather difficult to do by hand. However, if we restrict ourselves to simple linear regression, that inverse is rather straight-forward because \mathbf{X} is

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad (3.33)$$

With that, we have

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad (3.34)$$

The determinant of $\mathbf{X}'\mathbf{X}$ is

$$\det \mathbf{X}'\mathbf{X} = n \sum_{i=1}^n x_i^2 - n^2 \bar{x}^2 = n S_{xx} \quad (3.35)$$

$$(3.36)$$

Thus, the inverse is

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n S_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \quad (3.37)$$

Finally, the covariance matrix is

$$\mathbb{V}[\mathbf{b}] = \frac{\sigma^2}{n S_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \quad (3.38)$$

From this matrix, we see that the covariance between b_0 and b_1 is

$$\text{Cov}[b_0, b_1] = -n\bar{x} \frac{\sigma^2}{n S_{xx}} = -\sigma^2 \frac{\bar{x}}{S_{xx}} \quad (3.39)$$

Thus, the OLS estimators are independent if and only if $\bar{x} = 0$. \square

Note: As an extension, note that the sign of the covariance is the opposite that of \bar{x} .

Finally, while this last results may seem just slightly interesting, it is the basis of the Working-Hotelling (1929) procedure, which we will see later in Section 4.4.

This last result also suggests why many disciplines tend to center their x -values (subtract off \bar{x}) before doing regression. It ensures that the two estimators *are* independent.

centering

Example 1

Let us revisit Example 2.1 and show how to use the matrix representation to answer the same problem.

Solution: The first step is to create the two matrices. The dependent variable matrix is

$$\mathbf{Y} = \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \quad (3.40)$$

The independent variable matrix, also called the “data matrix” and the “design matrix,” is

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \quad (3.41)$$

Where did the column of 1s come from in \mathbf{X} ? Remember that the matrix equation is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (3.42)$$

and that this is equivalent (in simple linear regression) to

$$y_i = \beta_0 \cdot 1 + \beta_1 x_i + \varepsilon_i \quad (3.43)$$

The 1s column in \mathbf{X} is the multiplier of the β_0 in the \mathbf{B} matrix. As long as you have a β_0 in your model, you need that column of 1s.

Now that we have the two matrices, we can calculate \mathbf{b} .

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.44)$$

$$= \left(\begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}' \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}' \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \quad (3.45)$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix} \quad (3.46)$$

$$\Rightarrow (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{32} \begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix} \quad (3.47)$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \quad (3.48)$$

$$= \begin{bmatrix} 4 \\ -8 \end{bmatrix} \quad (3.49)$$

Thus, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.50)$$

$$= \frac{1}{32} \begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ -8 \end{bmatrix} \quad (3.51)$$

$$\mathbf{b} = \frac{1}{32} \begin{bmatrix} 32 \\ -32 \end{bmatrix} \quad (3.52)$$

And finally,

$$\mathbf{b} := \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (3.53)$$

From all of this, we have $b_0 = 1$ and $b_1 = -1$. ♦

The conclusion is exactly the same, $\hat{y} = 1 - x$. The process is different. Here, this process is much easier for computers to perform, as they can do matrix multiplication (and inverting) with little problem. We have to spend a lot of extra effort to perform those operations. Here it is in R:

```
|| y = matrix( c(3,0,2,-1), ncol=1 )  
|| x = matrix( c(1,1,1,1,-2,0,0,2), ncol=2 )  
|| solve( t(x)**x ) ** t(x) ** y
```

Also, if we have more than one independent variable, we need to calculate the OLS estimator equations again; the ones in equation 2.20 only hold for one independent variable. Using matrices, however, formula 3.13 holds for any number of independent variables.

3.2: Predictions and the Hat Matrix

Beyond modeling the relationship, one may also want to estimate or predict values of \mathbf{Y} for a given value of \mathbf{X} . In matrix terms, this requires solving the equation $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$. But note the following:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} \quad (3.54)$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (3.55)$$

Thus:

$$\hat{\mathbf{Y}} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \quad (3.56)$$

Note that the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ “puts a hat” on the \mathbf{Y} matrix. As such, it is called the “hat matrix,” \mathbf{H} . Thus, we have simple matrix equations for the estimators and the residuals:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \quad (3.57)$$

$$\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad (3.58)$$

Why is this important? It shows that the predictions and the residuals are orthogonal (see definition on page 497).

hat matrix

perpendicular

Theorem 3.2.1

The matrices \mathbf{H} and $\mathbf{I} - \mathbf{H}$ are orthogonal.

Proof. To show orthogonality, we need to show that the inner product is zero:

$$\mathbf{H}'(\mathbf{I} - \mathbf{H}) = \mathbf{H}(\mathbf{I} - \mathbf{H}) \quad = \mathbf{H} - \mathbf{H}\mathbf{H} \quad (3.59)$$

$$= \mathbf{H} - \mathbf{H} \quad (3.60)$$

$$= \mathbf{0} \quad (3.61)$$

□

In the proof, we used the fact that the hat matrix is symmetric idempotent. The next theorem proves this to be the case.

idempotent

Theorem 3.2.2

The matrix \mathbf{H} is symmetric idempotent.

Proof. Let us start with showing \mathbf{H} is symmetric.

$$\mathbf{H}' = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (3.62)$$

Recall from page 499 in Appendix M that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. Thus

$$(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \mathbf{X}''((\mathbf{X}'\mathbf{X})^{-1})'\mathbf{X}' \quad (3.63)$$

$$= \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1})'\mathbf{X}' \quad (3.64)$$

I leave it as an exercise to show that $\mathbf{X}'\mathbf{X}$ is symmetric, and so is its inverse.

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.65)$$

$$= \mathbf{H} \quad (3.66)$$

Now, let us show that \mathbf{H} is idempotent.

$$\mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.67)$$

$$= \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.68)$$

$$= \mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.69)$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.70)$$

$$= \mathbf{H} \quad (3.71)$$

□

Since \mathbf{H} is symmetric and idempotent, it is an orthogonal projection matrix that projects \mathbf{Y} -space onto the smaller $\hat{\mathbf{Y}}$ -space (Appendix M.4). Because it is an orthogonal projection, $\hat{\mathbf{Y}}$ is as close to \mathbf{Y} as possible in its subspace. That is, the errors are minimized. Figure 3.2 illustrates this.

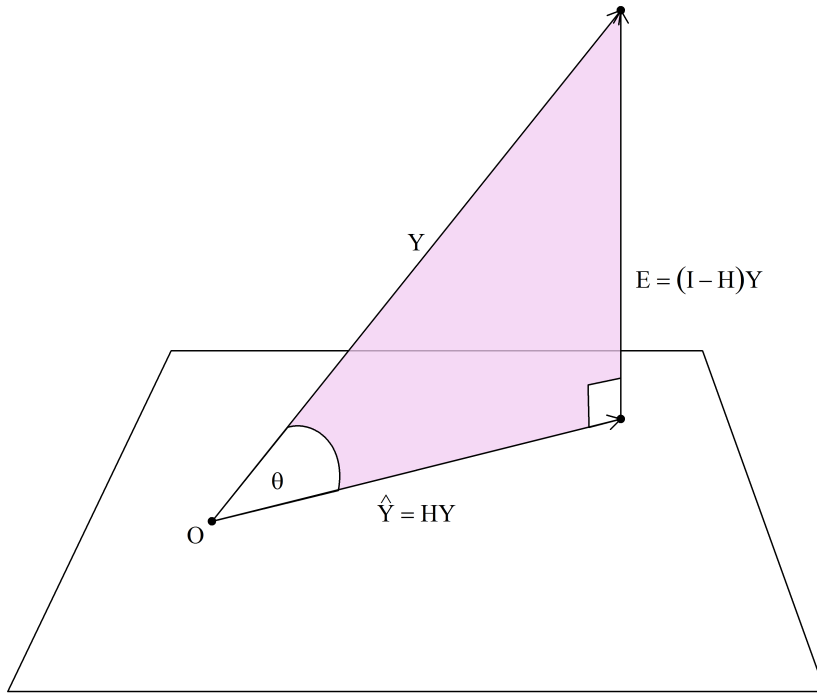


Figure 3.2: A schematic illustrating that \hat{Y} is as close to Y as possible, while remaining in its subspace (represented by the plane). In other words, the Y matrix exists in an n -dimensional space. The solution, \hat{Y} , is in a p -dimensional space, with $n > p$. Under the assumptions of ordinary least squares, the distance between Y and \hat{Y} (represented as the residuals, E) is as small as possible if you define “distance” in terms of the Euclidean distance, L_2 .

Theorem 3.2.3

The vectors \hat{Y} and E are orthogonal.

Proof. I leave this as an exercise. □

exercise

Since the predictions and residuals are orthogonal, we know the following is true by the Pythagorean Theorem:

$$Y'Y = \hat{Y}'\hat{Y} + E'E \tag{3.72}$$

Let us also prove this using matrices.

Theorem 3.2.4

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E}$$

Proof. Let us prove this without resorting to the Pythagorean Theorem. We know $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{E}$. Thus,

$$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \mathbf{E})'(\hat{\mathbf{Y}} + \mathbf{E}) \quad (3.73)$$

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \hat{\mathbf{Y}}'\mathbf{E} + \mathbf{E}'\hat{\mathbf{Y}} \quad (3.74)$$

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + (\mathbf{H}\mathbf{Y})'(\mathbf{I} - \mathbf{H})\mathbf{Y} + ((\mathbf{I} - \mathbf{H})\mathbf{Y})'\mathbf{H}\mathbf{Y} \quad (3.75)$$

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \mathbf{Y}'\mathbf{H}'(\mathbf{I} - \mathbf{H})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{Y} \quad (3.76)$$

Remember that \mathbf{H} and $\mathbf{I} - \mathbf{H}$ are symmetric. That gives us

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \mathbf{Y}'\mathbf{H}(\mathbf{I} - \mathbf{H})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{Y} \quad (3.77)$$

Finally, since $\mathbf{H}(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})\mathbf{H} = \mathbf{0}$, we have

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \mathbf{Y}'\mathbf{0}\mathbf{Y} + \mathbf{Y}'\mathbf{0}\mathbf{Y} \quad (3.78)$$

and

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} \quad (3.79)$$

□

This will come in handy when we add probability distributions to our mathematics, thus creating statistics.

By the way, we also can show that the residuals and predicted values are uncorrelated by showing their covariance is zero.

Theorem 3.2.5

$$\text{Cov}[\hat{\mathbf{Y}}, \mathbf{E}] = 0.$$

Proof. I will only give the first step to this proof. The rest will be up to you to figure out.

$$\text{Cov}[\hat{\mathbf{Y}}, \mathbf{E}] = \text{Cov}[\mathbf{HY}, (\mathbf{I} - \mathbf{H})\mathbf{Y}]$$

So, where to go from this? □

By the way, this result should not be surprising given that the prediction and residual vectors are orthogonal.

3.2.1 CONSEQUENCES In this section, we started with the matrix equation $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ and obtained the OLS estimator of \mathbf{B} . With that solution (and the requirement that \mathbf{X} be full column rank), we have another result.

Theorem 3.2.6

$$\mathbf{X}'\mathbf{E} = \mathbf{0}$$

Proof. Again, I will just start you off with this proof. Completing it is up to you.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$
□

mathematics

What does this result mean? Recall that $\mathbf{X}'\mathbf{E}$ is a $p \times 1$ matrix. The first column of \mathbf{X} is a column of 1s. Thus, the first element of $\mathbf{X}'\mathbf{E}$ is just the sum of the residuals. That means the residuals must sum to 0 when we use the OLS estimator.

The other elements in the $\mathbf{X}'\mathbf{E}$ matrix consist of the sum of the residuals times the values of each independent variable. This means that, under OLS, the residuals are necessarily linearly independent of each of the independent variables. It is a result of the mathematics used.

To see this in simple linear regression:

$$\mathbf{X}'\mathbf{E} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} \tag{3.80}$$

$$= \begin{bmatrix} \sum e_i \\ \sum x_i e_i \end{bmatrix} \tag{3.81}$$

This matrix is $\mathbf{0}$ only when all of its elements are also 0. Thus, we have $\sum e_i = 0$; the sum of the residuals in OLS is *mathematically* guaranteed to be zero.

We also have $\sum x_i e_i = 0$, which is equivalent to $\sum x_i e_i - n\bar{x}\bar{e}$ because $\bar{e} = 0$ and thus to $(n-1)\text{Cov}[x, e]$. This covariance is zero if x and e are linearly independent. This means that the residuals arising from OLS estimation are *linearly* uncorrelated with the predictor variables.

Note: Again, these are mathematical results from applying ordinary least squares. They are guaranteed simply because of the estimation method we selected. Had we chosen a different definition of “best fit,” then this section may not hold.

Everything follows from our chosen definition of “best fit.”

3.3: The PRE Measures

Now that we have reality (\mathbf{Y}) and our errors (\mathbf{E}), as pictured in Figure 3.2, we can create a measure of how well the model summarizes (fits) the data. In fact, we will create two of them! Both are “proportional reduction in error” measures; that is, they both are measures of how well the model reduces the unexplained variation in the dependent variable. The first is the venerable R^2 (“R-squared”) measure. The second is the \bar{R}^2 (“adjusted R-squared”) measure.

PRE

3.3.1 OTHER PREs Section 2.4 covered the two most frequently used PRE measures. They are not the only ones, however. In fact, we could use Figure 3.2 to create yet another PRE, this one based on the right triangle. Note that \mathbf{Y} is the target we are trying to describe and $\hat{\mathbf{Y}}$ is where we landed. Thus, a PRE measure could be the angle $\theta = \angle \mathbf{Y}\mathbf{O}\hat{\mathbf{Y}}$. This θ ranges between 0 and 90° , with 0 being an optimal fit and 90° being the worst fit.

PRE

As this measure is not intuitive as a measure of fit by itself (larger is not better), we can simply take its cosine and use $\cos \theta$ as our PRE. This value ranges between 0 and 1, with a 1 being the best fit ($\cos 0$) and a 0 being the worst ($\cos \pi/2$).

Question

So, how would you measure θ ?

3.4: Multicollinearity and Categorical Independent Variables

So far, our independent variable was either numeric (Example 2.1) or dichotomous (Example 2.1). Let us now look at the interesting case of a discrete independent variable with three levels.

Example 2

His Majesty Rudolph II would like some input on his next five-year plan. The primary crop in Ruritania is corn. To help optimize the profits made by farmers, Rudolph wants to know if that crop should be changed to summer wheat or to soybeans.

To help him, let us model the relationship between farmer profit and crop in Ruritania.

Solution: Collecting the data is not as difficult as it may seem at first. All three crops are currently grown in Ruritania. All we had to do was obtain a list of all farms and their primary crop and randomly select records from that. Table 3.1 provides our data.

Crop	Profit per Acre
Wheat	722
Wheat	965
Wheat	940
Wheat	756
Corn	763
Corn	765
Corn	565
Corn	621
Soybean	566
Soybean	658
Soybean	540
Soybean	485

Table 3.1: The data collected from Ruritania for Example 3.4. Note that this is the raw data with a categorical independent variable.

Wheat	Corn	Soybeans	Profit per Acre
1	0	0	722
1	0	0	965
1	0	0	940
1	0	0	756
0	1	0	763
0	1	0	765
0	1	0	565
0	1	0	621
0	0	1	566
0	0	1	658
0	0	1	540
0	0	1	485

Table 3.2: Data to be used for Example 3.4. This table differs from Table 3.1 by taking the original *Crop* variable and replacing it with three indicator variables. This form allows us to more easily calculate the ordinary least squares estimators by hand.

Note that the response variable is numeric, and the predictor variable is categorical. How do we code that variable so that we can use the methods of this chapter (and this class)???

In Example 2.1, it was easy to change our dichotomous variable into a numeric variable by selecting one level as the base level and measuring the other level from there. In other words, one level was given the value 0 (absence) and the other was given the value 1 (presence).

base level

In this case, we have *three* levels in our independent variable. It does not seem to make sense to select one level to represent with 0 (absence), one to represent with 1 (presence), and one to represent with 2 (huh????).

One method that always works is to create a series of dichotomous indicator variables from the one nominal variable. Thus, since there are three levels here, we would create three new dichotomous variables: corn, soybeans, and wheat.

This change is presented in Table 3.2. Note that each of the three dichotomous variables is now numeric. Each value indicates absence (0) or presence (1) of that trait (crop). With this change, we can use the methods of this chapter to calculate the values of the OLS estimators β_0 , β_1 , β_2 , and $\beta_3 \dots$ or can we?

To see why I ended that paragraph in an evil and foreboding voice, let us work through this using matrices.

Remember the formula to calculate the OLS estimators: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$.
Here, \mathbf{Y} is

$$\mathbf{Y} = \begin{bmatrix} 722 \\ 965 \\ 940 \\ 756 \\ 763 \\ 765 \\ 565 \\ 621 \\ 566 \\ 658 \\ 540 \\ 485 \end{bmatrix} \quad (3.82)$$

The design matrix, \mathbf{X} is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad (3.83)$$

So far, so good!

design matrix

Note: At this point, can you see why this matrix is termed the “design” matrix? From it, one can deduce the experimental design that gave rise to the data.

Next, let us calculate $\mathbf{X}'\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}' \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad (3.84)$$

$$= \begin{bmatrix} 12 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix} \quad (3.85)$$

Nice! That is a rather interesting matrix. From it, you can pick out the sample size ($n = 12$) and the sample sizes in each of the three levels ($n_i = 4$ in the diagonals). The next step is to calculate the inverse of this matrix.

Fantastický!

At this point, it is *soooooo* much easier to use technology to perform this calculation. However, when you do, you will get a notification that the matrix is singular. This means two things.

singular

1. Its inverse does not exist.
2. One column is a linear combination of the others.

Notice that the first column is the sum of the other three columns. Thus, there is redundant information. Thus, the columns are not linearly independent. To prove this last point, use the coefficient vector $a = \{1, -1, -1, -1\}$ in the definition on page 487.

Note: From an information standpoint, if one column is a linear combination of the others, then that column is redundant. The model can be repeated without that information.

information

This is one of the very few places in statistics where throwing away information helps. It is rather ironic that it helps solely in terms of the mathematics.

So, what do we do? We drop one of the redundant columns. The one we drop determines how we interpret the results.

interpretation



Dropping the first column is appropriate. It leads to this design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad (3.86)$$

This leads to

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \quad (3.87)$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 3383 \\ 2714 \\ 2249 \end{bmatrix} \quad (3.88)$$

Finally, this leads to

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.89)$$

$$= \begin{bmatrix} 845.75 \\ 678.50 \\ 562.25 \end{bmatrix} \quad (3.90)$$

Thus, from this decision, we have that the average profit for wheat is 845.75; for corn, 678.50; and for soybeans, 562.25.

means model

This is called the “**means model**” because the returned values are the means in each group.



Dropping the second column is also appropriate (the second column corresponds to the wheat design). When doing so, the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad (3.91)$$

Feel free to work through the calculation to obtain these estimates:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.92)$$

$$= \begin{bmatrix} 845.75 \\ -167.25 \\ -283.50 \end{bmatrix} \quad (3.93)$$

The interpretation here is that the average profit for wheat (the base category/-dropped column) is 845.75. The effect of corn over wheat is -167.25 , and the effect of soybeans over wheat is -283.50 . In other words, the expected corn profit is -167.25 over the wheat profit, and the expected soybean profit is -283.50 over the wheat profit.

Note that we dropped the first data column. Thus, the first result is the expected value of the first variable and the other results are the effects of those levels *as compared to* the base category (wheat).

Because the estimate are the effects of the other levels as compared to the selected base level, this is called an “**effects model**.”

effects model



Dropping the third column is appropriate, as well. When doing so, the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad (3.94)$$

Feel free to work through the calculation to obtain these estimates:

$$\mathbf{b} = \begin{bmatrix} 678.50 \\ 167.25 \\ -116.25 \end{bmatrix} \quad (3.95)$$

This interpretation is similar to the previous. The mean of the base category (corn) is 678.50 (first number). The effect of wheat over corn is 167.25. The effect of soybean over corn is -116.25.

Corn is the base category because the third column corresponds to the corn design. Note that this is *also* an **effects model**. The estimates are the effects in relation to the base category. ♦

If you look at these three sets of results, you will see a lot of commonalities. The one chosen depends on what you are trying to say about the relationship between the crop and the profit. Here, is it also very easy to move between the means model and the effects model.

Note: We are only investigating expected values (averages) in this analysis. Should we also decide to include the uncertainties in our estimates (as we should), the two models are complementary. It is very difficult to move between the standard error in the means model and the standard error in the effects model. It is so much easier to have the computer perform that computation for you.

For the record, here is the code I used for fitting the means model:

```
|| x = matrix( c(1,0,0, 1,0,0, 1,0,0, 1,0,0,
```



```

      0,1,0, 0,1,0, 0,1,0, 0,1,0,
      0,0,1, 0,0,1, 0,0,1, 0,0,1 ),
      ncol=3, byrow=TRUE)
Y = matrix( c(722, 965, 940, 756, 763, 765,
              565, 621, 566, 658, 540, 485) )

solve(t(X)%*%X)
t(X)%*%Y
solve(t(X)%*%X) %*% t(X)%*%Y

```

Here is the code I used for the first effects model:

```

X = matrix( c(1,0,0, 1,0,0, 1,0,0, 1,0,0,
              1,1,0, 1,1,0, 1,1,0, 1,1,0,
              1,0,1, 1,0,1, 1,0,1, 1,0,1),
            ncol=3, byrow=TRUE)
Y = matrix( c(722, 965, 940, 756, 763, 765,
              565, 621, 566, 658, 540, 485) )

solve(t(X)%*%X)
t(X)%*%Y
solve(t(X)%*%X) %*% t(X)%*%Y

```

Note that the only change is in the line that defines the data matrix, **X**.

Finally, here is the code I used when dropping the third column.

```

X = matrix( c(1,1,0, 1,1,0, 1,1,0, 1,1,0,
              1,0,0, 1,0,0, 1,0,0, 1,0,0,
              1,0,1, 1,0,1, 1,0,1, 1,0,1),
            ncol=3, byrow=TRUE)
Y = matrix( c(722, 965, 940, 756, 763, 765,
              565, 621, 566, 658, 540, 485) )

solve(t(X)%*%X)
t(X)%*%Y
solve(t(X)%*%X) %*% t(X)%*%Y

```

Again, the only change is in the line that defines the data matrix, **X**.

3.5: Conclusion

This chapter continued using our definition of “best.” However, we moved beyond simple linear regression and into multiple regression, where there is more than just one independent variable and more than just one parameter we need to estimate.

From those equations, we were able to learn even more about our estimators — things not obvious from the definition. For instance, we saw that the SLR estimators are only independent if $\bar{x} = 0$. But, then, that was the entire purpose of this chapter: to see that our results arise from applying mathematics to our selected definition of “best.” Had we chosen a different meaning, we may have arrived at different results.

In the next chapter, we will see what we can learn by taking the next step and applying statistics to the models. While the mathematics tells us the expected value... it is statistics that gives us an insight into the population based on our little sample.

3.6: End-of-Chapter Materials

Here are the expected materials to supplement the chapter. Since there is R code in this chapter, I am including an explanation of several helpful R functions.

3.6.1 R FUNCTIONS In this chapter, we were introduced to a couple R functions that will be useful in the future. These are listed here.

MATHEMATICS:

%*% This multiplies two matrices in R. Thus, running the command `A%*%B` will return the matrix product \mathbf{AB} (Section M.3.2). Be careful: `A*B` returns the Hadamard product (Section M.3), which is rarely what is needed.

c() This combines the several scalar values into a single vector of values.

matrix() This function creates a matrix from the given vector. The first slot belongs to the values in the matrix. After that is the number of rows (or columns) and whether you are entering the number by rows or by columns.

solve(m) This calculates the usual inverse of the provided matrix \mathbf{m} (page 484).

t(m) This calculates the transpose of the provided matrix \mathbf{m} (Section M.4).

3.6.2 EXERCISES I left many things as exercises for you. Here they are. You should be able to prove any and all of them using your prior knowledge of mathematics (matrices and calculus).

1. Prove that $\mathbb{V}[\mathbf{Y} | \mathbf{XB}] = \sigma^2 \mathbf{I}$.
2. Let \mathbf{A} be any full column rank matrix. Prove that $\mathbf{A}'\mathbf{A}$ is symmetric. Prove that its inverse is symmetric.
3. Prove that the vectors $\hat{\mathbf{Y}}$ and \mathbf{E} are orthogonal.

CHAPTER 4:

IMPROVED! NOW WITH PROBABILITIES

OVERVIEW:

This chapter extends the mathematics from last chapter by adding a probability distribution to the residuals. This results in the independent variable having a probability distribution.

Please keep in mind that the independent variables are not random variables. The researcher specifically selects their values. Adhering to this paradigm allows us to more easily determine the resulting distributions. As such, this chapter continues this requirement.

Should we not adhere to this requirement, the results of this chapter will technically be wrong, but will be close if the independent variable is statistically independent of the dependent variable.

Chapter Contents

4	Improved! Now with Probabilities	75
4.1	Probability Distributions	79
4.2	Test Statistics and Hypothesis Testing	91
4.3	Confidence Intervals	94
4.4	The Working-Hotelling Bands	100
4.5	Conclusion	101
4.6	End-of-Chapter Materials	102

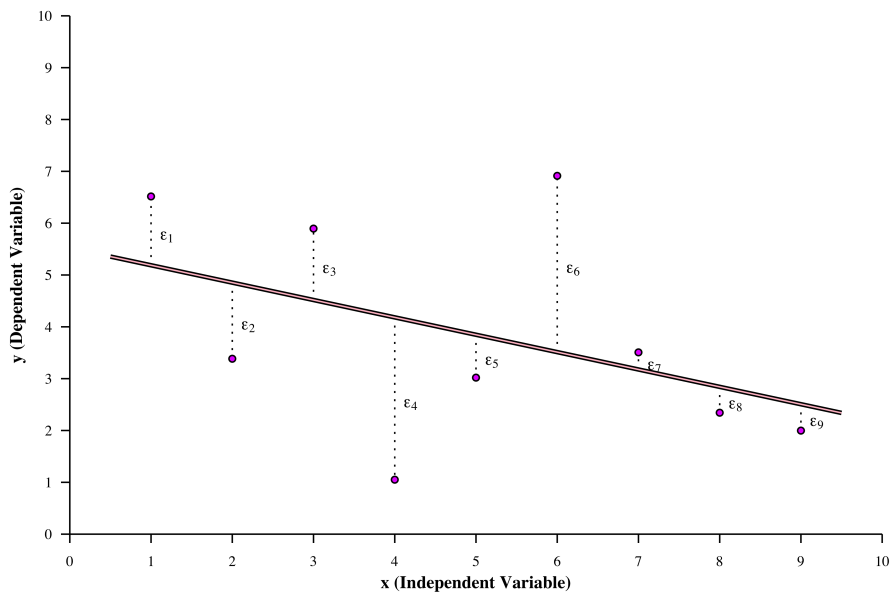


Figure 4.1: The basic scatter plot. This provides the observed values of the data as well as the line of best fit according to the Ordinary Least Squares method. The residuals are also indicated, with the values represented by dotted segments.



In the previous chapter, we explored the mathematical consequences of our choice of definition of “best.” In this chapter, we will acknowledge that the residuals are observations from a random variable, specify its distribution, and see where that takes us.

And so, let us return to our scalar model for our data (Figure 4.1):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (4.1)$$

and see what we can learn if we make the assumption that the ε_i are generated from a Normal distribution.

Specifically, in conjunction with our previous assumptions, let us assume:

$$\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad (4.2)$$

That single probability statement actually contains four parts:

- Normality** • The residuals follow a Normal distribution. No matter the values of the other variables, the residuals follow a Normal distribution.
- functional form** • The expected value of ε_i is a constant 0. No matter the values of the other variables, the expected value of the residual is 0 at that point.
- homoskedasticity** • The variance of the ε_i is a constant σ^2 . No matter the values of the other variables, the variance of the residual is σ^2 at that point.
- iid** • The abbreviation “iid” on top of the distribution sign means “independent and identically distributed.” It indicates that the ε_i are independent of each other, and that the distribution of each is the same, $\mathcal{N}(0, \sigma^2)$.

On the right-hand side (RHS) of Equation 4.1, the ε_i is the only random variable. The β_0 and β_1 are population parameters we are trying to estimate. The x_i are values selected by the experimenter, so they are also not random variables. This last sentence is rather important for a lot of the calculations we make. The values of the independent variable are selected by the researcher, they are *not* realizations of a random variable.

non-stochastic

Since the only thing on the right hand side that is a random variable is the ε_i , then it is rather easy to determine the distribution of Y . And, with that, we are able to determine the distribution of almost all parameters we find important.

Note: The RHS of Equation 4.1 is actually in two parts. The ε_i part is the source of the randomness, it is the “**stochastic**” part. The rest has no randomness associated with it. It is called the “**systematic**” part:

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{systematic}} + \underbrace{\varepsilon_i}_{\text{stochastic}} \quad (4.3)$$

4.1: Probability Distributions

From this one assumption/requirement, and the math from the previous chapter, we have many consequences. This section provides the results regarding the distribution of our estimators. The next sections build on this.

Theorem 4.1.1

The distribution of Y , conditional on the value of x , is

$$Y | x \stackrel{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2) \quad (4.4)$$

Proof. We are given $Y = \beta_0 + \beta_1 x + \varepsilon$, with the only random variable on the RHS being ε . Since ε follows a Normal distribution, so too does Y (see Corollary S.37). Since the Normal distribution has two parameters, the mean and the variance, we need to find those two values:

The expected value of $Y | x$ is

$$\mathbb{E}[Y | x] = \mathbb{E}[\beta_0 + \beta_1 x + \varepsilon] \quad (4.5)$$

$$= \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1 x] + \mathbb{E}[\varepsilon] \quad (4.6)$$

$$= \beta_0 + \beta_1 x + 0 \quad (4.7)$$

$$= \beta_0 + \beta_1 x \quad (4.8)$$

The variance of Y , conditional on x , is

$$\mathbb{V}[Y | x] = \mathbb{V}[\beta_0 + \beta_1 x + \varepsilon] \quad (4.9)$$

$$= \mathbb{V}[\beta_0] + \mathbb{V}[\beta_1 x] + \mathbb{V}[\varepsilon] \quad (4.10)$$

$$= 0 + 0 + \mathbb{V}[\varepsilon] \quad (4.11)$$

$$= \sigma^2 \quad (4.12)$$

Thus, putting this together, we have

$$Y | x \stackrel{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2) \quad (4.13)$$

□

Note that the Y are only ‘independently distributed’ and not ‘independent and identically distributed.’ This is because the expected value of Y depends on the value of x . Since the Y do not all have the same (identical) distribution, they are only ‘independently distributed.’

As for the results of the theorem above, they may not be too interesting. However, as our estimators depend on the Y_i , so too do their distributions. And *that* is where the interest arises.

We see this in the next theorem.

Theorem 4.1.2

The distribution of b_1 is $\mathcal{N}\left(\beta_1, \sigma^2 \frac{1}{S_{xx}}\right)$.

Proof. Before we start, we need to note that b_1 can be written as a linear combination of the Y_i :

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.14)$$

exercise

I leave the proof of *this* as an exercise.

Now, since our b_1 is a linear combination of Y_i , and since the Y_i come from independent Normal distributions, we have that b_1 also follows a Normal distribution (see Corollary S.37).

Again, since the Normal distribution has two parameters, the mean and the variance, we need to find those two values, as we do next.

The expected value of b_1 is

$$\mathbb{E}[b_1] = \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad (4.15)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[Y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.16)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.17)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.18)$$

$$= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.19)$$

$$= \beta_0 \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.20)$$

$$= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.21)$$

$$= \beta_1 \quad (4.22)$$

In this sequence, note that (be able to prove that):

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad (4.23)$$

and that

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x}) x_i \quad (4.24)$$

Thus, we know $\mathbb{E}[b_1] = \beta_1$; that is, our estimator is unbiased.

The final step is to determine the variance of b_1 :

$$\mathbb{V}[b_1] = \mathbb{V}\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad (4.25)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \mathbb{V}[Y_i]}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \quad (4.26)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \sigma^2 \quad (4.27)$$

$$= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 \quad (4.28)$$

$$= \sigma^2 \frac{1}{S_{xx}} \quad (4.29)$$

Recall that since we will be coming across $\sum_{i=1}^n (x_i - \bar{x})^2$ many, many, many times in the future, we denote it by S_{xx} . So, putting all of these parts together gives us

$$b_1 \sim \mathcal{N}\left(\beta_1, \sigma^2 \frac{1}{S_{xx}}\right) \quad (4.30)$$

□

Theorem 4.1.3

The covariance between our b_1 estimator and \bar{Y} is 0.

Proof. I leave this as an exercise. □

Theorem 4.1.4

The distribution of b_0 is $\mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$.

Proof. Remember that our estimator is

$$b_0 = \bar{Y} - b_1 \bar{x} \quad (4.31)$$

Since we have previously shown $\text{Cov}[\bar{Y}, b_1] = 0$ (Theorem 4.1.3), the proof is straight forward.

First, we note that b_0 is a linear combination of the Y_i . Thus, it follows a Normal distribution. (Again, see Corollary S.37 for a proof of this.) Because the Normal distribution has two parameters, we must find formulas for each:

Expected value:

$$\mathbb{E}[b_0] = \mathbb{E}[\bar{Y} - b_1 \bar{x}] \quad (4.32)$$

$$= \mathbb{E}[\bar{Y}] - \mathbb{E}[b_1 \bar{x}] \quad (4.33)$$

$$= (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} \quad (4.34)$$

$$= \beta_0 \quad (4.35)$$

Variance:

$$\mathbb{V}[b_0] = \mathbb{V}[\bar{Y} - b_1 \bar{x}] \quad (4.36)$$

$$= \mathbb{V}[\bar{Y}] + \mathbb{V}[b_1 \bar{x}] + 2 \text{Cov}[\bar{Y}, b_1 \bar{x}] \quad (4.37)$$

$$= \mathbb{V}[\bar{Y}] + \mathbb{V}[b_1] \bar{x}^2 + 2 \text{Cov}[\bar{Y}, b_1] \bar{x} \quad (4.38)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} \bar{x}^2 + 0 \bar{x} \quad (4.39)$$

Factoring out the σ^2 gives us

$$\mathbb{V}[b_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad (4.40)$$

Finally, putting these three parts together gives us what we want:

$$b_0 \sim \mathcal{N} \left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right) \quad (4.41)$$

□

There is another parameter in our model that we may like to estimate. That is the variance of ε . The ordinary least squares estimator of σ^2 is called the mean square error. It is defined as

MSE

$$\text{MSE} = \frac{1}{n-p} \sum_{i=1}^n \varepsilon_i^2 \quad (4.42)$$

Here, p is the number of parameters in the regression. So far, we have dealt with estimating β_0 and β_1 . Thus, $p = 2$ in simple linear regression.

Theorem 4.1.5

The distribution of the mean square error, MSE, can be written as

$$\frac{(n-p) \text{MSE}}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.43)$$

Proof. The first thing to do is remind ourselves of the definition of a χ^2 random variable. From Definition S.23, we have that if $Z_i \sim \mathcal{N}(0, 1)$, then $\sum Z_i^2 \sim \chi_v^2$, where v is the number of those Z_i that are independent (the degrees of freedom).

With this definition, we just need to find a random variable with a Normal distribution and transform it into the proper form. To that end:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad (4.44)$$

$$\frac{\varepsilon_i}{\sigma} \sim \mathcal{N}(0, 1) \quad (4.45)$$

$$\frac{\varepsilon_i^2}{\sigma^2} \sim \chi_{v=1}^2 \quad (4.46)$$

$$\frac{\sum \varepsilon_i^2}{\sigma^2} \sim \chi_{v=n-p}^2 \quad (4.47)$$

$$\frac{(n-p) \frac{1}{n-p} \sum \varepsilon_i^2}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.48)$$

$$\frac{(n-p) \text{MSE}}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.49)$$

$$(4.50)$$

And this is what we were to prove.

As usual, knowing the distribution of a sample statistic like the MSE allows us to create confidence intervals and perform hypothesis testing about the variance of the residuals, σ^2 . \square

Note: With that said, the importance of the previous theorem lies more in how we can use it to obtain confidence intervals and test hypotheses about the OLS estimators of the intercept and slope parameters.

By the way, the reason that equation 4.47 has $n - p$ degrees of freedom is that there are only $n - p$ independent terms. The other p terms can be determined (to within a constant) from the $n - p$ terms.

Theorem 4.1.6

The distribution of Y for an observed value of x_i , which we will term \hat{Y}_i , is

$$\hat{Y}_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right) \quad (4.51)$$

Note: What does this actually mean?

If we repeat this experiment (of collecting a sample of size n) an infinite number of times and estimate \hat{Y}_i for each of those experiments using our formulas, then those many \hat{Y}_i would follow the specified distribution.

stochastic

Proof. Remember that $\hat{Y}_i = b_0 + b_1 x_i$ and that x is non-stochastic (it is not a random variable). With this, we have that \hat{Y}_i is a linear combination of Normally distributed random variables (b_0 and b_1). As such, the name of the distribution of \hat{Y}_i is “Normal.” What remains is to calculate the expected value and variance.

$$\mathbb{E}[\hat{Y}_i] = \mathbb{E}[b_0 + b_1 x_i] \quad (4.52)$$

$$= \mathbb{E}[b_0] + \mathbb{E}[b_1 x_i] \quad (4.53)$$

$$= \beta_0 + \beta_1 x_i \quad (4.54)$$

As expected, the estimator is unbiased.

What about the variance? That is a bit more difficult, because we must deal with the covariance between b_0 and b_1 .

$$\mathbb{V}[\hat{Y}_i] = \mathbb{V}[b_0 + b_1 x_i] \quad (4.55)$$

$$= \mathbb{V}[b_0] + \mathbb{V}[b_1 x_i] + 2 \text{Cov}[b_0, b_1 x_i] \quad (4.56)$$

$$= \mathbb{V}[b_0] + \mathbb{V}[b_1] x_i^2 + 2 \text{Cov}[b_0, b_1] x_i \quad (4.57)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) + \sigma^2 \left(\frac{1}{S_{xx}} \right) x_i^2 + 2 \frac{-\bar{x} \sigma^2}{S_{xx}} x_i \quad (4.58)$$

Factoring things out to make it look more simple gives

$$\mathbb{V}[\hat{Y}_i] = \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} (\bar{x}^2 + x_i^2 - 2\bar{x}x_i) \quad (4.59)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} (\bar{x} - x_i)^2 \quad (4.60)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_i)^2}{S_{xx}} \right) \quad (4.61)$$

And so, putting these three things together gives us our hoped-for result

$$\hat{Y}_i \sim \mathcal{N} \left(\beta_0 + \beta_1 x_i, \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right) \quad (4.62)$$

... as we expected.

□

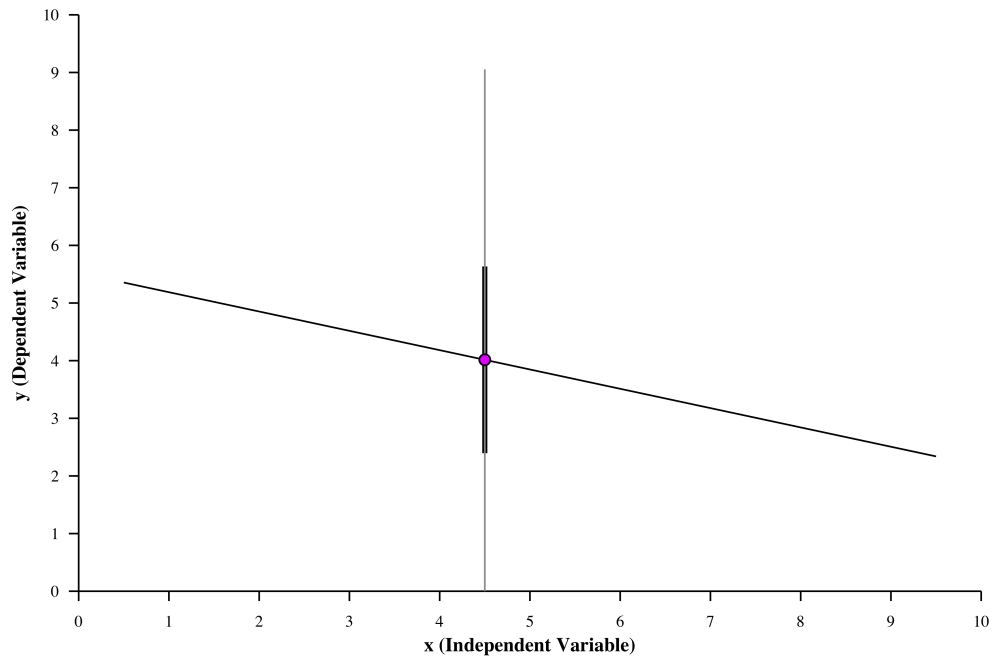


Figure 4.2: The basic scatter plot with the confidence and prediction intervals for $x = 4.5$ provided. Note that the prediction interval (thin line) is much wider than the confidence interval (thick line). This is because the prediction interval uncertainty includes both the uncertainty in the mean value (confidence interval) and the inherent variation in the residuals (σ^2).

Note: There are a couple of things interesting about this result. First, the uncertainty in \hat{Y}_i is a function of n , S_{xx} , and $\bar{x} - x_i$. Larger sample sizes (larger n) produce a more precise estimate.

Samples with larger values of S_{xx} also produce more precise estimates. To maximize S_{xx} , the researcher must have half of the x_i values at the minimum and half at the maximum.¹

Finally, the precision of the estimate also depends on how far that x value is from the center of gravity, (\bar{x}, \bar{y}) . Note that the uncertainty in \hat{Y}_i when $x = \bar{x}$ only comes from the uncertainty in the value of \bar{Y} . Convince yourself that this makes sense (non-mathematically).

¹Unfortunately, the drawback to doing this is that one is not able to detect a curvature in the expected values of Y . Thus, we again see that there is a trade off in statistics. The important part is to understand what you are trying to understand... and use your statistical understanding to understand it.

Theorem 4.1.7

The distribution of Y_{new} , a new observation, for a new value of x , is

$$Y_{new} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{new}, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}\right)\right) \quad (4.63)$$

Note: Before we begin this proof, remember that

$$Y_{new} = b_0 + b_1 x_{new} + \varepsilon = \hat{Y}_{new} + \varepsilon \quad (4.64)$$

Since we are estimating a new *observation* (as opposed to just an expected value), we need to include ε in our calculations. This is subtle and very important. It emphasizes the importance of ε .

subtle

Also, before we start the proof on the next page, compare and contrast this distribution with the distribution of \hat{Y}_i . What is the difference? Where does that difference come from?

Proof. And now for the expected proof. See that Y_{new} is a linear combination of Normally distributed random variables (b_0 , b_1 , and ε). Thus, Y_{new} follows a Normal distribution. All that remains is to calculate its expected value and its variance. To do so, we rely on the previous theorem.

$$\mathbb{E}[Y_{new}] = \mathbb{E}[b_0 + b_1 x_{new} + \varepsilon] \quad (4.65)$$

$$= \mathbb{E}[b_0 + b_1 x_{new}] + \mathbb{E}[\varepsilon] \quad (4.66)$$

$$= \mathbb{E}[b_0] + \mathbb{E}[b_1] x_{new} + \mathbb{E}[\varepsilon] \quad (4.67)$$

$$= \beta_0 + \beta_1 x_{new} + 0 \quad (4.68)$$

$$= \beta_0 + \beta_1 x_{new} \quad (4.69)$$

Next, for the variance:

$$\mathbb{V}[Y_{new}] = \mathbb{V}[b_0 + b_1 x_{new} + \varepsilon] \quad (4.70)$$

$$= \mathbb{V}[\hat{Y} + \varepsilon] \quad (4.71)$$

$$= \mathbb{V}[\hat{Y}] + \mathbb{V}[\varepsilon] + 2 \text{Cov}[\hat{Y}, \varepsilon] \quad (4.72)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_{new})^2}{S_{xx}} \right) + \sigma^2 + 0 \quad (4.73)$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{x} - x_{new})^2}{S_{xx}} \right) \quad (4.74)$$

Putting these parts together gives us the distribution of a new observation (a prediction):

$$Y_{new} \sim \mathcal{N} \left(\beta_0 + \beta_1 x_{new}, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}} \right) \right) \quad (4.75)$$

□

Note that the only difference in the uncertainties between Y_{new} and \hat{Y} is an additional term of σ^2 due to the inclusion of the residuals. Thus, all of the things that affect the variance of \hat{Y} also affect the variance of Y_{new} , and in the same way.

Note: Also note that the uncertainty in an observation is higher than the uncertainty in the expected value (see Figure 4.2).

observation

The important difference between this theorem and the previous is that this theorem models a new observation, while the previous models the expected value of an observation. The difference is important.

4.2: Test Statistics and Hypothesis Testing

The previous section provided the distribution of several important estimators. With those distributions, and our knowledge of probability distributions, we can test individual hypotheses. For this section, we rely heavily on the definition of Student's t distribution given as Definition S.24.

If we let $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_\nu$, with Z and V independent, then

$$T = \frac{Z}{\sqrt{V/\nu}} \quad (4.76)$$

follows a Student's t distribution with ν degrees of freedom.

You have, most likely, come across this ratio in your elementary statistics course when you were investigating hypotheses about a single population mean, given that the data came from a Normal distribution.

Theorem 4.2.1

The quantity

$$T = \frac{b_1 - \beta_1}{\sqrt{\text{MSE}/S_{xx}}} \quad (4.77)$$

follows a Student's t distribution with $n - p$ degrees of freedom.

Proof. To prove this statement, one must show that it can be written in the form of Equation 4.76. First, let us look at the numerator.

$$b_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{xx}) \quad (4.78)$$

$$\Rightarrow \frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0, 1) \quad (4.79)$$

Now for the denominator we use a previous theorem (Theorem 4.1.5):

$$\frac{(n-p) \text{MSE}}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.80)$$

Next, we put these together

$$T = \frac{b_1 - \beta_1}{\sqrt{\text{MSE}/S_{xx}}} \quad (4.81)$$

$$= \frac{\frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{\text{MSE}}{\sigma^2}}} \quad (4.82)$$

$$= \frac{\frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{(n-p) \text{MSE}}{\sigma^2}/(n-p)}} \quad (4.83)$$

Note that the numerator of Equation 4.83 follows a standard Normal distribution, while the denominator is the square-root of a chi-square distribution divided by its degrees of freedom. Thus, by Definition S.24, the quantity T follows a Student's t distribution with $n - p$ degrees of freedom. \square

Note: This result is important for two reasons. First, it allows us to test hypotheses regarding the β_1 parameter. Second, this result allows us to calculate confidence intervals for β_1 (see Section 4.3). This parameter is usually of most interest to researchers as it provides “the *effect* of the independent variable on the dependent variable.”

Since we know the distribution of this ratio, we can calculate p-values for any hypothesis about β_1 using the same rules as from your elementary statistics course (see Table 4.1).

Technically, we do need to show that b_1 and MSE are independent. If they are not, then Theorem 4.2.1 is not valid. For the proof, you will want to investigate Cochran's Theorem and its uses (Bapat 2000, Cochran 1934).

$H_0 : \beta_1 = \beta_{10}$	$H_A : \beta_1 \neq \beta_{10}$	p-value = $\mathbb{P} [t \leq - T] \times 2$
$H_0 : \beta_1 \leq \beta_{10}$	$H_A : \beta_1 > \beta_{10}$	p-value = $\mathbb{P} [t \geq T]$
$H_0 : \beta_1 \geq \beta_{10}$	$H_A : \beta_1 < \beta_{10}$	p-value = $\mathbb{P} [t \leq T]$

Table 4.1: Table of how to calculate p-values given the null and alternative hypotheses.

Theorem 4.2.2

The ratio

$$T = \frac{b_0 - \beta_0}{\sqrt{\text{MSE} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \quad (4.84)$$

follows a Student's t distribution with $n - p$ degrees of freedom.

Proof. I leave this as an exercise. □

This theorem allows us to easily prove the next.

Theorem 4.2.3

The ratio

$$T = \frac{\hat{y} - \hat{y}_0}{\sqrt{\text{MSE} \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}} \quad (4.85)$$

follows a Student's t distribution with $n - p$ degrees of freedom.

Proof. I leave this as an exercise. □

That sure is a lot of exercise.

How are those abs doing? Sore yet?

abs

4.3: Confidence Intervals

dual

In the previous section, we examined hypothesis testing. This required that we created a test statistic and determined its distribution. One can think of confidence intervals as the dual of test statistics. Test statistics are functions of an unknown population parameter and have a distribution. Confidence intervals are for that unknown population parameter, where a probability is known (assumed). Once a person has the test statistic and its definition, the confidence interval can be determined by inverting the test statistic function (solve for the parameter).

From your elementary statistic course, you knew that the distribution of $T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ followed a Student's t distribution with $n - 1$ degrees of freedom. Solving the formula for the parameter of interest, μ , gives

$$\mu = \bar{x} - T \frac{s}{\sqrt{n}} \tag{4.86}$$

The interpretation of T here is that it contains the values (quantiles) that correspond to the confidence level claimed (Figure 4.3). For instance, if you desire a 95% confidence interval for a sample of size 10, the central T values are ± 2.262 because the probability $\mathbb{P}[-2.262 < t < 2.262] = 0.95$.

Thus, the interpretation of μ in Equation 4.86 is that it contains the values that correspond to the endpoints of the confidence level claimed for the distribution of the right-hand side of the formula.

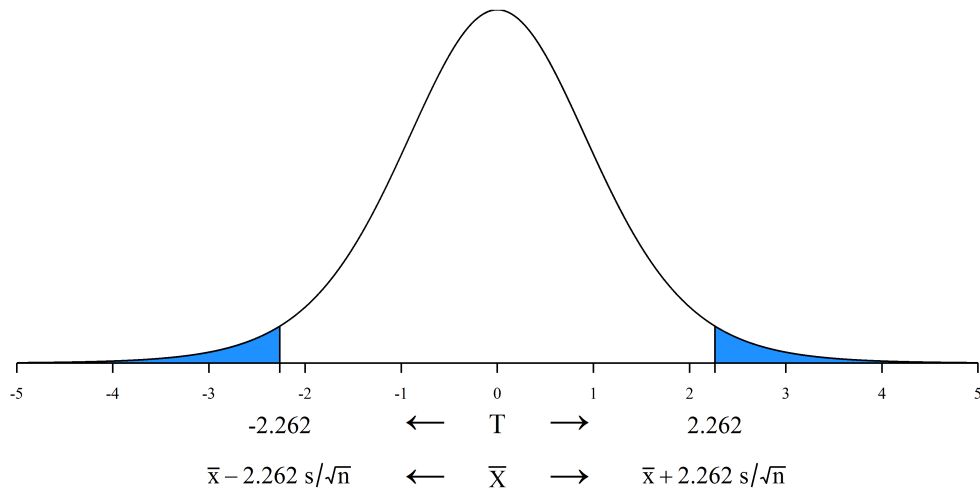


Figure 4.3: An illustration of a confidence interval seen from the standpoint of T or from \bar{X} . The unshaded area constitutes 95% of the area under the curve. Thus, the vertical segments delimit the endpoints of a central 95% confidence interval.

This interpretation holds for *all* confidence intervals.

With this discussion, it is rather straight-forward to calculate the endpoints of confidence intervals for all of the population parameters we have explored thus far. When the distribution of the test statistic is unimodal and symmetric, the central confidence interval is *also* the narrowest. This may be important if the researcher desires the most precise estimate of the population parameter.

Theorem 4.3.1

The endpoints of a central $(1 - \alpha) \times 100\%$ confidence for β_1 are defined by

$$b_1 \pm t_{\alpha/2, n-p} \sqrt{\text{MSE}/S_{xx}} \quad (4.87)$$

Proof. From Theorem 4.2.1, we know $T = \frac{b_1 - \beta_1}{\sqrt{\text{MSE}/S_{xx}}}$ follows a t distribution with $n - p$ degrees of freedom. Solving this for β_1 gives

$$\beta_1 = b_1 - T \sqrt{\text{MSE}/S_{xx}} \quad (4.88)$$

Because the distribution of T is symmetric unimodal, the endpoints of the minimum-width interval for T correspond to the two quantiles $t_{\alpha/2, n-p}$ and $t_{1-\alpha/2, n-p}$. These two endpoints are equivalent to $\pm t_{\alpha/2, n-p}$.

As such, the endpoints of a minimum-length $(1 - \alpha) \times 100\%$ confidence for β_1 are defined by $b_1 \pm t_{\alpha/2, n-p} \sqrt{\text{MSE}/S_{xx}}$. \square

This is a typical result when dealing with the Student's t distribution.

There is absolutely no reason we *need* a minimum-width confidence interval. It is, however, useful in maximizing the precision of the estimate.

When the distribution of the test statistic is unimodal symmetric, the central interval and the minimum-width interval are identical. When the distribution is *not* symmetric, they are not. The following illustrates this.

Theorem 4.3.2

The endpoints of a central $(1 - \alpha)100\%$ confidence for σ^2 are defined by

$$\frac{(n - p) \text{MSE}}{\chi^2_{1-\alpha/2, n-p}} \quad \text{and} \quad \frac{(n - p) \text{MSE}}{\chi^2_{\alpha/2, n-p}} \quad (4.89)$$

minimum-width

efficiency

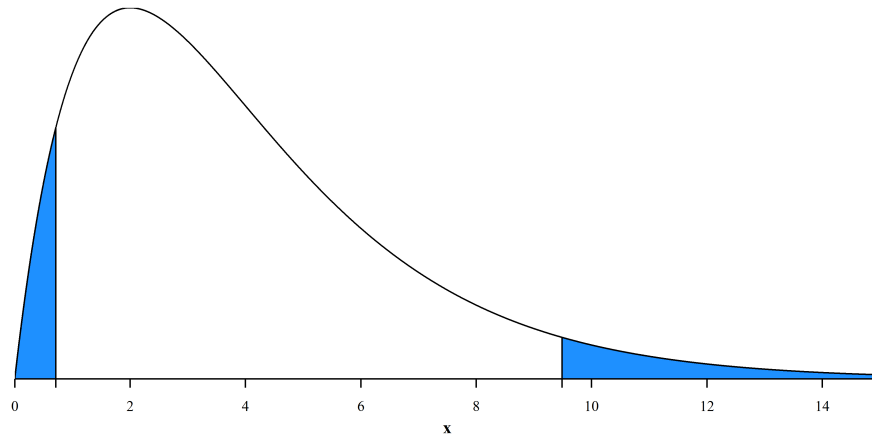


Figure 4.4: A plot of the chi-square distribution with 4 degrees of freedom. The unshaded area constitutes 90% of the area under the curve. Thus, the vertical segments delimit the endpoints of a central 90% confidence interval.

Proof. From Theorem 4.1.5, we know

$$\frac{(n-p) \text{MSE}}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.90)$$

Solving this for σ^2 gives

$$\sigma^2 = \frac{(n-p) \text{MSE}}{\chi_{n-p}^2} \quad (4.91)$$

Thus, a central $(1 - \alpha)100\%$ confidence interval (see Figure 4.4) is defined by the endpoints

$$\frac{(n-p) \text{MSE}}{\chi_{n-p,1-\alpha/2}^2} \quad \text{and} \quad \frac{(n-p) \text{MSE}}{\chi_{n-p,\alpha/2}^2} \quad (4.92)$$

□

Note: This is *not* the minimum-width interval. It is, however, the *usual* confidence interval provided. Calculating the minimum-width interval takes a little calculus that is beyond the scope of this section... and the typical coverage of this topic.

The minimum-width interval is illustrated in Figure 4.5. Note that the area in the shaded area to the right is not the same as that to the left. However, the two areas still account for 10% of the area, leaving 90% unshaded in the middle.

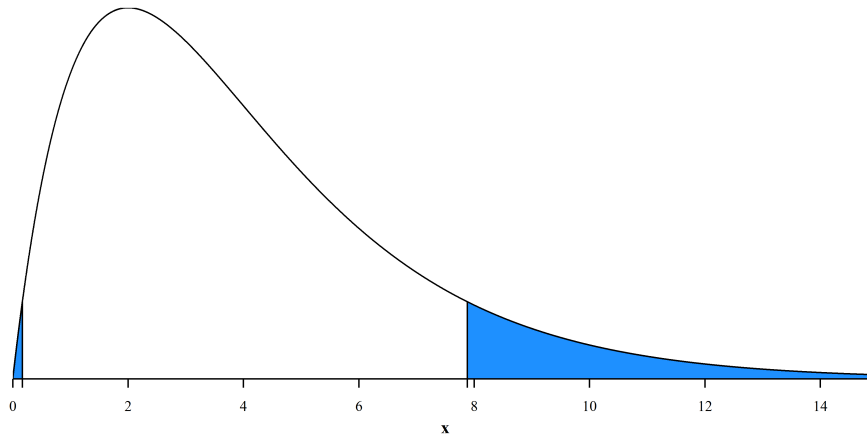


Figure 4.5: A plot of the chi-square distribution with 4 degrees of freedom. The shaded area constitutes 10% of the area under the curve. Thus, the vertical segments delimits the endpoints of a 90% confidence interval. This confidence interval, however, is the minimum-width interval.

The width of the central 90% confidence interval shown in Figure 4.4 is 8.777. This is wider than the width of the minimum-width confidence interval shown in Figure 4.5, which is 7.714. The minimum-width interval is 12% narrower than the central interval. That is an increase in estimator efficiency. It also requires some additional mathematics that we will skip.

However, as a teaser, notice that the value of the density function for each of the two endpoints is the same in the minimum-width interval. If the distribution is unimodal, then that observation will be true. That's enough of a hint. Feel free to explore this on your own. Calculus will serve you well here.

explore

Theorem 4.3.3

The endpoints of a central (and minimum width) confidence interval for β_0 are defined by

$$b_0 \pm t_{\alpha/2, n-p} \sqrt{\text{MSE} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \quad (4.93)$$

Proof. I leave this as an exercise. In fact, feel free to sketch the proof here. \square

Theorem 4.3.4

The endpoints of a confidence interval for \hat{y} are defined by

$$b_0 + b_1 x \pm t_{\alpha/2, n-p} \sqrt{\text{MSE} \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)} \quad (4.94)$$

Proof. I leave this as an exercise. In fact, feel free to sketch the proof here (without being sketchy). \square

Theorem 4.3.5

The endpoints of a prediction interval for y are defined by

$$b_0 + b_1x \pm t_{\alpha/2, n-p} \sqrt{\text{MSE} \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)} \quad (4.95)$$

Proof. I leave this as an exercise. In fact, feel free to sketch the proof here. \square

Note: This interval (Theorem 4.3.5) is termed a “prediction interval” because it is used to *predict a new observation* of y . It is not used to estimate the expected value of y — or trends in y . That would be the purpose of a confidence interval.

4.4: The Working-Hotelling Bands

One last confidence interval we may be interested in is a confidence interval for the **regression line**, itself.

Note that all of the confidence intervals in this chapter (except for σ^2) have been of the form

$$\text{point estimate} \pm K \times se \quad (4.96)$$

That is because they were confidence intervals for a measure of center. The Working-Hotelling (1929) confidence band for the regression line follows this format. It is

$$(b_0 + b_1x) \pm F(1 - \alpha, 2, n - 2) \times \sqrt{\text{MSE} \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)} \quad (4.97)$$

The proof is beyond the scope of this course.

With this being said, it *is* an interesting proof. The key is to focus on the joint distribution of b_0 and b_1 . This joint distribution is bivariate Normal. Thus, confidence intervals take the form of confidence *ellipses* with the same meaning and interpretation. However, as is common for confidence regions, the distribution of interest is the Chi-squared, instead of the Normal. Why? Answer: Think about the formula for an ellipse. Finally, the problem is transformed from the β_0 - β_1 plane to the x - y plane.

Believe it or not, the hardest part of the proof is the algebra.

So, where does the F distribution come from in the formula? The same place as the t distribution in the univariate case: the fact that we do not know the population variances involved.

Technically, Working and Hotelling only worked in the case of knowing the variances, which led to a Chi-square distribution in the formula. This is because the F distribution had not been invented (or discovered) yet. It was not until Snedecor in the 1940s that we were able to take that final step.

4.5: Conclusion

This chapter started with the mathematics of the previous chapter, the mathematics based on our definition of “best.” From that decision, we added a single assumption: $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

That assumption/requirement about the residuals gave us the entire chapter. Probability distributions for each of the estimators arose from the mathematics and the assumption. From those probability distributions, we created test statistics.

Having test statistics allows us to calculate p-values and confidence intervals for the parameters of interest. That is the flow of statistics. Once we have a distribution for a test statistic, we know everything we want to know for inferential statistics.

The difficulty comes in finding a test statistic with a known distribution. The assumption of Normality (and of iid) were key in allowing us to find those test statistics.

4.6: End-of-Chapter Materials

Here are the expected materials to supplement the chapter.

4.6.1 EXERCISES

1. Prove that $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x}) x_i$.
2. Prove that $b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$.
3. Prove that the covariance between our b_1 estimator and \bar{y} is 0.
4. Prove that the OLS estimators b_0 and MSE are independent.
5. Prove that the OLS estimators b_1 and MSE are independent.
6. Prove that the ratio $T = \frac{b_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$ follows a Student's t distribution with $n - p$ degrees of freedom.
7. Prove that the ratio $T = \frac{\hat{y} - \hat{y}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)}}$ follows a Student's t distribution with $n - p$ degrees of freedom.
8. Prove that the endpoints of a central confidence interval for β_0 are defined by $b_0 \pm t_{\alpha/2, n-p} \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$.
9. Prove that the endpoints of a confidence interval for \hat{y} are defined by $b_0 + b_1 x \pm t_{\alpha/2, n-p} \sqrt{MSE\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)}$.
10. Prove that the endpoints of a prediction interval for y are defined by $b_0 + b_1 x \pm t_{\alpha/2, n-p} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)}$.

4.6.2 THEORY READINGS

- R. B. Bapat (2000). *Linear Algebra and Linear Models* (Second ed.). New York: Springer.
- William G. Cochran (1934). “The distribution of quadratic forms in a normal system, with applications to the analysis of covariance.” *Mathematical Proceedings of the Cambridge Philosophical Society*. **30**(2): 178–191.
- Franklin A. Graybill and David C. Bowden (1968). “Linear Segment Confidence Bands for Simple Linear Models.” *Journal of the American Statistical Association*, 62(318): 403–408.
- Henry Scheffé (1959). *The Analysis of Variance*. New York: Wiley.
- Holbrook Working and Harold Hotelling (1929). “Applications of the Theory of Error to the Interpretation of Trends.” *Journal of the American Statistical Association*, 24(165A): 73–85.

The background of the slide is a photograph of a canal or river. The water is calm and reflects the surrounding greenery. On the right bank, there is a paved path lined with young trees, and further back, a road with a red curb. The sky is clear and blue.

CHAPTER 5:

DOOD! CHECK THE REQUIREMENTS

OVERVIEW:

There are several requirements for ordinary least squares regression to be applicable. In this chapter, we cover them, their tests, and their relative importance. Focus on the relationship between the assumption/requirement and the tests used. As these assumptions (requirements) are used for some other fitting methods, not just ordinary least squares, the lessons learned here are helpful in the future.

However, realize that each fitting method has its own set of assumptions/requirements. For instance, median regression (Chapter 9) requires only the constant expected residual value. Maximum likelihood for Poisson regression (Chapter 14) requires that, plus a specific relationship between the expected value and variance.

Chapter Contents

5.1	Normality	108
5.2	Constant Expected Value	122
5.3	Constant Variance	129
5.4	Multicollinearity	138
5.5	Conclusion	145
5.6	End-of-Chapter Materials	146



In Chapters 2 and 3, we created the ordinary least squares regression technique. The mathematics of the situation required $\det \mathbf{X}'\mathbf{X} \neq 0$. Because this corresponds to the requirement that the design matrix \mathbf{X} be of full column rank, the requirement is met when the independent variables are not linear combinations of each other.

Chapter 4 included the requirement

$$\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad (5.1)$$

This allowed us to move beyond the pure mathematics of Chapters 2 and 3 and begin making inferences about the population based on the sample.

The requirement that $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, equivalently $\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, has several parts to it: Normality, constant expected value (of zero), constant variance, and independence. This chapter takes these assumptions and explores them. Both graphical and numeric tests are presented. Throughout it all, we will rely heavily on R to generate the residuals, create the graphics, and perform the tests.

equivalently

Note: Some R functions require an additional package to be loaded. When such is the case, I indicate that in a special font. For example, the `runs.test` (Bradley 1968) is implemented in the `lawstat` package as the function `runs.test`. The `lmtest` package has the `bptest`, which performs the Breusch-Pagan test.

Some R functions are not included in any package and must be downloaded from the Internet each session. Such functions require you to run

```
|| source("http://rfs.kvasaheim.com/rfs.R")
```

at the start of the script. Once that line is run, you can then run several additional functions, like `overlay`, `hetero.test`, `shapiroTest`, and an advanced version of `runs.test`.

5.1: Normality

Let us tackle the requirement of Normality first. In Chapter 4, this assumption allowed us to exactly determine the distribution of the test statistics, which allowed us to exactly calculate p-values and confidence intervals. In this section, let us look at how to test your model that this assumption/requirement is met.

5.1.1 GRAPHICAL TESTS From your elementary statistics course, you were most likely introduced to a pair of graphical tests of Normality: Q-Q plots and histograms. In this section, we examine each when the residuals are generated from a Normal process and when they are not. To do this, let's do some experiments using R.

To begin, let us generate 100 residuals from a Normal process, with mean $\mu = 0$ and standard deviation $\sigma = 2$:

```
|| e = rnorm(100, m=0, s=2)
```

The `rnorm` function generates `n` random values from a Normal distribution with mean `m` and standard deviation `s`. Running this line only stores those 100 random values in the `e` variable.



Warning: Be aware that R, like most statistics programs, parameterizes the Normal distribution using the standard deviation instead of the variance.

With that one line, we have some “residuals” to play with that are generated “under the null hypothesis” that they are generated from a Normal distribution. This will allow us to better understand what the Normal distribution looks like.

Q-Q PLOT: So, let us look at how to generate a Normal-based quantile-quantile (Q-Q) plot using R.

```
|| qqnorm(e)
```

That's it. After running that one line, R creates the usual Q-Q plot for the Normal quantiles. It is a default graphic, so it does not look awesome, but it does get the point across to the statistician.

One shortcoming of the default Q-Q plot in R is that it does not provide the diagonal line. You can add it by also running the command

```
|| qqline(e)
```

quantile-quantile

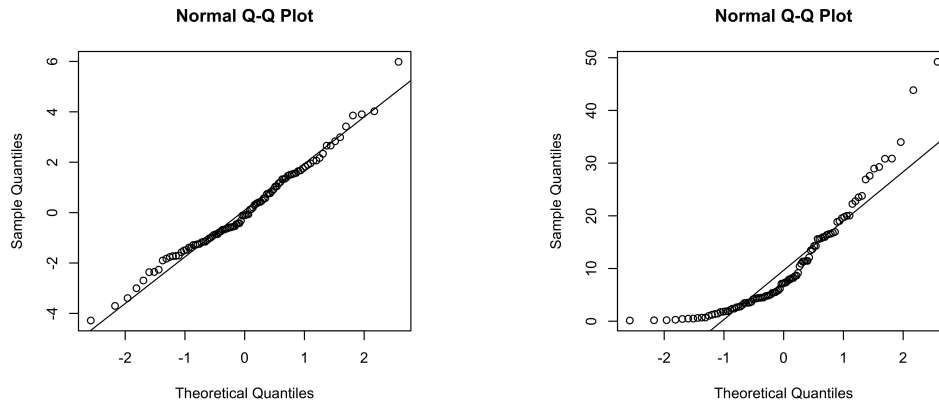


Figure 5.1: Default Normal quantile-quantile plots The left panel is for the randomly-generated Normal data. Note that the circles fall close to the diagonal target line. The right panel is for the randomly-generated Exponential data. Note that the circles do not tend to fall close to the diagonal target line. There is a distinct bow in them, which signifies the residuals have a right (positive) skew.

Figure 5.1, left panel, is the graphic produced by running these lines. Note that most of the circles cluster around the diagonal target line. Let us compare that graphic to a Q-Q plot from a non-Normal distribution. The non-Normal distribution will be the Exponential($\lambda = 0.10$) distribution.¹

```

||  enon = rexp(100, rate=1/10)
||  qqnorm(enon)
||  qqline(enon)

```

Compare the two graphics in Figure 5.1. Note that the shape of the Q-Q plot from the Normal residuals (left) is very different from the one generated from a skewed distribution (right). This particular shape indicates that the distribution of the residuals is positively skewed (right skewed).

HISTOGRAM: In the previous section, we examined quantile-quantile plots. We saw that a Q-Q plot with the dots aligning closely to the diagonal target line suggests Normality. In this section, we will use histograms to obtain a better view of the distribution of the data.

Again, let us generate Normally-distributed residuals.

```

||  e = rnorm(100, m=0, s=2)

```

¹This distribution is highly right-skewed. We will come back to the Exponential distribution several times to better understand how assumption violations affect our conclusions.

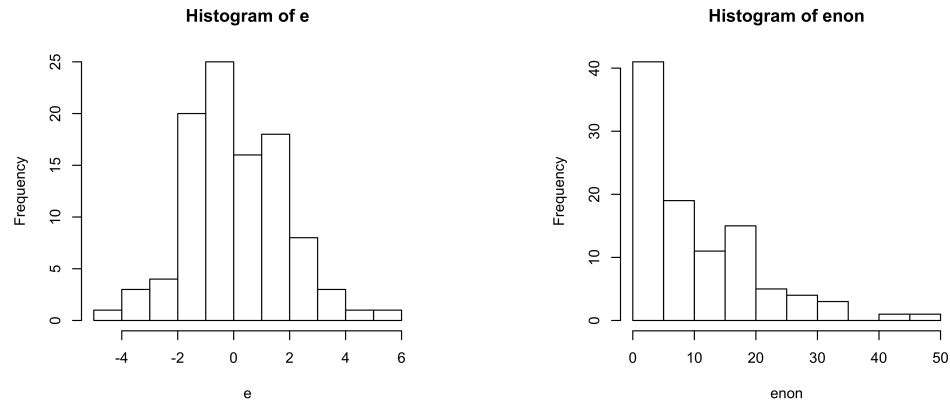


Figure 5.2: Default histograms for the randomly-generated residuals. In the left panel, note the basic bell shape to the histogram. Such a shape suggests that the residuals come from a Normal distribution. In the right panel, note the lack of a bell shape to the histogram. Such a shape suggests that the residuals do not come from a Normal distribution.

The function to create a basic histogram is just

```
|| hist(e)
```

or

```
|| overlay(e)
```

This basic histogram produced is provided as in the left panel of Figure 5.2. Note that it has the stereotypical “bell shape” to it, thus suggesting the data come from a Normal distribution. This is the shape you are seeking when using a histogram to explore the distribution of the residuals.

What does the histogram look like when the data come from a highly skewed distribution like the Exponential($\lambda = 0.10$) distribution above? See Figure 5.2, right panel.

Note the lack of bell shape in the left histogram of Figure 5.2 (and enhanced in Figure 5.3). In fact, one can easily see that the residuals are positively skewed. Recall that the direction of the skew is in the same direction as the long tail.

Note: I find using the histogram much easier than using the Q-Q plot. I can “see” how the residuals are distributed in the histogram. In the Q-Q plot, I have to interpret much more, remembering what the different shapes indicate.

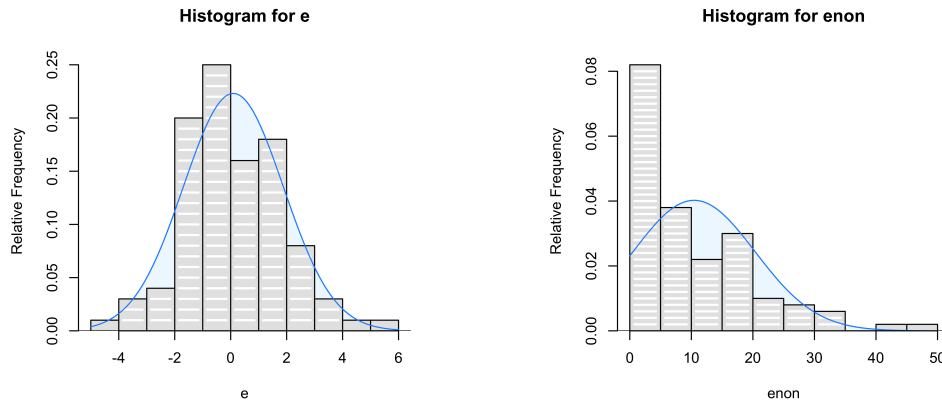


Figure 5.3: Enhanced histograms for the randomly-generated residuals. These are the same data as in Figure 5.2. The difference is that these graphics also overlay the Normal density function to aid in comparison.

Note: With that said, however, a Q-Q plot is much more useful than a histogram when there are few data points. So, if your sample size is small, you will want to use a Q-Q plot. Otherwise, a histogram may be best.

5.1.2 NUMERIC TESTS Because of the importance of the Normal distribution in Statistics, there are several Normality tests available, with more being created yearly. I mean, we gotta award those Ph.D. degrees for something, right?

PhDs? PhDd?

The reality is that it matters little which Normality test you use. As you will see in Section 5.1.3, the Normality assumption in OLS reflects the Normality assumption in t-tests and the like. The Central Limit Theorem ensures that a sufficiently large sample size makes the distribution of the *data* largely irrelevant; the distribution of sample sums is “close” to Normal.

CLT

So, for me, I default to using the Shapiro-Wilk test as my Normality test. In base R, this test is implemented as the function `shapiro.test`. In the `RFS` package (or `rfs` file), it is implemented as `shapiroTest`. I recommend the latter, because it adds additional functionality.

So, let us generate some Normally-distributed residuals and see how we can use the Shapiro-Wilk test.

Let us first set the random number seed. Why? Technically, there is no such thing as a truly random number when they are generated by a computer. This is because these pseudo-random numbers are just functions of a number called a **seed**.

prng

seed

Think of a seed as being the name of an entire list of “random-looking” numbers. Thus, if we use the same seed, then our “random numbers” will be the same.

To see this, run the following three lines:

```
|| set.seed(370)
|| e = rnorm(100, m=0, s=2)
|| head(e, 5)
```

The numbers you get are

```
|| -0.5566842 -1.7264618 1.3320962 -0.5392740 0.3477423
```

Setting the random number seed ensured that your values are mine. Even more interesting: if you rerun those three lines, you will get the same “random” numbers.

Question

In research, why would one *want* to set the random number seed?

Note: The `head` function returns the first six values in the variable, by default. A similar function is the `tail` function, which returns the *last* six values in the variable. I included the number 5 to have R only return five numbers.

To perform the Shapiro-Wilk test, just run the following line

```
|| shapiroTest(e)
```

The output you get will be

```
|| Shapiro-Wilk Normality test
|| data: e
|| W = 0.98554, p-value = 0.3472
```

The p-value is interpreted in the usual way: If the p-value is greater than α , then you fail to reject the null hypothesis. If the p-value is less than α , then you reject the null hypothesis. The null hypothesis in this case is that the data are generated from a Normal process (that the data come from a Normal distribution).²

Since the p-value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis. We do not have sufficient evidence that the data are non-Normal.

Note: We did *not* conclude that the data *do* come from a Normal distribution. We only concluded that there is no significant evidence to the contrary.

This point is very important. We first assume that the residuals are from a Normal distribution, then we test to see how reasonable that assumption is, given our data.

²The Normal distribution is identical to the Gaussian distribution. The only difference is the discipline. In much of statistics, it is known as the Normal distribution. However, when we get to generalized linear models (Chapter 11), this same distribution will be called the Gaussian distribution. In the Francophone world, this distribution is routinely called the Laplace-Gauss distribution.

golf or car?

To drive home this point, let us generate our data from a non-Normal distribution and perform the Shapiro-Wilk test:

```
|| set.seed(370)
|| e = rt(100, df=100)
|| shapiroTest(e)
```

By the way that we generated the data, we know that the residuals do not come from a Normal distribution; they come from a Student's t distribution with 100 degrees of freedom ($\nu = 100$).

Here is the output:

```
|| Shapiro-Wilk Normality test
|| data: y
|| W = 0.98979, p-value = 0.6474
```

Note that the p-value is greater than our α value of 0.05. As such, we again fail to reject the null hypothesis. There is not sufficient evidence that the residuals do not come from a Normal distribution.

However, we absolutely *know* they don't.



Warning: *The Lesson— Never accept the null hypothesis. The p-value depends on how reasonable the null hypothesis is, as well as how good the test is and how large the sample is.*

Note: Remember that when making decisions, there are really three usual decisions: Yes, No, and Maybe. In statistics, since we are just testing how reasonable the null hypothesis is, we only have two possible decisions: Reject and Don't Reject. These correspond to “the null hypothesis is wrong” and “the null hypothesis is right *or* we don't know.”

Let us now generate severely non-Normal residuals and use the Shapiro-Wilk test to test if they do come from a Normal distribution:

```
|| set.seed(370)
|| e = rexp(100, rate=1/10)
|| shapiroTest(e)
```

Here are the results:

```
|| Shapiro-Wilk Normality test
|| data: y
|| W = 0.68002, p-value = 1.891e-13
```

Note that the p-value of $1.891 \times 10^{-13} = 0.000\ 000\ 000\ 000\ 189\ 1$ is much less than $\alpha = 0.05$. That strongly indicates that the residuals were not generated from a Normal process.

Note: Remember the Central Limit Theorem (Section S.6.4) and the effect of sample size on the Normality of the sample sums. In the next section, we will explore the effects of non-Normality on our estimators and their distributions.

Why should I invoke the holy CLT in this context? Recall that the CLT speaks to the distribution of sums (and means) of independent random variables. As the sample size increases, the distribution of that sum approaches Normality (Section S.6.4).

Look at the formulas for b_0 and b_1 . They both include the *sum* of y_i . Thus, it is the distribution of $\sum y_i$ that we really care about. If the y_i are from a Normal distribution, then this requirement is met. **However**, if the sample size is large enough, this condition is *also* met, as long as the data are from a distribution with a finite variance. In other words, the CLT rules the world.

**the holy
hand grenade
of Antioch**

peace and love

Question

How “large” is “large enough”?

5.1.3 EXPLORATION OF THE EFFECTS OF NON-NORMALITY Mathematically speaking, if the Normality assumption does not hold, then nothing in Chapter 4 is absolutely true for the model. However, the Central Limit Theorem tells us that a large sample size mitigates the effects of non-Normality in the residuals.

Let us explore that here...

AN APPROPRIATE TEST?: Recall from your previous statistics course two main types of errors: Type I and Type II. A Type I Error happens when the null hypothesis is correct, but the test tells you to reject it. For a statistical test to be appropriate, the first requirement is that the Type I Error rate is equal (or sufficiently close to) the claimed α level.

Thus, to determine if OLS is an appropriate test, let us examine the Type I Error rate to check that it is (close enough to) α . This means we generate our data from the null hypothesis *while* meeting the requirements.

Here is the script to generate the data once

```

|| set.seed(30)
|| beta0 = 3
|| beta1 = 0
||
|| x = 1:20
|| e = rnorm(20, m=0, s=1)
||
|| y = beta0 + beta1*x + e

```

This tests the null hypothesis that $\beta_1 = 0$:

```

|| model = lm(y ~ x)
|| summary(model)

```

The p-value corresponding to the test of our hypothesis is 0.3784. We can have R just give that number to us:

```

|| summary(model)[[4]][2,4]

```

That is one p-value.

If the test is entirely appropriate, we would reject our true null hypothesis about α of the time. This statement is equivalent to the statement that the p-values follow a standard Uniform distribution (Section S.6.2).

$$P \sim \mathcal{U}(0,1) \quad (5.2)$$

Why?

appropriate

If the test is appropriate, and if we reject when the p-value is less than α , what is the probability of rejecting? It should be α . Why?

Question

Why *should* the probability of committing a Type I error be α ?

In statistical symbols, this is

$$\mathbb{P}[P \leq \alpha] = \alpha \quad (5.3)$$

This is just the cumulative distribution function for the standard Uniform distribution. Thus, $P \sim \mathcal{U}(0, 1)$.

The code above gave us one p-value. To investigate the distribution of p-values, you need to obtain thousands of p-values. The easiest way to do this is to loop through the above steps and save the p-value from each test.

or millions

That is what the following code does:

```
set.seed(30)

pval = numeric()

beta0 = 3
beta1 = 0
x = 1:20

for(i in 1:1e4) {
  e = rnorm(20, m=0, s=1)
  y = beta0 + beta1*x + e
  model = lm(y ~ x)
  pval[i] = summary(model)[[4]][2,4]
}
```

At the end of running this code, the variable `pval` contains 10,000 ($1e4 = 1 \times 10^4$) observed p-values.

To determine if it follows a standard Uniform distribution, we can create a histogram. When you do so, you note that it appears to closely follow a standard Uniform distribution, although not exactly. We would not expect it to follow the distribution *exactly*; those p-values are based on random samples and are therefore random values and the results will therefore be random.

Understanding the meaning of the p-value (above), we can check if the test is appropriate for $\alpha = 0.05$ by checking that the rejection rate is sufficiently close to

0.05. In other words, we can test if the proportion of p-values less than α is close enough to α . The Binomial test can accomplish this:³

```
|| binom.test( x=sum(pval<0.05), n=length(pval), p=0.05 )
```

The null hypothesis is that the proportion of rejections is equal to 0.05. The p-value of 0.9817 indicates that the claim that the Type I Error rate is 0.05 is reasonable; that is, the test seems reasonable. In other words, the test appears to be fine for $\alpha = 0.05$.

To determine if the test is appropriate for all possible α -values, we could repeat the above for every possible value of α . That would only take ∞ years. On the other hand, we can test all possible α values at once by recognizing that the p-values should follow a standard Uniform distribution and testing if they do.

We can perform a statistical test to determine if the distribution of the observed p-values is sufficiently non-Uniform:

```
|| ks.test(pval, "punif")
```

This line performs the Kolmogorov-Smirnov test (Massey 1951). Its null hypothesis is that the observed values follow the distribution stated. Because the p-value is 0.4361, we fail to reject the null hypothesis that the p-values follow a standard Uniform distribution.

In other words: the test appears to be universally appropriate. That is, it seems to be appropriate for *any* value of α you select.

Thus, we know that the test associated with testing the null hypothesis $\beta_1 = 0$ is appropriate for our usual value of α as well as for all values of α .

It is good to know that OLS works when the assumptions are met.

³The Binomial test is the exact test for checking that the rejection rate is equal to the claimed rate, 0.05. In your introductory course, you may have learned either the proportions test or the Wald test. Both are approximate tests that rely on the Normal approximation to the Binomial distribution.



Warning: When running tests, you will tend to just look for the p-value and draw conclusions based on that one number. If you do, be very clear what that p-value measures. In the previous example, there were 10,002 different p-values. The first 10,000 were p-values for the test that $\beta_1 = 0$ for 10,000 different sets of data. The 10,001th p-value determined if the distribution of those p-value was standard Uniform. The last p-value determined if the proportion of those p-values less than $\alpha = 0.05$ was 0.05.

There were 10,002 different tests in this example.

NON-NORMAL RESIDUALS I: Now, what about if the Normality assumption is not met? What if the residuals follow an Exponential distribution? The following code generates 10,000 p-values where the residuals follow an Exponential distribution with $\lambda = 1$. Be able to compare it to the previous code listing and find the *one* difference.

```
set.seed(30)

pval = numeric()

beta0 = 3
beta1 = 0
x = 1:20

for(i in 1:1e4) {
  e = rexp(20, rate=1)
  y = beta0 + beta1*x + e
  model = lm(y ~ x)
  pval[i] = summary(model)[[4]][2,4]
}

binom.test( x=sum(pval<0.05), n=length(pval), p=0.05 )

ks.test(pval, "punif")
```

The Binomial test returns a p-value of 0.422, which suggests to us that the OLS test is appropriate for $\alpha = 0.05$. Furthermore, the Kolmogorov-Smirnov test returns a p-value of 0.2867. Thus, even if the residuals are skewed this much ($\gamma_1 = 2$), the tests arising from the ordinary least squares estimation method (Theorem 4.2.1) appear universally appropriate.⁴

⁴The parameter γ_1 is a measure of skew. It is defined as the third standardized central moment. See Appendix S.6.5.

Note: The sample size in these test is $n = 20$. This is much lower than the usual “rule of thumb” of $n = 30$.

Also note that what we actually discovered is that OLS is very robust to violations of some of its requirements.

Not all of them, just some.

Exploration helps us determine which and how violated the requirements must be before the tests give useless results.

NON-NORMAL RESIDUALS II: Now, let’s create a different skewed distribution for the residuals that is even more skewed: a Chi-Square($\nu = 1$) distribution ($\gamma_1 = \sqrt{8} \approx 2.8$). Again, be able to compare it to the previous code listing and find the one difference.

```
set.seed(30)

pval = numeric()

beta0 = 3
beta1 = 0
x = 1:20

for(i in 1:1e4) {
  e = rchisq(20, df=1)
  y = beta0 + beta1*x + e
  model = lm(y ~ x)
  pval[i] = summary(model)[[4]][2,4]
}
```

The Kolmogorov-Smirnov test returns a p-value of 0.2667. Thus, even if the residuals are this skewed, the tests arising from OLS appear universally appropriate. The Binomial test returns a p-value of 0.0939, which tells us that the OLS test appears to be appropriate for $\alpha = 0.05$.

Thus, even when the residuals follow *this* heavily skewed distribution, the conclusions based on our OLS tests (Theorem 4.2.1) seem to be appropriate for a sample size of $n = 20$.

NON-NORMAL RESIDUALS III: Now, let’s create a symmetric distribution for the residuals. Because its variance is not finite, the Central Limit Theorem does not apply: It is the Cauchy distribution (Section S.4.7). Again, be able to compare it to the previous code listing and find the one difference.

```
set.seed(30)

pval = numeric()
```

```

beta0 = 3
beta1 = 0
x = 1:20

for(i in 1:1e4) {
  e = rcauchy(20)
  y = beta0 + beta1*x + e
  model = lm(y ~ x)
  pval[i] = summary(model)[[4]][2,4]
}

```

The Kolmogorov-Smirnov test returns a p-value of essentially 0. The Binomial test also returns a p-value of essentially 0. This tells us that the OLS test is not appropriate everywhere or even at $\alpha = 0.05$ when the residuals come from a Cauchy distribution.

Increasing the sample size will not fix this issue, either:

```

set.seed(30)

pval = numeric()

beta0 = 3
beta1 = 0
x = 1:5000

for(i in 1:1e4) {
  e = rcauchy(5000)
  y = beta0 + beta1*x + e
  model = lm(y ~ x)
  pval[i] = summary(model)[[4]][2,4]
}

```

In this code, the sample size is 5000. The conclusions are the same.

Looking at the histogram, you see that the first bar is much smaller than the others. This means the OLS tests reject at a lower rate than 0.05.

rejection rate

Note: When the underlying distribution does not have a finite variance, such as the Cauchy distribution, the Central Limit Theorem does not apply. That means increasing the sample size has absolutely no effect on the Normality of the sums. The sample sums are *never* Normally distributed.

5.2: Constant Expected Value

A second requirement of ordinary least squares (OLS) is that the expected value of the residuals is constant (and zero). From the gut, this means that the residuals evenly bounce above and below our estimates (regression curve). If the residuals are above (or below) our estimates more than expected, then the curve should be moved up (or down) to provide a “better” fit.

equivalent

We used this requirement in many places in Chapters 2 and 4. This requirement is entirely equivalent to the assumption that the underlying model (expected/predicted values) consistently fits the data, that there is no systematic error.

Note: Be aware, however, that the mathematics behind OLS will force the average residual to be zero. This means two things. First, the OLS model is “self-correcting,” in that the “line of best fit” will provide the best linear fit to the data. Second, the OLS model ensures that it is impossible to detect a systematic error in the measurements.

5.2.1 GRAPHICAL TEST Graph the residuals against each of the independent variables. Look for non-linear patterns in the plot (parabolas, cubics, etc.). If such exists, your model is misspecified. A fix is to transform the independent variable to eliminate that pattern. This is one place where the graphical “test” is superior to the numeric test. If you can identify the pattern, you have the fix.

residuals plot

For instance, if the residuals have the pattern in Figure 5.4, then the solution may be to use x^2 in place of (or in addition to) x . To see this, run the following code to obtain Figure 5.4.

```
set.seed(370)
x = seq(0, 3, length=20)
n = length(x)
e = rnorm(n)
y = 4 + 2*x^2 + e

mod = lm(y ~ x)
E = residuals(mod)

plot(x, E)                                ## residuals plot
```

Note the strong quadratic shape to the residuals plot (Figure 5.4). This strongly suggests that the model is misspecified.

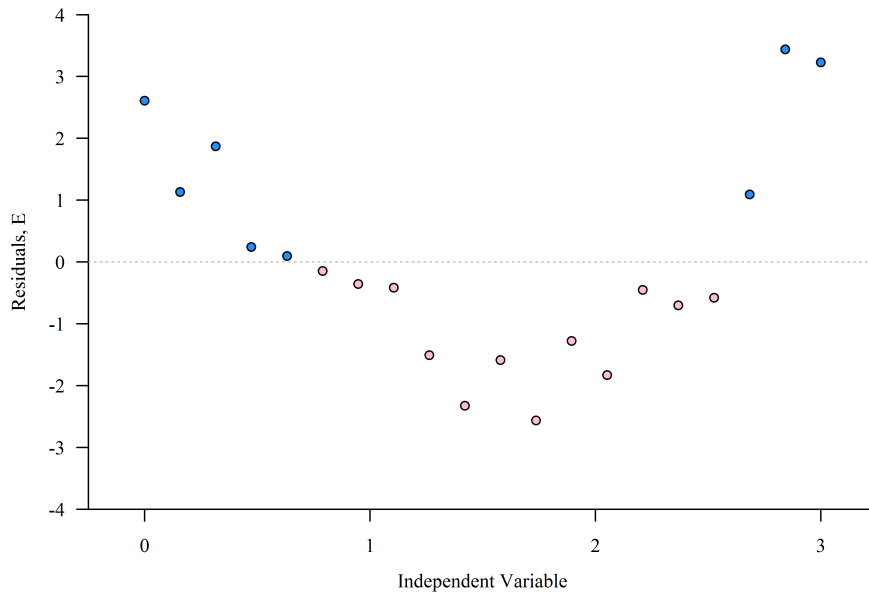


Figure 5.4: A residuals plot for a misspecified model. Note that the residuals show a definite quadratic form to them. Fixing this issue may be as simple as including x^2 as an additional independent variable.

Note: As an aside, note that there are only three “runs” in the residuals. A run is a sequence of values on one side of the prediction line. According to Figure 5.4, the first run consists of the first 5 values; the second, the next 12 values; and the third, the last 3 values.

run

Since the number of runs is based on the Binomial distribution, we can calculate the probability of observing this number of runs under the null hypothesis. Thus, we can calculate a p-value for the hypothesis that the model is properly specified (see Section S.6.3).

This example clearly shows that the model is misspecified. There is still some information contained in the residuals. It would be wrong to ignore that information.

information

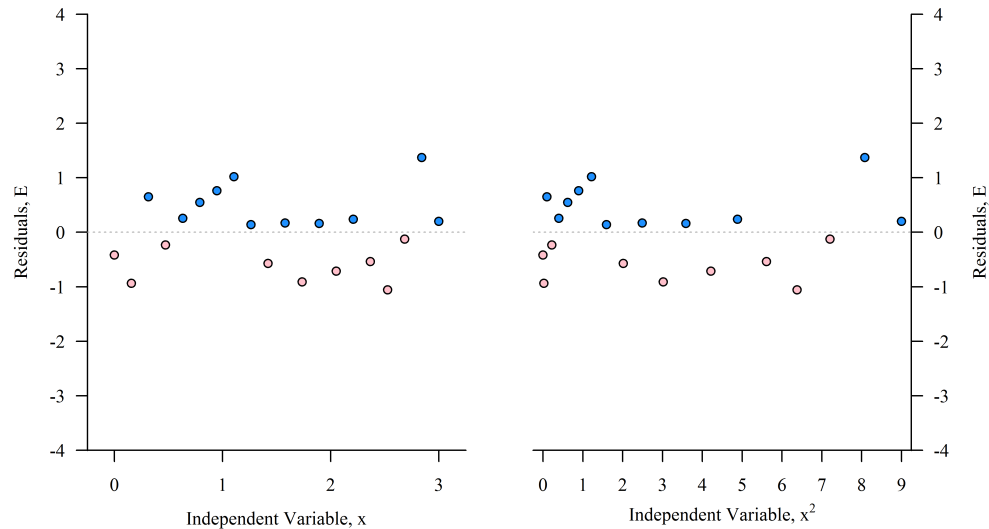


Figure 5.5: Residuals plots for the properly specified model. There is one residuals plot per independent variable. Note that the residuals in neither plot suggest anything other than a lack of pattern.

Since the residuals plot has a prominent quadratic shape, a solution is to include x^2 in the model:

```

|||  x2 = x^2
|||  mod = lm(y ~ x + x2)
|||  E2 = residuals(mod)
|||
|||  plot(x, E2)                ## residuals plot

```

one? With this change, we re-examine the residuals plot — one residuals plot for each independent variable. Since we have two (one?), we need to examine two (one?) residuals plots (Figure 5.5). Note the transformation was successful. Neither plot shows anything other than random bouncing across the line $y = 0$.

Happiness!

Note: There is a habit to feel sad that some requirement is not met by the model/data, such as above. However, do not feel sad. Feel happy, because you have learned something new about the relationships in the data! We know more! Celebrate!

5.2.2 A NUMERIC TEST That last sentence leads us to a numeric test. Compare the plots of Figure 5.4 and 5.5. The residual is colored blue if it is positive and pink otherwise. In Figure 5.4, there are long unbroken streaks (runs) of blue and pink. In Figure 5.5, the length of those runs is much reduced *and* the number is increased.

The test suggested by the above is not-surprisingly called the runs test (Section S.6.3). It is implemented in the `lawstat` package as the function `runs.test`. It takes just one piece of information: the residuals in the order of the independent variable.

It is also implemented in the `randtests` and `snpar` packages, as well as the `rfs` add-on, which is what I use in this book.

Here, I demonstrate the `runs.test` function in the `rfs` add-on. Note that this function currently requires the `lawstat` package to be installed. This restriction may change in the future.

```
source("http://rfs.kvasaheim.com/rfs.R")
library(lawstat)

set.seed(370)
x = runif(100)
e = rnorm(100)

runs.test(e, order=x)      ## The runs test
```

In this version of the `runs.test` function, the first slot goes to the residuals, and the second slot goes to the independent variable.

The output of this code is

```
Runs Test - Two sided

data: e, as ordered by x
Standardized Runs Statistic = 1.6081, p-value = 0.1078
```

As usual, check the p-value. Since the p-value of 0.1078 is greater than the α level of 0.05, we fail to reject the null hypothesis that the expected value of the residuals is constant and zero. Thus, since the p-value is greater than α , the model passes this test.

5.2.3 EXPLORATION OF THE EFFECTS OF NON-CONSTANT EXPECTED VALUE To see the effect of a non-constant expected value, let us revisit one of the proofs from Chapter 2.

What is the expected value of b_1 (the slope)?

$$\mathbb{E}[b_1] = \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad (5.4)$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \quad (5.5)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[Y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.6)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \mathbb{E}[\varepsilon_i])}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.7)$$

$$= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[\varepsilon_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.8)$$

$$= \beta_0 \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[\varepsilon_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.9)$$

$$= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[\varepsilon_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.10)$$

exercise

Now, this last line is β_1 if ε is independent of x . So, if the expected value is constant zero, the OLS estimator of β_1 is clearly unbiased. I leave it as an exercise to show that if it is constant, but non-zero, then the OLS estimator of β_1 remains unbiased. (Hint: If the expected value is constant, the $\mathbb{E}[\varepsilon_i] = c$ for some constant c .)

If the assumption of a constant expected residual value is violated, the OLS estimate of β_1 is biased. This is not a good thing. It means that your predictions are wrong... even "on average."

But, what about the OLS estimator of β_0 , the y-intercept? What effect does a non-zero expected value have on it? To see, let us revisit the proof of the unbiasedness of b_0 .

$$\mathbb{E}[b_0] = \mathbb{E}[\bar{Y} - \bar{x}b_1] \quad (5.11)$$

$$= \mathbb{E}[\bar{Y}] - \bar{x} \mathbb{E}[b_1] \quad (5.12)$$

$$= (\beta_0 + \bar{x}\beta_1 + \mathbb{E}[\varepsilon_i]) - \bar{x} \left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[\varepsilon_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \quad (5.13)$$

The last term is the expected value of b_1 from above. With that, we have

$$\mathbb{E}[b_0] = \beta_0 + \mathbb{E}[\varepsilon_i] - \bar{x} \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[\varepsilon_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.14)$$

Thus, there are two places that the non-constant expected value (poor model fit) affects the OLS estimator of β_0 . If $\mathbb{E}[\varepsilon_i] = 0$, the expected value of the errors is zero, then b_0 is unbiased for β_0 . If $\mathbb{E}[\varepsilon_i]$ is constant, but non-zero, then $\mathbb{E}[b_0] = \beta_0 + \mathbb{E}[\varepsilon_i]$. If $\mathbb{E}[\varepsilon_i]$ is a function of x , then $\mathbb{E}[b_0]$ is $\beta_0 + \mathbb{E}[\varepsilon_i]$ plus some function of x .

I leave it as an exercise for you to show that $\mathbb{E}[b_0] = \beta_0 + \mathbb{E}[\varepsilon_i]$ if $\bar{x} = 0$, regardless of whether the residuals are correlated with the independent variable. This is yet another reason some disciplines center their data before analyzing it.

exercise

Warning: The actual assumption is that the expected value of the residual is constantly zero. However, because of the mathematics of OLS, the mean residual is guaranteed to be zero (page 62). So, there is no way to test if the expected value of the residuals is constantly zero, only that it is constant.



Of all assumptions/requirements, this is the most important to meet. If your residuals depend on the value of x , then both the b_0 and b_1 estimators are biased. If the expected value of the residuals is not zero, then the b_0 estimator is biased.

important?

It is even worse. Because OLS mathematically forces $\bar{\varepsilon} = 0$, one cannot test if the expected value of the residuals *really* is zero (page 62). One must rely on the assumption that the data were collected without systematic error. That is, the statistician must trust the scientist to measure things correctly.

Question

When would the residuals be a function of the residuals?

This is an excellent question that you need to grapple with. It fundamentally means that you are missing an important variable from your model. Perhaps that variable is not an independent variable. Perhaps it is a confounding variable. Perhaps it is another *dependent* variable, which requires multivariate regression (beyond the scope of this book). It definitely means your model is weak and should be rejected if possible.

5.3: Constant Variance

The last assumption/requirement we will explore is the assumption that the residuals have a constant variance, σ^2 . We used this in many places in Chapter 4. In fact, every place you saw a σ^2 , we relied on the assumption it was a constant, that it was not really σ_i^2 .

If the variance is not constant with respect to the independent variables, neither the test statistic nor the confidence intervals will be correct.

Note: A note on vocabulary. If the variance of the residuals is constant, we claim the residuals are “homoskedastic.” Otherwise, they are “heteroskedastic.” Recall that “homo” means same, “hetero” means different, and “skedastic” means scatter.

5.3.1 GRAPHICAL TEST The graphical test is very similar to that for checking constant expected value. In that assumption, a residual plot was created. The middle of the vertical spread was traced out and checked to see if it was always near the zero line.

For testing constant variance, a residual plot can be used. The vertical spread of the data is traced out and checked that it does not vary too much. Note that the “vertical spread” is not the range, but the middle portion that contains about two-thirds of the data.

There are three stereotypical shapes that suggest heteroskedasticity. These three shapes, along with a shape suggesting homoskedasticity, are presented in Figure 5.6.

5.3.2 NUMERIC TEST In addition to what your eyes may tell you, it may be better to *also* perform a numeric test of homoskedasticity.⁵ There are a couple. The main test used in the regression area is the Breusch-Pagan test.

The way that the Breusch-Pagan test works is to refit the model including higher-order terms of the independent variables and compare the ratio of the two *SSE* values to a specific distribution. Since the null hypothesis is homoskedasticity (constant variance), a ratio close to one (large p-value) indicates that the model passes this test (the higher-order terms add little to the predictive ability of the

⁵I recommend performing both, when possible. If the p-value on the numeric test is too low, then the graphical test either gives you clues on where the problem is, or that the problem is practically minor and can be ignored.

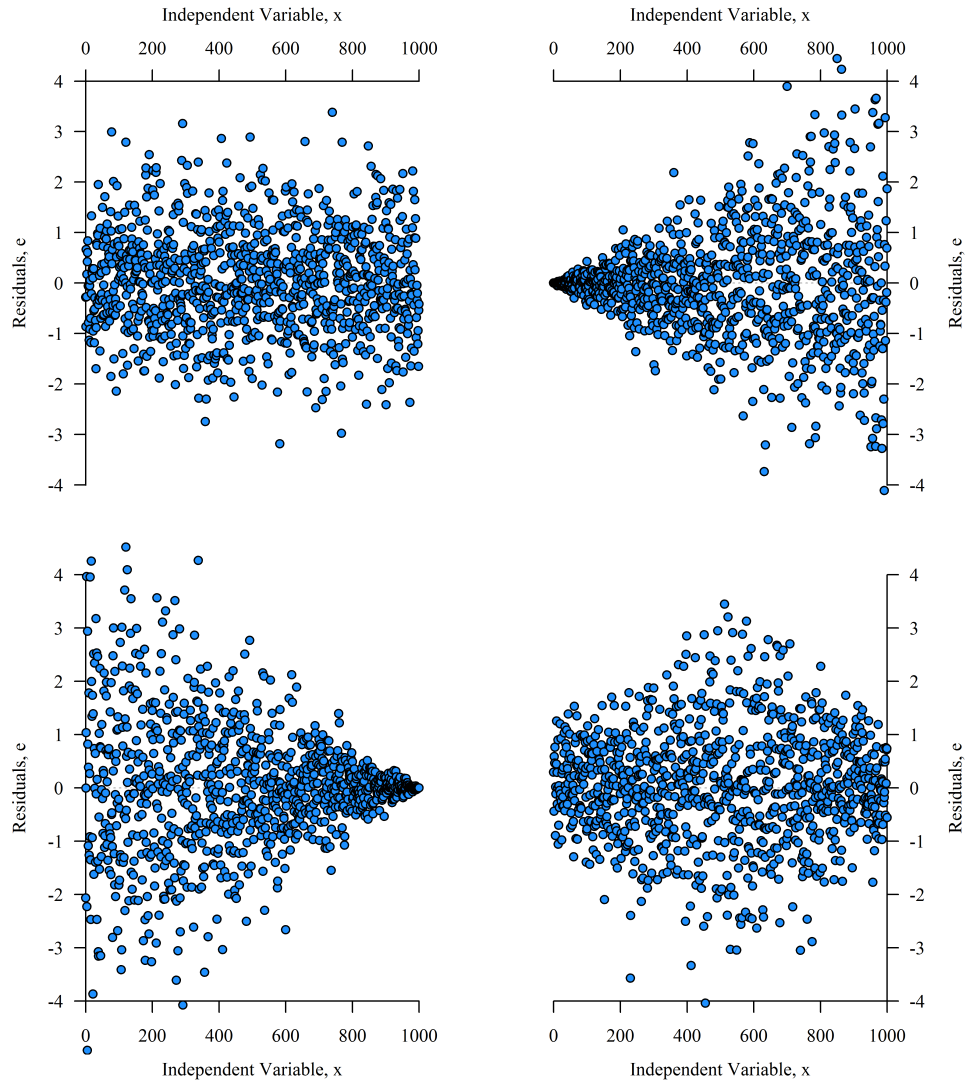


Figure 5.6: Residual plots illustrating homoskedasticity (top-left) and three typical types of heteroskedasticity: trumpet (top-right), funnel (bottom-left), and bulge (bottom-right).

model). A sufficiently small p-value indicates the presence of heteroskedasticity (those larger-order terms should be included to the model), which means the model needs improvement.

So, to examine the Breusch-Pagan test, let us generate our data, first according to our assumptions, then in violation of them. The Breusch-Pagan test is contained in the `lmtest` package as the function `bptest`.

bptest

```
|| set.seed(30)
|| x = seq(0,100)
|| n = length(x)
|| e = rnorm(n, m=0, s=1)
|| y = 3 + 2*x + e
||
|| plot(x,y, pch=21, bg="dodgerblue")
```

The scatter plot of this data does not really seem to suggest a changing variation in the residuals. Let us perform the Breusch-Pagan test to see if this numeric test also fails to detect a problem with the constant-variance assumption.

```
|| mod = lm(y ~ x)
|| bptest(mod)
```

If you get

```
Error: could not find function "bptest"
```

then you need to load (or install and load) the `lmtest` package.

Once you do, your output should look exactly like this

```
|| studentized Breusch-Pagan test
||
|| data: mod
|| BP = 0.23429, df = 1, p-value = 0.6284
```

Because the p-value is greater than α , we cannot reject the null hypothesis of homoskedasticity. The model passes the test.

Let us look at what happens when the residuals *are* heteroskedastic.

```
set.seed(30)
x = seq(0,100)
n = length(x)
e = rnorm(n, m=0, s=sqrt(x))
y = 3 + 2*x + e

plot(x,y, pch=21,bg="dodgerblue")
```

The scatter plot of this data definitely suggests increasing variation in the residuals, a trumpet shape that indicates heteroskedasticity. Let us perform the Breusch-Pagan test to see if this numeric test also detects a problem with the constant-variance assumption.

```
mod = lm(y ~ x)
bptest(mod)
```

Your output should look like this

```
              studentized Breusch-Pagan test

data:  mod
BP = 8.0992, df = 1, p-value = 0.004428
```

Because the p-value is less than α , we reject the null hypothesis of homoskedasticity. The model does not pass the test.

Note: Recall from your previous statistics course that a true null hypothesis will be rejected α proportion of the time and a false null hypothesis will be accepted β proportion of the time, where α is the Type I error rate and β is the Type II error rate. Neither of these numbers can be zero without making the other 1. So, be aware that you may be wrong.

5.3.3 EXPLORATION OF THE EFFECTS OF NON-CONSTANT VARIANCE (COVERAGE) Instead of looking at proofs or the distribution of p-values, let us generate data and look at the confidence intervals estimated using OLS. While it is easy to see that there *is* an effect using proofs, this method may make it easier to see *how serious* those effects are.

The first thing to do is to generate heteroskedastic data, calculate the 95% confidence intervals, and check if the true population parameters β_0 and β_1 are in the intervals. The second thing is to repeat this step many, many, many times to approximate the probability that the confidence intervals contains the population parameters. If the coverage of the estimated intervals are close to 95%, then the effect of heteroskedasticity is minor.

coverage

Here is the entire code

```
set.seed(30)
b0covered = numeric()
b1covered = numeric()

for(i in 1:1e4) {
  x = seq(0, 10, 0.5)
  n = length(x)
  e = rnorm(n, m=0, s=sqrt(x))
  y = 3 + 2*x + e
  mod = lm(y ~ x)
  ci = confint(mod)
  b0covered[i] = (3>ci[1,1]) && (3<ci[1,2])
  b1covered[i] = (2>ci[2,1]) && (2<ci[2,2])
}
```

Running this may take anywhere from a few seconds to a minute or more. At the end, you will have two variables, `b0covered` and `b1covered`, that contain 10,000 Boolean values (TRUE and FALSE) each. A TRUE indicates the population parameter was covered by the confidence interval for that random sample; a FALSE indicates it was not. The proportion of TRUE values can quickly be calculated using `mean(b0covered)` and `mean(b1covered)`.

The proportion of times the β_0 parameter was covered by the confidence interval was 0.996. Thus, the estimated coverage for β_0 when the data are heteroskedastic *in this manner* is 99.6%. The coverage rate for β_1 is about 95.1%. Both *should* be close to $1 - \alpha$, 95%.

interpretation

From these results, we know a couple of things. First, this amount of heteroskedasticity did not (practically) significantly affect the confidence intervals for the slope parameter, β_1 . Second, it *did* (practically) significantly affect it for the intercept parameter. In fact, because the reported confidence interval is much wider than the “true” confidence interval, a researcher will reject too *infrequently*. This means the power of the test is too low.

Let us try a greater level of heteroskedasticity. And, let us make it funnel instead of the above trumpet:

```
set.seed(30)
b0covered = b1covered = numeric()

for(i in 1:1e4) {
  x = seq(0,10,0.5)
  n = length(x)
  e = rnorm(n, m=0, s=15-x)
  y = 3 + 2*x + e
  mod = lm(y ~ x)
  ci = confint(mod)
  b0covered[i] = (3>ci[1,1]) && (3<ci[1,2])
  b1covered[i] = (2>ci[2,1]) && (2<ci[2,2])
}
```

Note what changed: the standard deviation of the residuals. It starts high and gets smaller — a funnel shape.

interpretation

This scheme produces a coverage estimate of 88.8% for the intercept parameter and 93.9% for the slope parameter. Thus, the confidence interval for the intercept is again affected more than that of the slope parameter. In fact, the slope parameter is relatively unchanged.

Let us try once more. This time, let us look at bulge heteroskedasticity.

```
set.seed(30)
b0covered = b1covered = numeric()

for(i in 1:1e4) {
  x = seq(0,10,0.5)
  n = length(x)
  e = rnorm(n, m=0, s=11-2*abs(x-5))
  y = 3 + 2*x + e
  mod = lm(y ~ x)
  ci = confint(mod)
  b0covered[i] = (3>ci[1,1]) && (3<ci[1,2])
  b1covered[i] = (2>ci[2,1]) && (2<ci[2,2])
}
```


Again, note the common code and the slight changes. Here, the variance starts low (1), gets higher (11), then decreases again to 1 — a bulge heteroskedasticity.

The coverage for both parameters are far from 95%. For β_0 , the coverage is 99.3%; for β_1 , 99.8%. Both indicate that the researcher will reject the hypotheses far too infrequently.

interpretation

TEST STATISTICS: The test statistics for β_0 and β_1 are calculated using the MSE. That test statistic, if all of the assumptions are true, follows the Student's t distribution with $n - p$ degrees of freedom. However, if the variance is a function of the independent variable, the distribution of the test statistic is also a function of the independent variable. This is a problem, because the "correct" p-values and confidence intervals would also be functions of x .

The analysis of the effect of heteroskedasticity can be repeated for the distribution of test statistics. Or, we can realize that we reject the null hypothesis when the confidence interval does not contain our hypothesized value. That is, the analysis for the confidence interval is sufficient for our understanding of the p-value:

rejection

1. If the confidence interval is too wide, then the p-values will be too high (reject the null hypothesis too infrequently).
2. If the confidence intervals are too narrow, then the p-values are too low (reject too frequently).

Of the two possibilities, the first (which produces a higher Type I error rate) may be better from the researcher's point of view in some cases: While the researcher would end up rejecting too infrequently, those rejections are more sure because the true p-value is less than the observed p-value. One may prefer this when it worse to commit a Type I error (rejecting a true null hypothesis).

On the other hand, however, the second may be better in certain circumstances. It will have a lower Type II error rate. If it is more important to protect against a Type II error, then this will be the better scenario.

Note: These findings regarding the hypothesis tests about β *strictly* relate only to these particular three types of heteroskedasticity at these three particular levels. If you find your heteroskedasticity is much higher than those used in these examples, you should run a coverage analysis similar to these to see how bad the heteroskedasticity really affects the confidence intervals.

The above analysis examined the effects of heteroskedasticity on the parameter estimates. It did *not* examine its effect on the prediction of y-values. Those effects are *much more pronounced* and are also a function of the variance *at the x-value*.

To see this in one example, let us generate trumpet-shaped heteroskedastic residuals, predict the value of y at three points along the x -axis, and determine how frequently those points are covered by the calculated confidence intervals.

Before we do this, let us see if we can determine what to expect. For low values of x , there is very little true variation in the predicted value. Thus, we would expect the predicted value to fall in the calculated confidence interval very frequently.

For middling values of x , the true and estimated variances are close to each other. Thus, we would expect coverage to be close to the nominal 95%.

For large values of x , the true variation is much higher than the estimated variation. Thus, a lot of the predicted y -values should fall outside the confidence interval. We should expect the coverage to be relatively low. Let's see if these expectations are met by reality.

coverage

Here is the entire code

```
set.seed(30)

y1covered = numeric()
y2covered = numeric()
y3covered = numeric()

for(i in 1:1e4) {
  x = seq(0,10,0.5)
  n = length(x)
  e = rnorm(n, m=0, s=sqrt(x))
  y = 3 + 2*x + e

  mod = lm(y ~ x)

  ypr = predict(mod, newdata=data.frame(x=c(0,5,10)), interval=
    "confidence")

  y1covered[i] = ( 3 > ypr[1,2] ) && ( 3 < ypr[1,3] )
  y2covered[i] = (13 > ypr[2,2] ) && (13 < ypr[2,3] )
  y3covered[i] = (23 > ypr[3,2] ) && (23 < ypr[3,3] )
}

mean(y1covered)
mean(y2covered)
mean(y3covered)
```

interpretation

The coverage for y when x is very low is 99.6%, which is very high. The coverage for y when x is in the middle is 94.5%, which is about what we want. The coverage for y when x is very high is 88.7%, which is relatively low.

These results are what we expected.

Note: Realize again the connection between p-values and confidence intervals. To ensure that our p-values are “protected” — are no more than estimated — we need to limit ourselves to the areas where the true spread is no larger than the average.

5.3.4 HUBER-WHITE HETEROSKEDASTIC ESTIMATORS Because heteroskedasticity is common in several areas of study, it would be helpful to find a method for reducing its effect. This was the task that Professors Huber and White set for themselves.

Ultimately, their solution arises from using matrix multiplication to make the residuals homoskedastic.

sgf sf sfg fgs
gf sfg sdf gs dfgs dfg sdf g
dgs s fs g dfs

5.4: Multicollinearity

Recall that there was just one requirement of the mathematics on ordinary least squares estimation: The independent variables could not be linear combinations of each other (Section 3.1.1). If this is the case, then we cannot do the mathematics behind OLS estimation.

supercollinearity

However, such perfect multicollinearity is not the only problem that can occur. If two (or more) independent variables are highly correlated, then problems can arise. These problems can be examined in terms of either the computer science or the experimental logic.

5.4.1 THE CS OF MULTICOLLINEARITY Note that we are using a computer to perform our calculations. Because a computer is not able to store most numbers in memory perfectly, rounding errors creep into the calculations. While this will always happen when the number does not have a finite binary representation, it is most important to understand when the decimal is close to zero, when it is less than the “machine epsilon.” When this is the case, the number rounds to zero. In other words, if the value is between $-\epsilon$ and $+\epsilon$, the computer treats it as a zero.

On 64-bit computers, the value of epsilon is approximately $2.220446 \times 10^{-16} = 0.000\ 000\ 000\ 000\ 000\ 222\ 044\ 6$. If the determinant of the matrix is this value or less in magnitude, the computer will claim it is singular (Appendix page 484).

```
|| solve( matrix( c(1,0, 0,2.220446e-16), ncol=2) )
```

Mathematically, we can calculate the determinant to be 2.220446×10^{-16} , which means the matrix is actually *not* singular. However, calculating the inverse returns the following:

```
|| Error in solve.default(matrix(c(1, 0, 0, 2.220446e-17), ncol
|| = 2)) :
|| system is computationally singular: reciprocal condition
|| number = 2.22045e-17
```

The lesson to take beyond this specific case is that the matrix does not have to be singular for the computer to tell you it is. All that is needed is that the determinant of the $X'X$ matrix be close to zero.

I term this a Computer Science result because it is based on the vagaries of computers instead of the vagaries of mathematics. The next section looks at what multicollinearity means in terms of the logic of experimental design and interpretation.

5.4.2 THE LOGIC OF MULTICOLLINEARITY The previous section examined the effects of multicollinearity on calculations, specifically of the inverse of the $X'X$ matrix. If its determinant is sufficiently close to zero, then the computer will treat it as a zero, meaning the matrix will be effectively singular. However, this is only a CS problem. A good statistician will pay attention to the conditions that *cause* multicollinearity... even minor amounts of it.

Recall that multicollinearity occurs when a column in the data matrix is a linear combination of the other columns; that is, it happens when one variable is a linear combination of the others; that is, **it happens when one variable adds no new information beyond what the other variables contain**. For instance, if one variable is a person's height in inches and another variable is a person's height in centimeters, then multicollinearity exists. The first variable offers no information that is not contained in the second.

A statistician cares about the independent variables in the model. They are designed to explain the dependent variable. Each independent variable is supposed to be independent of the others, because each is designed to explain a *different aspect* of the response variable. If two explanatory variables are highly correlated with each other, then it will be logically impossible to determine which of the two is causing the change in the dependent variable:

- Is it the logarithm of a person's height in inches or the logarithm of the *square* of a person's height in inches that can be used to estimate weight?
- Is it average daily temperature or ice cream consumption that can be used to estimate the violent crime rate?
- Is it educational attainment or parental income that can be used to estimate a person's future income?

These three exemplify the issue with multicollinearity in practice. The first example produces mathematical ("super-") multicollinearity because the logarithm of the square of a variable is exactly twice the logarithm of the variable. The first column is twice the second.

The second example does not exemplify mathematical multicollinearity. There is no function of average daily temperature that gives the ice cream consumption. However, there is a very strong linear relationship between the two. Because of this,

one cannot *statistically* tell if it is the temperature or the ice cream that is affecting violent crime. With this being said, unless the dairy farmers are attacking the very foundation of society, the substantive scientific theory suggests that the temperature is likely the factor affecting the crime rate, not the frigid dairy.

The third example is more subtle. There is also a strong relationship between a person's education attainment and the parent's income (at least in the United States). Because of this, we are unable to statistically determine if it is the person's educational attainment or the income of the person's parents that affects the person's future income. Social science theories suggest each. The statistics with each explanatory variable also suggest each. What can we do in this case?

INDICATIONS OF MULTICOLLINEARITY: To see some statistical indications of multicollinearity, try the following code.

```
set.seed(30)

b0 = 3
b1 = 2
b2 = 3

x1 = seq(0,10,length=8)
x2 = c(1,2,3,4,6,7,8,9)
e = rnorm(8)

y = b0 + b1*x1 + b2*x2 + e

mod1 = lm(y ~ x1)
mod2 = lm(y ~ x2)
modA = lm(y ~ x1+x2)
```

Clearly, from how this experiment is set up, we know the following:

- There is a strong relationship between x_1 and y .
- There is a strong relationship between x_2 and y .
- There is a strong relationship between x_1 and x_2 .

Running `summary(mod1)` shows us that the first statement is true, *if we ignore the effect of x_2* . Similarly, running `summary(mod2)` shows us that the second statement is true, *if we ignore the effect of x_1* . Combining the two explanatory variables in `modA` is confusing if we do not think about multicollinearity. `summary(modA)` gives the following results:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.814      1.898    1.483   0.198
```

```

x1          2.145      1.681   1.276   0.258
x2          2.861      2.009   1.424   0.214

Residual standard error: 1.386 on 5 degrees of freedom
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9924
F-statistic: 458.7 on 2 and 5 DF,  p-value: 2.164e-06

```

Note that this model shows that *neither* independent variable is valuable in modeling the dependent variable. Since a scientist will usually put all the explanatory variables in the model, this is a lesson for us to pay attention to the relationships among the independent variables.

Note: It is also important to look at *all* of the regression output. Note that neither of the two independent variables have small p-values. However, the R^2 value is very close to 1. Thus, this inconsistency also suggests that something is wrong.

A TEST OF MULTICOLLINEARITY: How do we statistically detect this type of multicollinearity? A simple correlation test will not suffice if we have more than two independent variables because correlation is between only two variables.

The answer comes from the cause of the multicollinearity: If there is multicollinearity, then one independent variable should be linearly related to the others. A linear regression will be able to detect this. *Technically*, a linear regression for each independent variable will detect this. To make this process easy, there is the `vif` function in the `car` package. This function calculates the “variance inflation factor” for each independent variable. The variance inflation factor for independent variable i is defined as

$$\text{VIF}_i := \frac{1}{1 - R_i^2} \quad (5.15)$$

Here, R_i^2 is the R-squared value for the model regressing the independent variables on independent variable i .

In our example above, one can calculate the VIF by hand:

```

summary( lm(x1 ~ x2) )
1/(1-0.9921)

```

The higher the value of R^2 , the more independent variable i can be explained by the other independent variables. In other words, the higher the VIF, the less new information that variable adds to the model.

As in much of the field, this description leads to the question

How high is too high?

The “rule of thumb” depends on the discipline. Typical cut-offs are 5, 8, and 10. If the VIF for any of the variables is greater than the cut-off, then there is “too much multicollinearity in the the model.”

FIXING MULTICOLLINEARITY: So, let’s say that you have detected multicollinearity in your model. What can you do?

The presence of multicollinearity means that one of your independent variables is highly correlated with a linear function of the others. It adds little to the understanding of the response variable. However, is it variable i that should be examined or the others?

From a statistical standpoint, the model is not too helpful. Multiple variables are trying to explain the same aspects of the dependent variable. In other words,

- Is it educational attainment or parent’s income that affects the respondent’s income?
- Is it race or poverty that affects violent crime?
- Is it intelligence or birth position (oldest, youngest, middle, etc. child) that affects success?
- Is it the ranch or the cattle feed that affects the weight?
- Is it race or income or religion or parental income or home state that affects voting behavior?
- Is it Nordic ancestry or blood type or neanderthal genes that affect the severity of CoViD-19?

In each of these, the explanatory variables are highly correlated and have been used to model the response variable. Because of the correlations, conclusions about what *really* affects the dependent variable are unclear. Statistically, the answer is “Yes, each does.”

However, since each of the independent variables above are correlated, their effects overlap. This is represented as the purple overlapping area in Figure 5.7. While each variable has an effect on the dependent variable (red and blue), that effect is also split with the other variable (or variables). As such, the key is trying to separate the three sections to determine whether it is the red, the blue, or the purple that is affecting the response variable.

Unfortunately, this is beyond the scope of this course. For those who are interested, you may want to investigate factor analysis (FA) and principal component analysis (PCA). These are two methods for dealing with that overlap (purple area). The first focuses on estimating the purple area; the second, on creating two other variables that combine the two independent variables into their independent

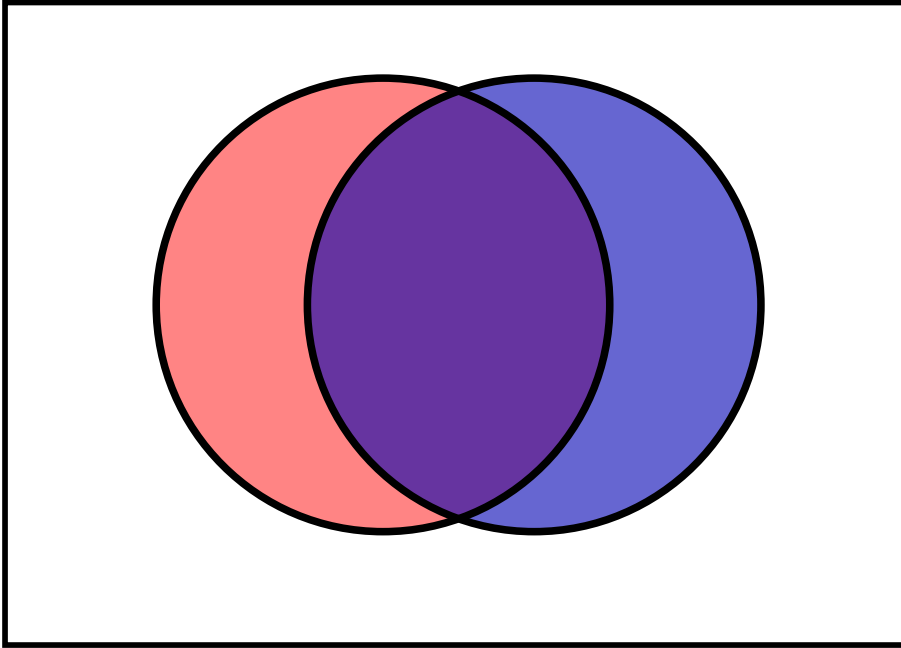


Figure 5.7: Diagram to illustrate multicollinearity. The left circle is the effect of the first independent variable on the dependent variable; the right, the second. The purple overlap represents the similarity between the two variables. The rectangle represents everything that affects the dependent variable, which means the colored portion represents everything in the model that affects the dependent variable.

components (the parts that are purely red and blue). The advantage is that the independent variables become independent ($VIF = 1$). The disadvantage is that the newly-created independent variables are only related to the original explanatory variables; thus, interpretation is made more complicated.

Note: By the way, Figure 5.7 illustrates several things about the model. Everything that affects the dependent variable is represented by the rectangle, whereas the colored part represents *only* what the model explains. Thus, one may think of the ratio of the colored area to the total area as being the R^2 value.

Question

Given that the size of the circles is fixed, how would you arrange them to cover the most area of the rectangle? What would that mean in terms of the variables?

5.5: Conclusion

In this chapter, we looked at violating the primary assumptions of ordinary least squares, as well as the effects of multicollinearity. Once we violated those requirements, we looked to see what effect that violation had on our parameters of interest.

We saw that violations of the Normality assumption cause very minor changes to our test statistics, unless the residuals were generated from a distribution with non-finite variance. In such a case, the tests were all but worthless.

Violations of constant expected value (proper model fit) were bad. They cause the OLS estimators, test statistics, *and* confidence intervals to be biased. The lesson here is to make sure that your measuring stick is properly calibrated.

Heteroskedasticity is not an issue for the estimators. They remain unbiased. They are an issue, however, for any testing done. This includes both the test statistic and the confidence interval. This is because the standard error used is the average, not the actual value for a general point.

Multicollinearity arises from independent variables that measure highly similar concepts. Thus, the statistical effect is that the standard errors are inflated. The logical effect is that we are unsure which of the two variables actually affects the dependent variable.

Warning: *Again, be aware of the multiple comparisons issue discussed in Appendix S.6.2. It explains why you need to either adjust your p-value or your alpha level when performing multiple tests, such as when you are testing both $\beta_0 = 4$ and $\beta_1 = 0$.*



5.6: End-of-Chapter Materials

5.6.1 R FUNCTIONS In this chapter, we were introduced to many, many, many R functions that will be useful in regression. In fact, this chapter uses more R functions than any other chapter in this book. Here are the many.

PACKAGES:

car This package provides several statistical tests used in the book “An R Companion to Applied Regression” by J. Fox and S. Weisberg. It is a great package that provides a lot of additional functionality for R.

lawstat This package provides several statistical tests used in law and public policy analysis. It provides the `runs.test` function for us.

lmtest This package provides many tests related to linear models. It provides an implementation of the Breusch-Pagan test, `bptest`, which tests for heteroskedasticity in the residuals.

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

STATISTICS:

source(filename) This function runs an R script from a separate file. That file may be local or on the Internet.

runs.test(E, order) This alteration to the `lawstat` function tests whether the variable `E`, as ordered by `order` exhibits fit issues.

shapiroTest(E) This tests the null hypothesis that the variable `E` comes from a Normal distribution. It is based on the `shapiro.test` function in the basic R installation. It adds capabilities to test Normality in several groups.

lm(formula) This is the function that performs ordinary least squares estimation on linear models.

bptest(mod) This function from the `lmtest` package performs the Breusch-Pagan test for heteroskedasticity.

confint(mod) This calculates confidence intervals for the parameters in ordinary least squares regression.

mean(x) This calculates the mean of a sample.

summary(x) This produces the six-number summary or a frequency table of the provided variable, depending on the type of variable.

summary.lm(mod) When applied to a linear model fit using either the `aov` function or the `lm` function, provides estimates of the effects of the numeric variables and the levels of the categorical variables in the model.

summary.aov(mod) When applied to a linear model fit using either the `aov` function or the `lm` function, provides estimates of the statistical significance of the variables in the model.

predict(mod) This predicts the values of the dependent variable at each point in the dataset *or* for the values specified.

fligner.test(formula) This tests for heteroskedasticity when the independent variable is categorical.

aov(formula) This function performs ordinary least squares estimation on linear models.

vif(model) This function calculates the variance inflation factor (VIF) for each of the independent variables in the model.

set.base(var,level) This `RFS` package function redefines the base category in the provided level. By default, the base category is the first according to the alphabet.

PROBABILITY:

set.seed(x) This sets the random number seed.

rexp(n, rate) This generates n random values from an Exponential distribution with the specified rate parameter.

rnorm(n, mean, sd) This generates n random values from a Normal distribution with specified mean and standard deviation. By default the mean is 0 and the standard deviation is 1.

runif(n, min, max) This generates n random values from a Uniform distribution with specified minimum and maximum values. By default, the minimum is 0 and the maximum is 1.

MATHEMATICS:

head(x) This returns the first six values in the variable.

foot(x) This returns the last six values in the variable.

seq(from, to, by, length) This returns a vector of sequential values, where `by` indicates the step size and `length` specifies the vector length.

length(x) This calculates the length of a vector (variable), which is the sample size, n .

residuals(mod) This calculates the residuals in the model, which is the difference between the observed and the predicted.

GRAPHICS:

qqnorm(x) This creates a Normal quantile-quantile plot for the given values.

qqline(x) This adds the diagonal line to the quantile-quantile plot.

overlay(x) This, from the `RFS` package, produces a histogram with a Normal curve overlaying it.

par(...) This sets parameters on the next graphic started. Look through the help page for this function to see all you can specify.

plot(x,y) This produces a scatter plot of the y -values against the x -values.

axis(side) When a plot is already drawn, this adds values along axis number `side`.

title(...) When a plot is already drawn, this adds the x - and y -labels.

lines(x,y) When a plot is already drawn, this draws lines between each subsequent (x,y) pair.

points(x,y) When a plot is already drawn, this draws points at each (x,y) pair.

PROGRAMMING:

attach(dataframe) This allows you to access the variables in the `dataframe` without having to prefix each with `dataframe$`.

library(package) This loads an external package that you have already installed on your computer. It allows access to all functions and data sets in the `package` package.

as.character(x) This changes the values in variable x to be characters.

as.numeric(x) This changes the values in variable x to be numbers.

5.6.2 EXERCISES

1. Show that if the expected value of the residuals is constant, but non-zero, then the OLS estimator of β_1 remains unbiased.
2. Show that $\mathbb{E}[b_0] = \beta_0 + \mathbb{E}[\varepsilon]$ if $\bar{x} = 0$, regardless of whether the residuals are correlated with the independent variable.

5.6.3 THEORY READINGS

- George E. P. Box. (1976) “Science and Statistics,” *Journal of the American Statistical Association*, 71(): 791–799.
doi:10.1080/01621459.1976.10480949.
- James V. Bradley. (1968) *Distribution-Free Statistical Tests*. New York: Prentice-Hall.
- John Fox and Harvey Sanford Weisberg. (2010) *An R Companion to Applied Regression*, second edition. Thousand Oaks, CA: SAGE Publications.
- Frank J. Massey, Jr. (1951) “The Kolmogorov-Smirnov Test for Goodness of Fit.” *Journal of the American Statistical Society*. 46(253): 68–78.
- Abraham Wald and Jack Wolfowitz. (1940) “On a test whether two samples are from the same population.” *The Annals of Mathematical Statistics* **11**: 147–162.



CHAPTER 6:

A TIME FOR SOME EXAMPLES

OVERVIEW:

We have covered a lot of theory and mathematics over the past several chapters. Here, we will apply what we have learned to help settle the theory into our minds. In other words, we will perform the analysis process with the information and skills we now have.

This means we will use data to answer our research questions. Of course, we will need to examine the research question to determine the appropriate model, check the assumptions — both statistically and graphically — and properly interpret the results.

That is a lot of summarizing to do!

Chapter Contents

6.1	Full Example: Violent Crime	153
6.2	Full Example: Violent Crime, Wealth, Region	158
6.3	Full Example: Cows in the City of Děčín	166
6.4	Conclusion	184
6.5	End-of-Chapter Materials	185



And so, we have completed a majority of the important mathematics underlying ordinary least squares estimation. Be aware that OLS is how we estimate the parameters. The model itself is referred to as the classical linear model. It makes the usual four assumptions. The observations follow the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad (6.1)$$

and the residuals follow this distribution

$$\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0; \sigma^2) \quad (6.2)$$

From those assumptions, we were able to use OLS to calculate formulas for the estimators of $\beta_0, \beta_1, \dots, \beta_{p-1}$. The next chapter used the distribution to determine the distribution of those estimators. This led to confidence intervals for the parameters and test statistics for testing hypotheses about the parameters. It also led to distributions and intervals and test statistics for estimated and predicted values of y .

All of that from four small assumptions.



This chapter will apply these results to different research questions to illustrate the statistical research process. So, turn the page and begin seeing applications of what we have done.

6.1: Full Example: Violent Crime

To help settle all of this, let's see a simple extended example of modeling the violent crime rate in 2000 using just the violent crime rate in 1990.

The preamble is the part of the code that imports the extra functions, loads the data, and gives us an overview of it. This is a typical preamble

preamble

```
### Preamble
source("http://rfs.kvasaheim.com/rfs.R")
library(lawstat)
library(lmtest)

dt = read.csv("http://rur.kvasaheim.com/data/crime.csv")
attach(dt)

summary(dt)
```

Note that there are *many* variables in this data set. Since we are modeling the violent crime rate in 2000 using the rate in 1990, we will only use the variables `vcrime00` and `vcrime90`. To fit the model and estimate the parameters using ordinary least squares, run this line:

OLS

```
|| crimeMod = lm(vcrime00 ~ vcrime90)
```

Nothing gets outputted by this line. R just echoes it if you typed it correctly. However, a lot has happened behind the scenes. Inside R, the model was fit using ordinary least squares (using matrices). The parameters were estimated. All of this was done behind the scenes.

The next step is to check that the model does not violate any of the assumptions/requirements.

6.1.1 NORMALITY OF THE RESIDUALS The first we will check is the Normality of the residuals:

```
|| e = residuals(crimeMod)
|| # Normal Residuals?
|| overlay(e)
|| shapiroTest(e)
```

The histogram overlaid with the normal curve suggests the residuals are slightly skewed to the right. The Shapiro-Wilk test strongly indicates a lack of Normality (p-value = 0.004424). The sample size of $n = 51$, however, definitely seems large enough to ensure the sums of the residuals closely follows a normal distribution

(this is the *actual* requirement). If you would like to check this, run the following code (first think what it does and why it answers this problem):

```
|| et = numeric()
|| for(i in 1:1e3) {
||   x = sample(e, replace=TRUE)
||   et[i] = sum( x )
|| }
|| shapiroTest(et)
```

Since the reported p-value is much greater than α , we can conclude that the sample sums are sufficiently Normal. And, it is the sample sums that affect the distribution of b_0 and b_1 .

Thus, the model passes the normality requirement.

■ 6.1.2 CONSTANT EXPECTED VALUE (FUNCTIONAL FORM) The second assumption to test is constant expected value (proper model form):

```
|| # Constant Expected Value of Residuals
|| plot(vcrime90,e)
|| runs.test(e, order=vcrime90)
```

The residuals plot seems a bit inconclusive to me. This is mainly due to the single point far to the right (the District of Columbia). The runs test, however, indicates that there is no significant evidence the residuals follow anything other than a horizontal line (p-value = 0.6732).

Thus, the model does not violate the second assumption.

■ 6.1.3 CONSTANT VARIANCE The third assumption is that the variance of the residuals is constant:

```
|| # Constant Variance of Residuals
|| plot(vcrime90,e)
|| bptest(crimeMod)
```

For me, the graphic is inconclusive because of DC. The Breusch-Pagan test did not detect significant heteroskedasticity (p-value = 0.1041).

Thus, the model passes the third and final requirement.

■ 6.1.4 THE FINAL MODEL This model seems appropriate, and we can now see the estimates:

```

Call:
lm(formula = vcrime00 ~ vcrime90)

Residuals:
    Min       1Q   Median       3Q      Max
-241.32  -42.84  -18.04   40.97  208.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.52716   21.42679   5.112 5.27e-06 ***
vcrime90     0.58065    0.03107  18.689 < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.10 1

Residual standard error: 85.55 on 49 degrees of freedom
Multiple R-squared:  0.877, Adjusted R-squared:  0.8745
F-statistic: 349.3 on 1 and 49 DF,  p-value: < 2.2e-16

```

The first section reports the model you presented. Use it to double-check you typed things in correctly or to remind you what the model is examining. The second part produces the five-number summary of the residuals. Since the mean (0) is greater than the median (-18.04), there is evidence of a positive skew to the residuals. This, we discovered above.

The third section is the “regression table.” Each row corresponds to a different independent variable (or the intercept). The columns are the estimates, the standard errors, the test statistic (estimate divided by standard error), and the p-value.

In this example, there is very strong evidence that the relationship between the violent crime rate in 1990 and in 2000 is positive ($b_1 = 0.58065$). If State A had a higher violent crime rates in 1990 than State B, then it also tended to have a higher violent crime rates in 2000.

The intercept, $b_0 = 109.52716$ indicates that for a state with 0 violent crime in 1990, the expected violent crime rate in 2000 is 109.52716 crimes per 100,000 people. However, since no state was close to having a violent crime rate in 1990 of 0, this interpretation does not make statistical sense.

Remember that we should only use our models to predict and estimate for values of the 1990 violent crime rate within the domain of the `vcrime90` variable in our data.

interpolation

6.1.5 GRAPHIC The following lengthy code produces a graphic like that at the top of the page:

```

|| plot.new()
|| plot.window( xlim=c(0,2500), ylim=c(0,2500) )

```

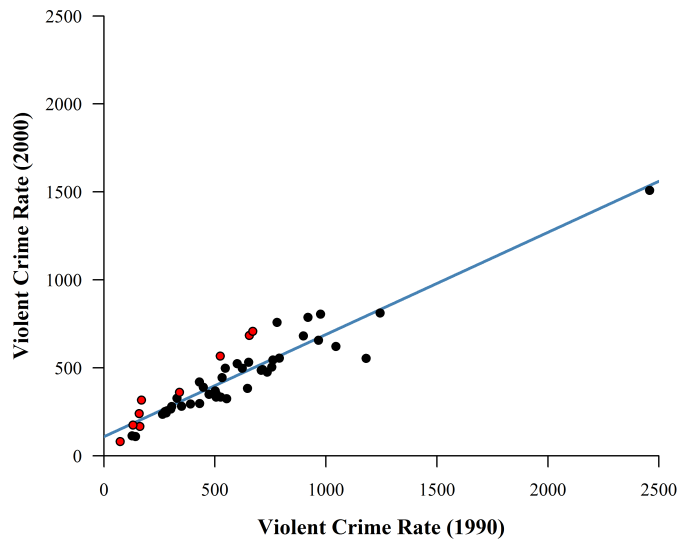


Figure 6.1: Plot of the violent crime rate in 2000 against that in 1990. The ordinary least squares line of best fit is included. Red-colored states are those whose violent crime rate increased from 1990 to 2000.

```
axis(1); axis(2)

title(xlab="Violent Crime Rate (1990)", line=2.75)
title(ylab="Violent Crime Rate (2000)", line=3.25)

xx = seq(0,2500)
yy = predict(crimeMod, newdata=data.frame(vcrime90=xx))
lines(xx,yy, col="steelblue", lwd=2)

points(vcrime90,vcrime00, pch=21, bg=1+(vcrime00>vcrime90))
```

Note that the graphic also indicates which states had their violent crime rate increase. It comes from this line:

```
|| points(vcrime90,vcrime00, pch=21, bg=1+(vcrime00>vcrime90))
```

The first two slots are the x- and y-values. The third slot specifies the plotting character. A `pch` of 21 is a dot with its insides colorable. The fourth slot, `bg`, specifies the color to fill the inside of the dots (`bg` = “background”).

The part `(vcrime00>vcrime90)` takes on value 1 if the violent crime rate increased and 0 otherwise. Adding 1 to each ensures that the two colors are 1 and 2 — black and red.

■ **6.1.6 CONFIDENCE INTERVAL FOR β_1** In addition to calculating the point estimates of the slope and intercept, we can also calculate confidence intervals:

```
|| confint(crimeMod)
```

From this output, we are 95% confident that the effect of the violent crime rate in 1990 on 2000 is between 0.518 and 0.643.

■ **6.1.7 CONFIDENCE INTERVAL FOR Y** We can estimate the value of Y for a given value of x :

estimation

```
|| predict(crimeMod, newdata=data.frame(vcrime90=100), interval="confidence")
```

We are 95% confident that the expected value of Y when $x = 100$ is between 129.5 and 205.6, with a point estimate of 167.6.

■ **6.1.8 PREDICTION INTERVAL FOR Y** Finally, we can predict the value of Y for a new value of x :

prediction

```
|| predict(crimeMod, newdata=data.frame(vcrime90=100), interval="prediction")
```

We are 95% sure that the next *observation* of the violent crime rate in 2000 for a state with a violent crime rate in 1990 of 100 is between -8.5 and 343.7 , with a prediction of 167.6.

6.2: Full Example: Violent Crime, Wealth, Region

That was fun! Let's now try this with two independent variables. We will model the violent crime rate in 2000 using the GSP per capita in 1990 *and* the region of the state. This will give us the opportunity to reiterate and emphasize that these methods are not constrained to numeric independent variables. As in Example 3.1.3 on page 53, we can represent categorical independent variables appropriately and model using ordinary least squares estimation.

The following creates the interaction model between a numeric and a categorical variable. This particular type of interaction analysis is referred to as the Analysis of Covariance, ANCOVA:

interaction

```
|| modEd1 = lm(vcrime00 ~ gspcap90 * census9)
|| summary.aov(modEd1)
```

The interaction model allows for the effect to vary between the levels. In terms of this problem, the interaction model allows the effect of the 1990 violent crime rate on the 2000 to be different for the Midwest, the Northeast, the South, and the West.

It does not *force* it to be different. It only allows it to be.

Because of the writings of a 14th-Century monk by the name of William of Ockham, there is a bias in science to create models that are as simple as possible, without being too simple (his doctrine of efficient reasoning).¹

Occam's Razor

Non sunt multiplicanda entia sine necessitate.

The usual translation is “Things are not to be multiplied without necessity.” In other words, simpler models tend to be more helpful than complicated ones. Realize that they are more “helpful” and not more “correct.” To drive this point home, allow me to quote George E. P. Box (1976):

Since all models are wrong the scientist cannot obtain a “correct” one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

The results from the above code are given here

```
|| Df Sum Sq Mean Sq F value Pr(>F)
```

¹Note, however, that this doctrine/belief did not originate with William. It goes back to — at least — Aristotle in his *Posterior Analytics*: “We may assume the superiority *ceteris paribus* of the demonstration which derives from fewer postulates or hypotheses.”


```

|| gspcap90          1  816163  816163  26.149  1.32e-05  ***
|| census9          8  794820  99353   3.183   0.0087  **
|| gspcap90:census9 8  273868  34233   1.097   0.3901
|| Residuals       33 1029986  31212

```

Note that the p-value (last column) for the interaction term is greater than our usual $\alpha = 0.05$ (p-value = 0.3901). This tells us that the interaction term is not statistically significant. In other words, we can remove it from our model without adversely affecting our model.

When the model has no interaction terms, it is called an additive model. The following code fits the additive model.

additive

```

|| modEd2 = lm(vcrime00 ~ gspcap90 + census9)

```

As usual, the next step is to test the assumptions. Note that this process is the same, even if the particulars differ. The `census9` variable is categorical, not numeric. That requires we think a bit more about how to perform the assumption testing.

6.2.1 NORMALITY Again, the first assumption to test is the Normality of the residuals:

```

|| e = residuals(modEd2)
||
|| # Normality checking
|| overlay(e)
|| shapiroTest(e)

```

According to the Shapiro-Wilk test there is no significant evidence that the residuals come from a non-Normal distribution (p-value = 0.1732). Thus, the model passes this test.

6.2.2 CONSTANT EXPECTED VALUE (FUNCTIONAL FORM) The second requirement we test is that the expected value of the residuals is constant against *each of* the independent variables.

```

|| # Expected Value
|| plot(gspcap90, e)
|| runs.test(e, order=gspcap90)

```

The runs test indicates that there is no evidence the expected values are not constant (p-value = 0.2038). Thus, this test is passed, too. Yippee!

But wait! This only tested for constant expected value of the residuals against one of the two independent variables. It is required that the expected value is constant against *all* of them.

Here is the problem with just using the runs test when the independent variable is categorical: The ordering within each level is not uniquely defined. For a hint² on what to do, look at the graphic:

```
|| plot(census9, e)
```

Holy side-by-side box-and-whiskers plot, Batman! This makes sense, because we are plotting a numeric variable (e) across nine levels. We need to test that the expected value (mean) is the same in each group.

From your prior statistics class, this screams ANOVA!

```
|| summary(aov(e ~ census9))
```

The p-value returned is 1, which is greater than α . So we fail to reject the null hypothesis of equal expected values.

Well, this result should not be surprising. Because of the mathematics of OLS, the means in each group will be centered at zero. Thus, you should expect p-values of 1 whenever doing this test.

6.2.3 CONSTANT VARIANCE The last requirement is that the residuals have a constant variance against each of the independent variables. For the numeric variable, this is not a problem:

```
|| # Heteroskedasticity
|| plot(gspcap90, e)
|| hetero.test(e, gspcap90)
```

With a p-value of 0.7008, this test is passed for the numeric independent variable.

²When not sure what to do, plotting things frequently helps. It seems to force the researcher into determining what needs to be examined. When in doubt: Graph and Interpret.

For the categorical variable, remember that we need to test equality of variance *across several groups*. From your previous statistics course, you may recall that the Fligner-Killeen test does just this:

```
|| plot(census9, e)
|| fligner.test(e ~ census9)
```

According to the Fligner-Killeen test, there is no evidence of heteroskedasticity (p-value = 0.7869).

Thus, this model passes the homoskedastic requirement. Vahooo!

Note: We could have taken care of both tests for heteroskedasticity using the Breusch-Pagan test. However, if the model fails that test, we have no clue as to how to fix it. Breaking it into two parts allows us the additional information of which variable caused the issues.

Personally, I run the Breusch-Pagan test to determine *if* there is a violation, then the separate tests to get information on *where* the issue lies.

6.2.4 THE FINAL MODEL This model is appropriate, and we can now see the estimates. However, if we are in the ‘model creation’ or ‘model selection’ mode, we need to determine if both variables are statistically significant. If either is not, then that variable needs to be dropped and the new model tested.

To get p-values for the variables, just run

```
|| summary.aov(modEd2)
```

Yeppers, that is `summary.aov` that you are using. It provides statistical significance of the *variables*, while `summary` and `summary.lm` provide the statistical significance of the levels of the categorical variables.

The output from the `summary.aov(modEd2)` command is

```
||      Df  Sum Sq Mean Sq F value    Pr(>F)
|| gspcap90    1  816163   816163  25.664 9.07e-06 ***
|| census9     8  794820    99353   3.124 0.00753 **
|| Residuals  41 1303854    31801
```

The p-value for the `gspcap90` variable is less than alpha, so that variable has a significant effect on the violent crime rate. The p-value for the `census9` variable is also less than alpha, so it too needs to be included in the final model.

In short, this is the model we need to use. The abbreviated regression table from this model is

```
||| Coefficients:
      Estimate      Pr(>|t|)
(Intercept)  1.125e+02    0.308
gspcap90     1.550e-02    6.7e-05 ***
census9East South Central  7.518e+01    0.535
census9Middle Atlantic  -6.777e+01    0.608
census9Mountain        -1.707e+02    0.101
census9New England     -1.708e+02    0.122
census9Pacific         -1.316e+02    0.263
census9South Atlantic   1.590e+02    0.125
census9West North Central -4.953e+01    0.638
census9West South Central  1.197e+02    0.323
```

From this, we can conclude that there is a statistically significant, and positive (!), effect of average state wealth on the violent crime rate. We get that conclusion from the `gspcap90` line in the table.

base category

The rest of the table compares the effect of each level of the `census9` variable to the base category, East North Central. As no p-values is less than alpha, we can conclude that none of the regions is statistically different in its effect from the East North Central region.

What about when compared to the Mountain region?

First, we have to specify that we want the Mountain region to be the base category against which everything else is calculated. Then, we need to re-fit the model with the new base.

```
||| census9 = set.base(census9, "Mountain")
      modEd3 = lm(vcrime00 ~ gspcap90 + census9)
      summary(modEd3)
```

The regression table now indicates that the violent crime rate in the Mountain region is significantly lower than that in the East South Central, South Atlantic, and West South Central regions.

How does the violent crime rate in the different regions compare to the Pacific region?

```
|| census9 = set.base(census9, "Pacific")
|| modEd3 = lm(vcrime00 ~ gspcap90 + census9)
|| summary(modEd3)
```

The violent crime rate in the Pacific region is significantly lower than that in the South Atlantic and the West South Central regions.

How do the regions compare to the South Atlantic region?

```
|| census9 = set.base(census9, "South Atlantic")
|| modEd3 = lm(vcrime00 ~ gspcap90 + census9)
|| summary(modEd3)
```

The violent crime rate in the South Atlantic region is significantly higher than in the Mountain, New England, Pacific, and West South Central regions.

Note: Be aware of the multiple comparisons issue (see Section S.6.2). Remember that these individual analyses only work if you perform only one of them. Multiple comparisons require adjustment of the alpha-level. For a reminder, see Appendix Section S.6.2.

6.2.5 THE GRAPHIC The following code generates the graphic at the top of the next page:

```
|| par(mar=c(4,4,0,1)+0.5, family="serif", las=1)
|| par(xaxs="i", yaxs="i")
|| par(cex.lab=1.2, font.lab=2)
||
|| plot.new()
|| plot.window( xlim=c(0,75), ylim=c(0,2500) )
||
|| axis(1); axis(2)
||
|| title(xlab="GSP per Capita (1990) [$000]", line=2.75)
|| title(ylab="Violent Crime Rate (2000)", line=3.5)
||
|| xx = seq(15,70)*1000
|| yyPac = predict(modEd2, newdata=data.frame(gspcap90=xx, census9
||       ="Pacific"))
|| yyMtn = predict(modEd2, newdata=data.frame(gspcap90=xx, census9
||       ="Mountain"))
```

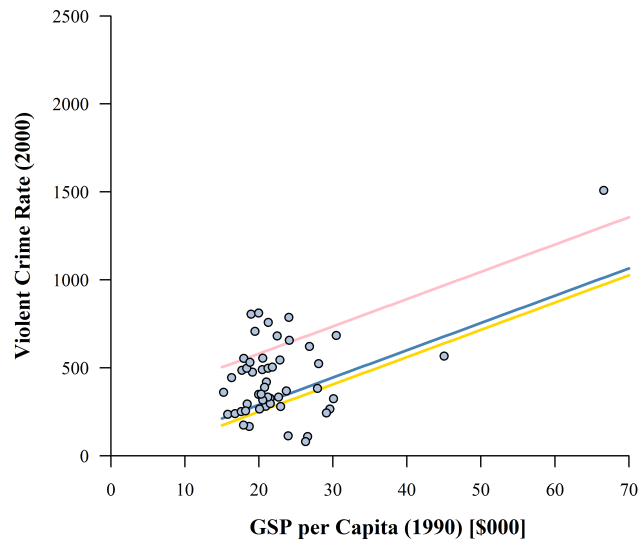


Figure 6.2: Plot of the violent crime rate in 2000 against the GSP per capita in 1990 (in thousands of dollars). The ordinary least squares line of best fit is included. The red-colored line is the estimate for the South Atlantic states; blue, Pacific states; and gold, Mountain states.

```

yySat = predict(modEd2, newdata=data.frame(gspcap90=xx, census9
      ="South Atlantic"))

lines(xx/1000,yyPac, col="steelblue", lwd=2)
lines(xx/1000,yyMtn, col="gold", lwd=2)
lines(xx/1000,yySat, col="pink", lwd=2)

points(gspcap90/1000,vcrime00, pch=21, bg="lightsteelblue")

```

It would be helpful to have a legend, but let us leave that for another day!

Also, you need to be able to determine what each line of this script does.

6.2.6 CONFIDENCE INTERVAL FOR SLOPE We can obtain confidence intervals for the effects using the same method as before. The interpretation follows the same rules, but the table is *much* bigger:

```
|| confint(modEd2)
```

We are 95% confident that the effect of the GSP per capita on the violent crime rate is between 8 and 23 additional violent crimes (per 100,000 population) for every \$10,000 increase in GSP per capita.

The confidence intervals for the effects of each of the levels is *as compared to* the base level. Thus, we are 95% confident that the average violent crime rate in the West North Central region is between 21 and 396 lower than that in the South Atlantic region (the base level).

6.2.7 CONFIDENCE INTERVAL FOR \hat{y} Again, we can estimate the expected value of a violent crime rate, given the GSP per capita and the region.

```
|| predict(modEd2, newdata=data.frame(gspcap90=50000, census9="
|| Pacific"), interval="confidence")
```

We are 95% confident that the expected violent crime rate in a Pacific-region state with a GSP per capita of \$50,000 is between 537 and 975, with a point estimate of 756 violent crimes per 100,000 people.

6.2.8 PREDICTION INTERVAL FOR y_{new} Finally, we can also calculate a prediction interval for a new observation:

```
|| predict(modEd2, newdata=data.frame(gspcap90=50000, census9="
|| Pacific"), interval="prediction")
```

We are 95% sure that the violent crime rate for a new observation of a Pacific-region state with a GSP per capita of \$50,000 is between 334 and 1177, with a best guess of 756 violent crimes per 100,000 people.

Note: As is always the case, the width of the prediction interval is larger than the width of the confidence interval. Remember why this is the case.

6.3: Full Example: Cows in the City of Děčín

To illustrate the process of model selection, let us examine Děčín's ballot measure of 2009. That ballot measure sought to constitutionally restrict the number of cows that can be housed within the city limits. While this extended example seems rather dated, it does cover some interesting issues in statistical modeling and questions we can answer with our model.

Example 1

The voters of Děčín are being sent to the polls to vote on a constitutional referendum that proposes to limit the number of cows that could be housed within the city limits. This was not the first time that Ruritanians were sent to the polls to vote on this or a closely related issue. Given the information from previous votes, and the demographics of Děčín voters, what is the probability that this ballot measure will pass?

Before attempting any analysis, there needs to be a search of the literature to inform us as to which variables should be present, and which directions those variables should affect the dependent variable. From that literature review, we hypothesize that the vote in favor of such ballot measures depends on three variables: age of the population, religiosity of the population, and whether the ballot measure also restricts chickens. The effect direction for each is that kraj that are more religious should vote against cow-housing at a higher rate; Measures that also ban chickens should have a harder time passing; Measures passed more recently should have a more difficult chance of passing, as the young tend to support cows, and the elderly tend to oppose them (wanting quiet, dung-free neighborhoods).

directional hypotheses

With this theory and the resulting hypotheses, we can take our next step: Getting to know the data.

	Year Passed	Chicken Ban	Religious Percent
Minimum	1998	0	51.00
Maximum	2008	1	85.00
Median	2004	1	67.50
Mean	2004	0.5938	66.75
Variance	6.0650	0.2490	88.1935
Coefficient of Variation	0.5794	0.8404	0.1407

Table 6.1: Descriptive statistics on the variables in the *cows* dataset.

6.3.1 GET TO KNOW THE DATA Before we begin trying to answer this question, we must get to know our data. There are several functions available to us to visualize the data: histogram, scatter plots, and quantile-quantile plots. In addition to visualizing the data, we should calculate several of the descriptive statistics for the variables of interest.

```
source("http://rfs.kvasaheim.com/rfs.R")

cows = read.csv("http://rur.kvasaheim.com/data/cows.csv")
summary(cows)
```

VARIABILITY: Since we have multiple independent variables, we should calculate both univariate and bivariate descriptive statistics. Table 6.1 provides the univariate descriptive statistics. The primary univariate question to ask about the independent variables here is whether there is sufficient variation. The two measures we need to examine are the variance and the coefficient of variation. If both of these numbers are small, then there may be an issue.

variation

In this data, the variance of the Chicken Ban variable is small and potentially worrisome; however, its coefficient of variation (a scaled standard deviation,

coefficient of variation

$$c_v = \left| \frac{s}{\bar{x}} \right| \quad (6.3)$$

indicates that there is no serious issue (the value is close to 1).³ None of the three variables have small enough variation to cause us concern.

³As this is a dichotomous variable, the mean is the percent of the values equal to 1. Thus, there are about 60% of the values 1 and 40% of the values 0 — more than sufficient variation.

	Year Passed	Chicken Ban	Religious Percent
Year Passed		0.1903	0.2399
Chicken Ban	0.1903		0.5146
Religious Percent	0.2399	0.5146	

Table 6.2: The correlations between the variables in the `cows` data. The correlation between Chicken Ban and Percent Religious is statistically significant ($t = 3.2869; \nu = 30; p = 0.0026$). This is the sole statistically significant correlation.

RELATIONSHIPS: After getting to know the variables individually, it is important to get to know the relationships between the variables. This can be done through correlation tests and bivariate scatter plots. Independent variables with strong correlations with the dependent variable should be considered for inclusion in the model. Independent variables with strong correlations with other independent variables should be of concern. Remember that one of the assumptions of OLS regression is that the independent variables are statistically independent of each other. If independent variables are highly correlated, the statistical properties of the method weaken.

correlated

Question

If independent variables are highly correlated, the statistical properties of the method weaken. Why?

The pairwise correlations are provided in Table 6.2. Of the three independent variables, only Chicken Ban and Religious Percent have a statistically significant correlation ($t = 3.2869; \nu = 30; p = 0.0026$). Should the level of correlation be a concern? Perhaps. While their correlation is $r = 0.5146$, this corresponds to an R^2 value of just 0.2648. As such, the correlation may not be large enough to severely affect our coefficient estimates (see Sections ?? and ??). Let us just remember this relationship for the future.

Note: The issue is actually more than a statistics issue. If two independent variables are highly correlated with each other, it is logically impossible to determine *which* affects the dependent variable or how much of the effect to partition to each independent variable. Statistics is, however, able to tease out the independent relationships better than not. As a rule of thumb, if the correlation is greater than $r = 0.90$, there may be a serious logical issue. If two variables are so highly correlated, which of the two is the “correct” inde-

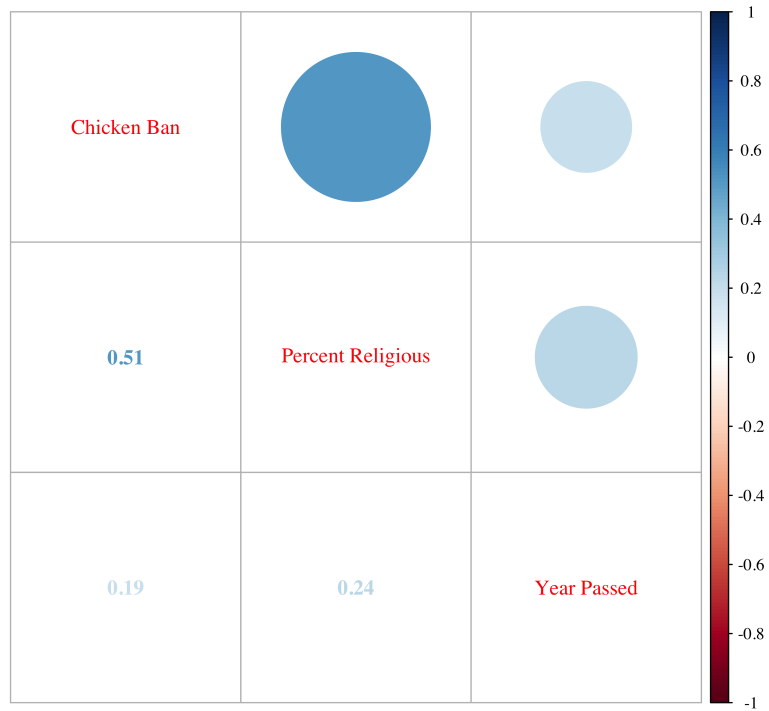


Figure 6.3: Correlation plots between the three independent variables. The correlation between Civil Ban and Percent Religious was statistically significant according to the Pearson product-moment correlation test. This is evident in this graph, as well.

pendent variable? How can one tell? Can both be good? Is the commonality between them the real independent variable?

6.3.2 VARIANCE INFLATION FACTOR This problem is a bit more extensive than suggested above. Recall that one of the mathematical requirements is that the rank of the design matrix equals p (the number of parameters to be estimated). This can happen if one variable is perfectly correlated with (linear function of) another variable. It can also happen if a variable is a linear function of the other variables.

Thus, while checking the bivariate correlations is helpful, it is not the answer we need. We need to check if the independent variable is a linear function of (or close to) a combination of the others.

To determine the level of multicollinearity we can use the “variance inflation factor” (VIF). Recall from Section 5.4.2 how to calculate the VIF for a given independent variable: Regress all other independent variables on it, calculate the R_i^2 , and calculate the VIF from

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \quad (6.4)$$

Once you have calculated the VIF for each independent variable, compare the values to the “usual” Rule of Thumb.⁴

There is nothing magical about using the R^2 value to calculate your variance inflation factor. One could advocate using the adjusted- R^2 . However, in doing so, the Rules of Thumb may need to be adjusted.

Of course, if we think about the effects of multicollinearity, rather than just detecting “severe levels,” then we may wish to eschew such tests, assume multicollinearity is an issue, and adjust for it. On the other hand, if there is no reason to think that the model should suffer from multicollinearity, we will want to avoid such adjustments.

These are science questions, not statistics questions.

Know the science behind your theory.

⁴This Rule of Thumb depends on the discipline. The three typical boundaries are 5, 8, and 10. If all of your VIF scores are greater than 10, then there is an issue with multicollinearity. If all are less than 5, then there is no issue. If it is between those extremes, then you should think about the effect of multicollinearity on your estimators. What do those VIF values correspond to? A VIF of 5 means the R^2 value is 0.80. A VIF of 10 means the R^2 is 0.90. Keep that in mind. In other words, the *other* independent variables explain 90% of the variation in this independent variable.

~	Separates the dependent variable (left-hand side) and the independent variables (right-hand side)
+	Indicates the following variable is added to the formula
-	Indicates the following variable is removed from the formula
:	Indicates the following and the preceding variable are multiplied in the formula
*	Indicates the following and the preceding variable are crossed in the formula
^	Includes the specified level of interactions
I ()	Replaces the formula grammar of what is in the parentheses with algebraic grammar

Table 6.3: *The symbols and their meanings in the grammar of formulas. I sure wish I could locate the book that created these, but I cannot find it anymore. It is in the Oklahoma State University library... somewhere.*

6.3.3 MODEL THE DATA The example asked us to determine the probability that the ballot measure will pass. Before we can answer that question, we need to model the proportion of the vote in favor of the ballot measure using our independent variables; that is, we need to be able to predict the proportion of the vote in favor of the ballot measure with the information we have.

prediction

Thus, the dependent variable will be `propWin` and the independent variables will be `yearPassed`, `chickens`, and `religPct`. For now, let us assume a linear relationship between the independent variables and the dependent variable.

MODEL SELECTION: Unless you have a lot of independent variables, I recommend you start with the interaction model.⁵ The interaction model includes the effects of each independent variable singly (main effects) as well as all possible combinations of those variables (interaction effects).

interaction model

R uses the usual formula grammar (Table 6.3). Its use takes a little practice

grammar

⁵Some will disagree and recommend starting with the simplest model and building complexity from that. There tends to be little difference between the two model-building methods. On either case, one has to worry about the multiple comparisons issue (Appendix S.6.2). How we should address it in the realm of model building is still unknown. We are certain of two things, however. First, the Bonferroni procedure is far too conservative. Second, doing nothing is not an acceptable option.

to get the hang of, but it is entirely logical (for the most part). For instance, if you wish to fit the model $y = \beta_0 + \beta_1 x + \varepsilon$, you would use $y \sim x$.

However, if you wish to force the y-intercept to be 0, that is to fit the model $y = \beta_1 x + \varepsilon$, then you have a choice:

$$y \sim x - 1$$

$$y \sim x + 0$$

The first is logically interpreted as the usual model, “less the intercept term” (hence the $- 1$). This is the usual method I use. The second exists for backward compatibility. **I strongly encourage you to use the first method.**

Some other examples of this grammar include:

Algebraic form	Formula form
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$y \sim x_1 * x_2$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$y \sim x_1 + x_2 + x_1 : x_2$
$y = \beta_0 + \beta_1 x_1 x_2$	$y \sim x_1 : x_2$
$y = \beta_0 + \beta_1 x_1^3 + \beta_2 \sin(x_2)$	$y \sim I(x_1 \wedge 3) + I(\sin(x_2))$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3$	$y \sim x_1 * x_2 * x_3$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$y \sim (x_1 + x_2) \wedge 2$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$	$y \sim (x_1 + x_2 + x_3) \wedge 2$

Note that the first two examples above are the same model. These illustrate what comprises a “crossed” model. Also note that the “wedge” operator, \wedge , indicates how many variables are multiplied in the terms. So $\wedge 2$ indicates you want all one-way and two-way interactions for the model, and $\wedge 3$ means you want all one-, two-, and three-way interactions.

Thus, we can see the following pairs are equivalent:

$$y \sim (x_1 + x_2 + x_3) \wedge 3 \text{ and } y \sim (x_1 * x_2 * x_3)$$

$$y \sim (x_1 + x_2 + x_3) \wedge 2 \text{ and } y \sim (x_1 * x_2 * x_3) - x_1 : x_2 : x_3$$

With this brief introduction to the grammar of formulas, we can return to our example. We have three independent variables; the formula to give a full three-way interaction model is

asterisk

```
propWin ~ yearPassed * chickens * religPct
```

As we will use this model a bit, we save the linear regression results into a variable. Thus, the two lines to run are

```
mod1 = lm(propWin ~ yearPassed * chickens * religPct)
summary(mod1)
```

These lines give the following output (well the first, fourth, and fifth column of that output):

	t value	Pr(> t)
(Intercept)	1.148	0.262
yearPassed	-0.901	0.377

chickens	-1.084	0.289
religPct	1.557	0.133
yearPassed:chickens	0.950	0.352
yearPassed:religPct	0.510	0.615
chickens:religPct	0.979	0.338
yearPassed:chickens:religPct	-0.895	0.379

three-way

Occam

The line starting `yearPassed:chickens:religPct` is the three-way interaction term. As it is the highest interaction, it is the *only one* we can interpret here. Note that it is not statistically significant ($p = 0.379$). Thus, removing that term will do two things. First, it will simplify the model. Second, it will not significantly harm the model's descriptive (or predictive) ability.

That second model can be written as either

```
|| mod2 = lm(propWin ~ yearPassed * chickens * religPct -  
||     yearPassed:chickens:religPct)
```

or as

```
|| mod2 = lm( propWin ~ (yearPassed + chickens + religPct)^2 )
```

The two formulas are equivalent.

Note that the `summary.aov(mod2)` command indicates that none of the three two-way interactions are statistically significant. Thus, these two-way interactions should be removed from the model.⁶ This leaves a model with no interactions—an additive model. Fitting the additive model and checking the statistical significance of the variables is as above

```
|| mod3 = lm(propWin ~ yearPassed + chickens + religPct)  
||     summary.aov(mod3)
```

Note that all three variables are significant according to this output (the chicken variable is statistically significant because we specified an effect direction). Thus, this is our provisional model.

formula grammar

two-ways

additive model

provisional model

Question

Given that the additive model is the appropriate model, does the effect of religiosity (`religPct`) change over time?

Question

Given the analysis above, is there evidence that the effect of religiosity (`religPct`) changes over time?

⁶Again, some would alternatively advocate removing just the least significant effect, then refit the new model. Others would suggest refitting with three different models, one for each combination of interaction. There is no “always best” answer, other than the one that your science suggests.

THE ADDITIVE MODEL: That is, the equation we will use to fit the data is

$$\text{propWin} = \beta_0 + \beta_1(\text{yearPassed}) + \beta_2(\text{chickens}) + \beta_3(\text{religPct}) + \varepsilon \quad (6.5)$$

If $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then we know

$$\mathbb{E}[\text{propWin}] = \beta_0 + \beta_1(\text{yearPassed}) + \beta_2(\text{chickens}) + \beta_3(\text{religPct}) \quad (6.6)$$

CHECK THE ASSUMPTIONS: But, does this model violate any of the assumptions of OLS regression? All of the usual tests (Shapiro-Wilk, Breusch-Pagan, and runs) pass.

What about multicollinearity? Remember that the effect of multicollinearity is to inflate the standard errors (reduce the t-value, increase the p-value). Thus, if multicollinearity exists, fixing it will make the variables even more statistically significant.

How do we test it? We can do it the hard way or the easy way. The hard way is to estimate three regression equations, calculate the individual R^2 values, and calculate the resulting VIF values.

```
|| vif1 = lm(yearPassed ~ chickens + religPct)
|| vif2 = lm(chickens ~ yearPassed + religPct)
|| vif3 = lm(religPct ~ yearPassed + chickens)
||
|| 1/(1-summary(vif1)$r.squared)
|| 1/(1-summary(vif2)$r.squared)
|| 1/(1-summary(vif3)$r.squared)
```

Or, you can use the `car` package:

```
|| library(car)
|| vif(mod3)
```

The results of these VIF checks are

```
|| yearPassed  chickens  religPct
|| 1.067965    1.368963    1.399954
```

None of these three are even close to the lowest “Rule of Thumb.” As such, multicollinearity is not an issue in this model.

	Estimate	Std. Error	t-value	p-value
Constant Term	0.1512	0.0659	2.293	0.0295
Year Passed (post 2000)	-0.0201	0.0036	-5.618	$\ll 0.0001$
Banned Chickens	-0.0373	0.0200	-1.868	0.0723
Percent Religious	0.0095	0.0011	8.801	$\ll 0.0001$

Table 6.4: Results table for the regression of proportion support of a generic ballot limiting the number of cows housed in the city against the three included variables. The R^2 for the model is 0.7801; the \bar{R}^2 , 0.7565. The p-values calculated are based on two-tailed test. The hypotheses were one-tailed hypotheses. As such, all three explanatory variables are statistically significant at the standard level of significance ($\alpha = 0.05$).

6.3.4 RESULTS The regression table for model `mod3` is given in Table 6.4. Notice that all three variables of interest are statistically significant at the $\alpha = 0.05$ level.⁷ Additionally, the model has an \bar{R}^2 of 0.7565, which is a great fit in most of the social sciences. The direction of the coefficients also agrees with theory.

Thus, the equation for the line of best fit is approximately

$$\begin{aligned} \mathbb{E}[\text{propWin}] = & 0.1512 \\ & - 0.0201(\text{yearPassed}) \\ & - 0.0373(\text{chickens}) \\ & + 0.0095(\text{religPct}) \end{aligned} \quad (6.7)$$

From this, we can make the following observations. First, the expected vote proportion is declining as time passes by about 2 percentage points per year. Second, ballot measures that also ban chickens are less likely to pass. Finally, the more religious a kraj, the more likely a cow ban is to pass.

Question

What does the 0.1512 represent in this context (in real terms)?

⁷You may claim that the Chicken variable is not statistically significant at the $\alpha = 0.05$ level. However, the provided p-values are two-tailed (non-directional) p-values. Our hypotheses were all directional hypotheses (one-tailed). Thus, to get the one-tailed p-values just halve the two-tailed p-values. With that, all three independent variables are statistically significant.

directional hypothesis

research hypotheses

prediction line

6.3.5 PREDICTING DĚČÍN According to this model, what is the expected vote in Děčín? To answer this, we need information about the Děčín ballot measure, specifically the value of the independent variables:

```
yearPassed = 9,  
chickens = 0, and  
religPct = 48.
```

With this information, and under the assumption that the model is correct, we have our prediction that 42% of the Děčín voters will vote in favor of this ballot measure.

Thankfully, R does not require us to do this calculation by hand. The R code for predicting the percent of Děčín voters voting in favor of this ballot measure can be

```
DECIN = data.frame(yearPassed=9, chickens=0, religPct=48)  
predict(mod3, newdata=DECIN)
```

The first line was used to make the code more readable. It is also helpful to first define the variable `DECIN` if you are going to make predictions for Děčín using several models.

If neither of these appeal to you and you wish to do this in one line, that line would be

```
predict(mod3, newdata=data.frame(yearPassed=9, chickens=0,  
religPct=48))
```

predict

Note the inclusion of the `predict` function, which predicts the dependent variable value given values for each of the independent variables (read the help file on `predict`; we will use this function frequently).

Question

How does the estimate change if you specify more variables than are in the model for Decin? For instance:

```
DECIN = data.frame(yearPassed=9, chickens=0, religPct  
=48, age=45.2, channels=12, turnout=0.72)
```

6.3.6 GRAPHING THE RESULTS Now that we have confidence in our model, we can use it to predict the effects of each of the three independent variables on the vote in favor of these ballot measures. There are three independent variables, so we cannot create a single graph that displays the results. However, as one of the variables is dichotomous, we can show the results in just two graphs (the number of continuous independent variables).

Both of these graphs will have the vote in favor as the dependent variable (vertical axis). One of the two graphs will have percent religious as the primary independent variable, whereas the other will have the year passed as the primary independent variable. The chicken variable will be present in both graphs, signified by two separate curves, one where the ballot measure banned chickens and one where it did not (Figure 6.4).

The graphs illustrate the results of the model — this is their purpose. Although the graphs “illustrate the story,” we must still “tell the story” of the graphics, including numbers from the prediction table (Table 6.4). The following paragraphs explain the graphics.

Both graphics show that the effect of adding a chicken ban to the referendum tends to reduce the vote in favor of the referendum. All things being equal, a ballot measure banning chickens will have 3.7% fewer people vote for it than a like measure not banning chickens ($s = 1.9988, t = -1.87, p = 0.0723$).

The top graphic illustrates the effect of passing time on the proportion of the vote in favor of these referenda: As the year increases by one, the proportion voting in favor of the referendum decreases by 2% on average ($t = -5.62, p \ll 0.0001$).

The bottom graphic shows the effect of religiosity on the ballot outcome: those kraj with higher levels of religiosity tend to vote in favor of these measures at a higher level than kraj with lower levels of religiosity. In fact, increasing the level of religiosity in the kraj by 1% will tend to increase the vote in favor of the ballot measure by 0.95% ($t = 8.80, p \ll 0.0001$).

Note the interweaving of the graphic discussion with concrete, numerical effects (and statistical significance in parentheses) from the prediction table. This combination aids the reader in interpreting the graphic(s) in terms of statistical language.

tell the story

regression table

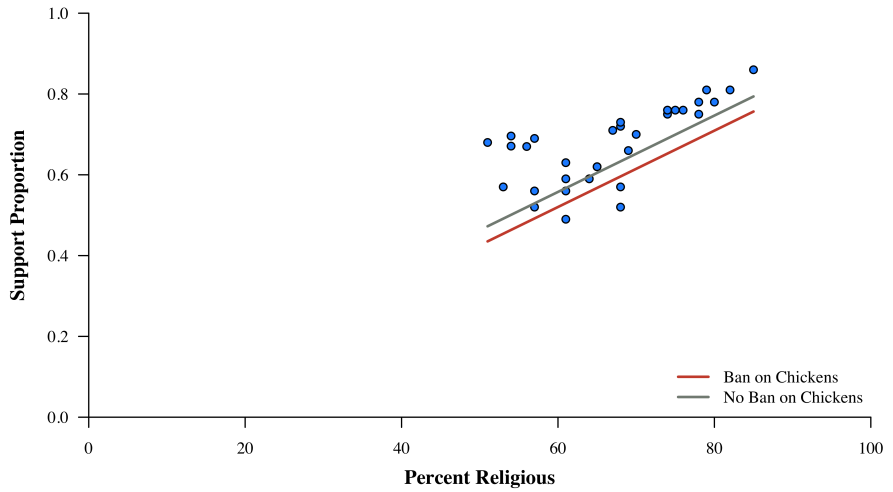
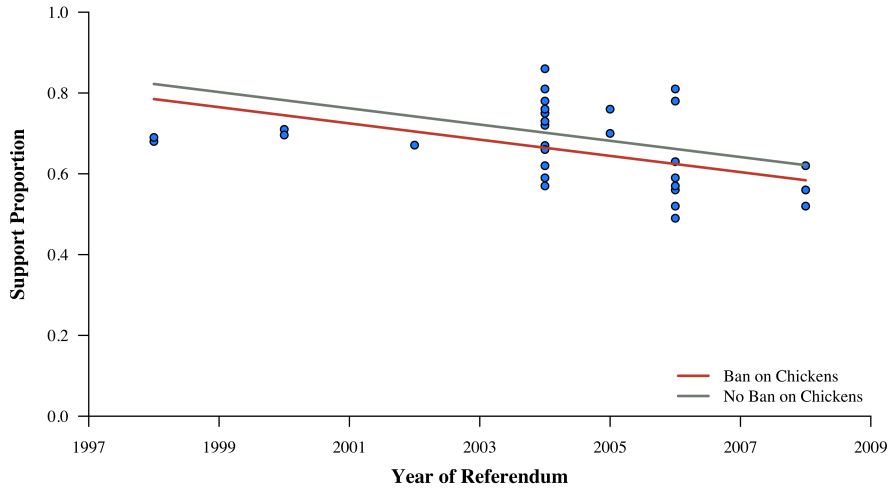


Figure 6.4: Prediction graphs of our *cows* model. These graphs contain two independent variables plotted against the dependent variable, with the dichotomous independent variable included as separate lines. Note that the effect of each of the three independent variables is made manifest by these two graphs.

6.3.7 ANSWERING THE QUESTION WITH THE PARAMETRIC BOOTSTRAP* Thus, we have a prediction of 42% of the voters will support the ballot measure. However, this is *not* an answer to the original question, which asked about the *probability* of the ballot measure passing. From a modeling standpoint, this probability depends on the coefficient estimates, which are just estimates of the true population value, and the standard errors, which are measures of our certainty in those estimates.

point prediction

In the ordinary least squares method, those parameter estimates are random variables, since they are *functions of the data*. In other words, if we re-ran human history, the estimated effect would be different, since reality would be different. Furthermore, as these are random variables, they have an associated distribution — the normal distribution. In fact, the distribution of each parameter estimate is normal, with expected value equal to the estimate and standard deviation equal to the standard error. Thus, for example, the effect of `yearPassed` is $\hat{\beta}_1 \sim \mathcal{N}(\mu = -0.0201, \sigma = 0.0036)$; of `chicken`, $\hat{\beta}_2 \sim \mathcal{N}(\mu = -0.0373, \sigma = 0.0200)$; and of `pctRelig`, $\hat{\beta}_3 \sim \mathcal{N}(\mu = 0.0095, \sigma = 0.0011)$.

random variable

distribution of estimators

Let us leverage these facts to (virtually) re-run human history multiple times, get the parameter estimates for each history, and predict the outcome of the ballot measure in Děčín.⁸ In other words, let us perform a Monte Carlo analysis. The steps are the same as with any Monte Carlo analysis we have done (Kennedy 2008). The only difference is what we do within the loop. Here, we draw random numbers from the appropriate distribution and calculate the predicted vote.

Monte Carlo

Before you look at the following algorithm, write your own and compare it to the one below:

1. Initialize variables
2. Perform loop
 - a) Draw from the four distributions
 - b) Predict the Děčín outcome
3. Calculate the proportion of times the ballot measure garnered more than 50% of the vote

One can also store the random numbers inside the loop and predict outside the loop. Also, if the statistical program allows it, you can avoid the loop and just draw all the numbers at once. This last has the advantage of being *very* fast.

It is also the method I use here, in the R script:

⁸Note that this process assumes the parameter estimates are independent of each other. This is not the case. See Theorem 3.1.3. The effects are dependent on each other, as is the intercept. As such, treat this sub-section as a **pedagogical exercise** rather than a statistical exercise. There are a lot of questions dealt with here that help better understand things.

```

# Initialize variables
outcome <- numeric()
trials <- 1000000

# Coefficient estimates
b.intc <- 0.151221
b.year <- -0.020095
b.cban <- -0.037331
b.rpct <- 0.009452

# Coefficient standard errors
s.intc <- 0.065938
s.year <- 0.003577
s.cban <- 0.019988
s.rpct <- 0.001074

# Distributions (the "loop")
e.intc <- rnorm(trials, m=b.intc, s=s.intc)
e.year <- rnorm(trials, m=b.year, s=s.year)
e.cban <- rnorm(trials, m=b.cban, s=s.cban)
e.rpct <- rnorm(trials, m=b.rpct, s=s.rpct)
outcome <- e.intc + e.year*9 + e.cban*0 + e.rpct*48

```

At this point, the variable `outcome` holds the proportion of people voting in favor of the ballot measure in one million simulated elections. To answer the question, we just need to determine the proportion of those elections in which the `outcome` is greater than 0.50: `mean(outcome>0.50)` will work.

Of course the numbers are nice, but a histogram may tell a better story. The following code will give a histogram similar that in Figure 6.5.

```

hist(outcome, main="", xlab="Proportion Vote for Ballot Measure",
     breaks=-1:99/100)
hist(outcome[outcome>0.50], main="", yaxt="n", breaks=-1:99/100,
     col=2, add=TRUE)
axis(1, at=0.50, labels="50%")

```

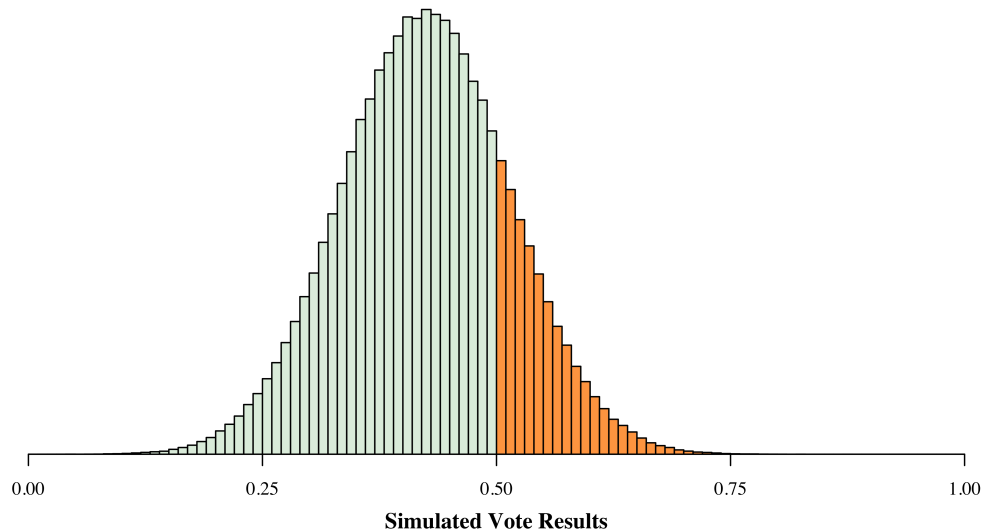



Figure 6.5: Plot of the predicted vote outcomes from the Monte Carlo experiment described in the text. Note that, while the expected proportion of the vote in favor of the ballot measure is 42%, there is still a 20% chance of the ballot measure passing, given that our model is correct.

The histogram of the Děčín predictions is presented in Figure 6.5. Note that the expected outcome is still 42%, which we found above, but that there is a spread to that prediction the histogram makes manifest, which the single prediction did not. In fact, prior to this analysis, we may have concluded that there was no possibility that the ballot measure would pass in Děčín based on our model; now, we see that there is a 20% chance of the ballot measure passing.⁹

confidence interval



Thus, we have an estimated answer to our original question. Given that our model is correct, there is approximately a 20% chance that the ballot measure to limit the number of cows in the city will pass in Děčín, with a point prediction of 42% in favor of the bill.

point estimate

The actual results of the 2009 ballot measure in Děčín was that the ballot measure passed with 53% of the vote. This result is well within the 95% prediction interval suggested by Figure 6.5. Also, the fact that the ballot measure passed should not be too surprising, since this model gave it a 20% probability of passing, and 20% is not a rare event by *any* stretch of the imagination.

⁹As with all statistical analysis, the *caveat* is that the model and the assumptions must be correct.

6.3.8 A FUNDAMENTAL PROBLEM There is a really big problem with these results, however. Run the following code and interpret it.

```
|| mean(outcome>1) + mean(outcome<0)
```

This is an important exercise: Always check that predictions make sense.

In a future chapter, we will revisit this issue and propose a solution.

6.4: Conclusion

In this chapter, we performed full analyses, demonstrating the entire process.



Warning: Again, be aware of the multiple comparisons issue discussed in Appendix S.6.2. It explains why you need to either adjust your p -value or your alpha level when performing multiple tests, such as when you are testing both $\beta_0 = 0$ and $\beta_1 = 0$.

6.5: End-of-Chapter Materials

6.5.1 R FUNCTIONS In this chapter, we were introduced to many, many, many R functions that will be useful in regression. In fact, this chapter uses more R functions than any other chapter in this book. Here are the many.

PACKAGES:

car This package provides several statistical tests used in the book “An R Companion to Applied Regression” by J. Fox and S. Weisberg. It is a great package that provides a lot of additional functionality for R.

lawstat This package provides several statistical tests used in law and public policy analysis. It provides the `runs.test` function for us.

lmtest This package provides many tests related to linear models. It provides an implementation of the Breusch-Pagan test, `bptest`, which tests for heteroskedasticity in the residuals.

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

STATISTICS:

source(filename) This function runs an R script from a separate file. That file may be local or on the Internet.

runs.test(E, order) This alteration to the `lawstat` function tests whether the variable `E`, as ordered by `order` exhibits fit issues.

shapiroTest(E) This tests the null hypothesis that the variable `E` comes from a Normal distribution. It is based on the `shapiro.test` function in the basic R installation. It adds capabilities to test Normality in several groups.

lm(formula) This is the function that performs ordinary least squares estimation on linear models.

bptest(mod) This function from the `lmtest` package performs the Breusch-Pagan test for heteroskedasticity.

confint(mod) This calculates confidence intervals for the parameters in ordinary least squares regression.

mean(x) This calculates the mean of a sample.

summary(x) This produces the six-number summary or a frequency table of the provided variable, depending on the type of variable.

summary.lm(mod) When applied to a linear model fit using either the `aov` function or the `lm` function, provides estimates of the effects of the numeric variables and the levels of the categorical variables in the model.

summary.aov(mod) When applied to a linear model fit using either the `aov` function or the `lm` function, provides estimates of the statistical significance of the variables in the model.

predict(mod) This predicts the values of the dependent variable at each point in the dataset *or* for the values specified.

fligner.test(formula) This tests for heteroskedasticity when the independent variable is categorical.

aov(formula) This function performs ordinary least squares estimation on linear models.

vif(model) This function calculates the variance inflation factor (VIF) for each of the independent variables in the model.

set.base(var,level) This `RFS` package function redefines the base category in the provided level. By default, the base category is the first according to the alphabet.

PROBABILITY:

set.seed(x) This sets the random number seed.

rexp(n, rate) This generates n random values from an Exponential distribution with the specified rate parameter.

rnorm(n, mean, sd) This generates n random values from a Normal distribution with specified mean and standard deviation. By default the mean is 0 and the standard deviation is 1.

runif(n, min, max) This generates n random values from a Uniform distribution with specified minimum and maximum values. By default, the minimum is 0 and the maximum is 1.

MATHEMATICS:

head(x) This returns the first six values in the variable.

foot(x) This returns the last six values in the variable.

seq(from, to, by, length) This returns a vector of sequential values, where `by` indicates the step size and `length` specifies the vector length.

length(x) This calculates the length of a vector (variable), which is the sample size, n .

residuals(mod) This calculates the residuals in the model, which is the difference between the observed and the predicted.

GRAPHICS:

qqnorm(x) This creates a Normal quantile-quantile plot for the given values.

qqline(x) This adds the diagonal line to the quantile-quantile plot.

overlay(x) This, from the `RFS` package, produces a histogram with a Normal curve overlaying it.

par(...) This sets parameters on the next graphic started. Look through the help page for this function to see all you can specify.

plot(x,y) This produces a scatter plot of the y-values against the x-values.

axis(side) When a plot is already drawn, this adds values along axis number `side`.

title(...) When a plot is already drawn, this adds the x- and y-labels.

lines(x,y) When a plot is already drawn, this draws lines between each subsequent (x, y) pair.

points(x,y) When a plot is already drawn, this draws points at each (x, y) pair.

PROGRAMMING:

attach(dataframe) This allows you to access the variables in the `dataframe` without having to prefix each with `dataframe$`.

library(package) This loads an external package that you have already installed on your computer. It allows access to all functions and data sets in the `package` package.

as.character(x) This changes the values in variable `x` to be characters.

as.numeric(x) This changes the values in variable `x` to be numbers.

6.5.2 EXERCISES

1. In the two panels in Figure 6.4, the lines of best fit do not go beyond the data. Why?
2. Section 6.3.8 mentioned that there was a really big problem with this analysis. Run the following code.

```
|| mean(outcome>1) + mean(outcome<0)
```

What value is given, what does it mean, and why does it imply there is something fundamentally wrong with the analysis?

CHAPTER 7:

FIXING THE VIOLATIONS

OVERVIEW:

In Chapter 5, we examined the assumptions of ordinary least squares and how to check that they are not violated by your model. The requirements (assumptions) have different importance to our estimation method. The most important requirement is that the model uniformly fits the data (constant expected value of the residuals). In this chapter, we see some ways to fix those violations.

Much of this chapter will deal with transforming the dependent variable, because mis-identified models is the greatest problem in modeling. Frequently, fixing this problem also fixes other problems with assumption violations.



Chapter Contents

7.1	The Issue of Boundedness	192
7.2	Full Example: The South Sudanese Referendum	209
7.3	Heteroskedastic Adjustments	216
7.4	Conclusion	218
7.5	End-of-Chapter Materials	219



In the previous chapters, we introduced the ordinary least squares (OLS) estimation method for the classical linear model (CLM) — and its assumptions (requirements). Last chapter, we looked at how to test that the requirements are sufficiently met in our data and model. We also looked at the importance of the assumptions. In this chapter, we determine some methods for dealing with *some* violations of those requirements. Hopefully, this extends the usefulness of this simple and straightforward estimation method.

Recall that the ordinary least squares estimation method (OLS) requires that the error terms have a constant expected value, have a constant variance, and are generated from a Normal (Gaussian) process. But, what happens when these requirements are not met?

There are essentially three ways of handling violations depending on the type and the severity: First, you can ignore it. Ignoring the violations is usually not *too* bad when you are dealing with predicting within the domain of the observed data, as the increase in bias and the loss of efficiency are usually minor. However, if it is important to estimate parameters, you definitely should not ignore this violation. Furthermore, if the assumption of a constant expected value is practically violated, you *need* to fix it.

Second, we can use other methods (and modeling paradigms) to perform regression. Two popular alternatives to the Classical Linear Model paradigm are the Generalized Linear Model (GLM) and the Generalized Additive Model (GAM). The former paradigm will be covered in Chapters 11 through 15. The latter is well examined in Wood (2006). The strength of these models (and estimation methods) is that they extend the CLM to include (for instance) discrete dependent variables and non-linear relationships (Nelder and Wedderburn 1972; Wood 2006). These unified paradigms allow the computer to estimate the effect coefficients using a very powerful method (called Maximum Likelihood Estimation). The drawback is that not all problems lend themselves to fitting using Maximum Likelihood Estimation (MLE; Chapter 10). Luckily, most do. Even more luckily, new estimation methods are developed frequently.

However, if we desire to stay within the realm of the classical linear model, estimating the parameters using ordinary least squares, we can fix many violations

simply by transforming the dependent variable — especially if the violations are minor.

These transformations are very flexible. Once you get used to working in two different systems of units, you can easily use transformation methods to ‘Normalize’ many restricted dependent variable. Unfortunately, one cannot transform an arbitrary dependent variable; there are types that cannot be fit using this technique, such as categorical. To handle these types of dependent variables, we will need to introduce a new modeling paradigm (Chapter 11).

two systems

Finally, you can make adjustments to the estimates and their standard errors to “fix” or “adjust for” the violation. This is a common practice in the presence of heteroskedasticity (Section 7.3) and multicollinearity (Section 5.4.2).

Unfortunately, these do *not* work for violations of model fit (non-constant expected residuals).

7.1: The Issue of Boundedness

We finished Chapter 6 with a model of vote proportions for ballot measures concerning keeping cows in the city (Section 6.3). We applied that model to an upcoming vote in Děčín to predict the outcome. Finally, we used Monte Carlo methods to estimate the probability that the ballot measure would pass. In the end, we predicted that the ballot measure had a 20% chance of passing, with a point-prediction of 42% of the voters in favor of the bill.

Results, however, suggest that there may be something gravely wrong with this model (Section 6.3.8). To see this more clearly, let us predict the proportion of voters in support of a hypothetical 1994 ballot measure in Venkovský (religious percent = 85) that also banned chickens (the results table from our Cow-Vote model is in Table 6.4 on page 177).

From the results summarized in the table, the point-prediction for this 1994 Venkovský ballot measure is

$$\hat{p} = 0.1512 + -0.0201(\text{yearPassed}) + -0.0373(\text{chicken}) + 0.0095(\text{religPct}) \quad (7.1)$$

$$= 0.1512 + -0.0201(-6) + -0.0373(1) + 0.0095(85) \quad (7.2)$$

$$= 1.0379 \quad (7.3)$$

Thus, this model predicts that the ballot measure will pass with over 103% of the vote — a physically impossible outcome. What went wrong? How can we fix this model so that this cannot happen?

First, nothing “went wrong,” *per se*. The model did *exactly* what it was supposed to do. The prediction, however, is based on assuming the effect (slope) is *constant*. If the slope is constant, one can find large enough (or small enough) values for the independent variables to make the prediction arbitrarily large or small. When we are predicting a bounded dependent variable, this will necessarily lead to an impossible prediction, such as a 103.79% support rate.

Thus, the issue is either with the linear (constant slope) aspect of the prediction equation *or* with the bounded nature of the dependent variable (bounded below by 0 and above by 1).

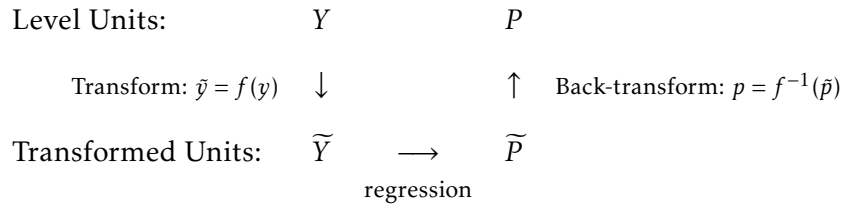


Figure 7.1: Schematic of a variable transformation procedure, such as described in the text. Here, Y is the original values of the dependent variable, \tilde{Y} is the transformed values of the dependent variable, \tilde{P} is the result from the regression in transformed units, and P is the result in the original (level) units.

So, to improve the model, we can either model using non-linear coefficient functions (Chapter 11) or eliminate the boundedness. At this point, the easier of the two is to eliminate this boundedness; that is, we need to change the dependent variable so that *all* values make physical sense. This is done through the process of **variable transformation**. There are three steps:

- First, transform the dependent variable from a restricted range to an unrestricted range.
- Second, perform the analysis on this transformed variable.
- Finally, back-transform the estimated values (not estimated effects) into the original units.

The overview of this plan is shown in Figure 7.1.

The key is the transformation. It must change the range of Y from its current limited version to an unlimited version, denoted \tilde{Y} . Luckily, there are two transformations that take care of most of our needs, in general: the logit (*LOH-jit*) and the logarithm transformations.

7.1.1 DATA BOUNDED BY 0 AND 1 One type of data you may come across in your research is proportion data, data where the values are bounded below and above (by 0 and 1, respectively); that is, if Y is the dependent variable, then $0 < Y < 1$. One appropriate function that transforms this bounded domain into an unbounded range is the logit function:

$$\tilde{y} = \text{logit}(y) := \log\left(\frac{y}{1-y}\right) \tag{7.4}$$

The logit function transforms (maps) variables bounded by 0 and 1 into unbounded variables; in symbols,

$$\text{logit} : (0, 1) \mapsto \mathbb{R} \tag{7.5}$$

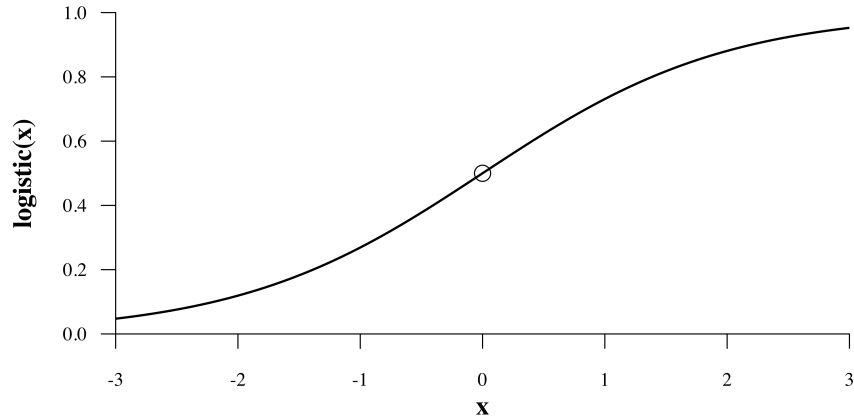


Figure 7.2: *Graphic of the logistic function. The logit function is the inverse of the logistic. Note that the graph is symmetric about the point (0, 0.5).*

The logit's inverse, which maps it from logit units back into level units is called the logistic function:

$$y = \text{logistic}(\tilde{y}) := \frac{1}{1 + \exp(-\tilde{y})} = \frac{\exp(\tilde{y})}{1 + \exp(\tilde{y})} \quad (7.6)$$

The logistic function transforms unbounded variables into variables bounded by 0 and 1:

$$\text{logistic} : \mathbb{R} \mapsto (0, 1) \quad (7.7)$$

Figure 7.2 shows a graphic of the logistic function. The logit is the inverse.

While other transforms are available, the logit is frequently used for the following three reasons:

1. The transformation *and* its inverse are both functions (the transform is a bijective function). This means that the results are always commensurate to the original problem.
2. The transformation is symmetric. This means that stretching above 0 is the same as below.
3. The function is exact, as opposed to the probit transform which requires numerical approximations. This increases the speed and accuracy of your predictions.

A careful reader will note that the domain of Y includes neither 0 nor 1. This is because there is no way of transforming a closed (or a half-closed) interval into an open interval such as \mathbb{R} while ensuring that the inverse is also a function. This is a provable fact of mathematics (Strichartz 2000).

But, what do we do if there are y -values that *are* zero or one? One solution is to add (subtract) an extremely small number, δ , to the zero (one). A second solution is to completely drop those data from the analysis.

Note: Neither of these solutions is perfect. *If* you insist on using linear regression, then you should use both methods and see how much your answer changes.

A general rule of thumb is that if your underlying research model is correct, then the results should not vary wildly based on similar models. That is, if we know Y depends on X_1 and X_2 , then all appropriate modeling techniques should give *approximately* the same results. If they do not, then there is something seriously wrong with our assumptions about the underlying relationships — the model.

rule of thumb

A third solution is to change the proportion into a bounded count and use a different paradigm (Chapter 13). While this is the best option, it requires more background before we can cover it.

Example 1

Let us return to the `cows` data file and the example of Section 6.3. The voters of Děčín are being sent to the polls to vote on a constitutional referendum that proposes to limit the number of cows kept in the city. This was not the first time that Ruritaniens were sent to the polls to vote on this or a closely related issue. Given the information from previous votes, what is the estimated probability that this ballot measure will pass in Děčín?

Solution: Let us now answer this question more correctly. Recall that without performing a transformation of the dependent variable, there existed predictions which fell outside possible reality. To fix this, let us transform the dependent variable using the logit function, repeat the analysis, back-transform these transformed results to the original units, and compare results (a la Figure 7.1).

steps

The first step is to transform the dependent variable. As the dependent variable is a proportion, let us use the `logit` transform (from the `RFS` package). If we decide to call the new variable `logitWin`, then the command will be

```
|| logitWin = logit(propWin)
```

Now, this is our new dependent variable. As such, we perform the same analysis as in Chapter 6:

	Estimate	Std. Error	t-value	p-value
Constant Term	-1.8909	0.2898	-6.53	≪ 0.0001
Year Passed (after 2000)	-0.0885	0.0157	-5.64	≪ 0.0001
Contains a Chicken Ban	-0.2318	0.0878	-2.64	0.0134
Percent Religious in Kraj	0.4750	0.0047	10.06	≪ 0.0001

Table 7.1: Results table of the results of regression on the dependent variable, using a logit transformation of the dependent variable.

```
|| modLgt = lm(logitWin ~ yearPassed + chickens + religPct)
```

The `summary(modLgt)` command provides the results summarized in Table 7.1. Note that all three independent variables are more statistically significant than in the non-transformed model, Table 6.4. Also note that the effect directions are the same as before. ♦

How shall we interpret the results? There are a few ways. The graphic is the best. However, an older manner relies on the “log odds ratio.” The odds ratio is frequently used to illustrate the strength of the association between two variables. For every increase of 1 in the percent religious in the kraj, the log of the odds of the vote passing increases by 0.4750. Said another way, the *odds* of it passing increases by approximately 60.80% for each increase of 1 percentage point in religiosity. (Note: $\exp[0.4750] = 1.6080$.)

An increase of 2pp in religiosity increases the odds by about 157%. (Note: $\exp[2 \times 0.4750] = 2.5857$.) Thus, if the original odds were 3-to-1 against, increasing the religiosity by 2pp means that the odds are now about 7.75-to-1 against.

Note: Beyond this, one *cannot* directly compare the magnitudes of these coefficients with the magnitudes of the previous coefficients; these effect estimates are in different units. The coefficients seen in Table 6.4 predict in the original units (proportions). The coefficients in Table 7.1 predict in logit (of proportions) units.

Furthermore, merely taking the logistic of the coefficients will not put them in level units; the transform is non-linear, as we designed, thus the effect of *any* depends on the values of *all*. In order to compare the two models, we need to perform predictions (remembering to back-transform them). Refer to Figure 7.3 for the steps we use in this particular example.

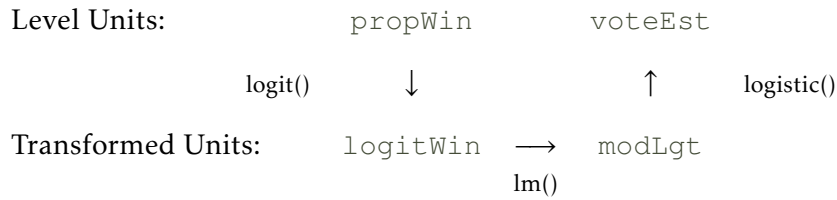


Figure 7.3: Schematic of the variable transformation procedure used in Example 7.1.1. Note that the results table, Table 7.1, displays the coefficients of `modLgt`, which is in the transformed units, not the original units. As such you cannot compare these magnitudes with the magnitudes in Table 6.4.

Predicting the proportion of the vote for the Děčín ballot measure is almost as easy as it was before. The only additional step is that we need to back-transform the prediction to get it in proportion units.

So, according to this transformed model, what is the expected vote in Děčín? To answer this, we need the Děčín information: `yearPassed = 9`, `chickens = 0`, `religPct = 48`. With this information, and under the assumption that the model is correct, we have our prediction of -0.4091 logits. Back-transforming this value gives a prediction of $\text{logistic}(-0.4091) = 40\%$ of the population will vote in favor of this ballot measure — just slightly different from our original prediction of 42%.

```

||| DECIN = data.frame(yearPassed=9, chickens=0, religPct=48)
||| voteLgt = predict(modLgt, newdata=DECIN)
||| voteEst = logistic(voteLgt)

```

However, remember that the original question was not this point estimate, it was a *probability* of the ballot measure passing. To determine this probability, we just need to repeat the same steps as we did answering this question before (Section 6.3.7), but remembering to back-transform the results.

The Monte Carlo results of the transformed model indicate that there is a 15% chance that the ballot measure will pass in Děčín. The histogram of a million predictions is presented as Figure 7.4. From this information, we can conclude that there is a definite possibility that the cow ballot measure will pass in Děčín (15%), with a predicted 40% vote in favor.

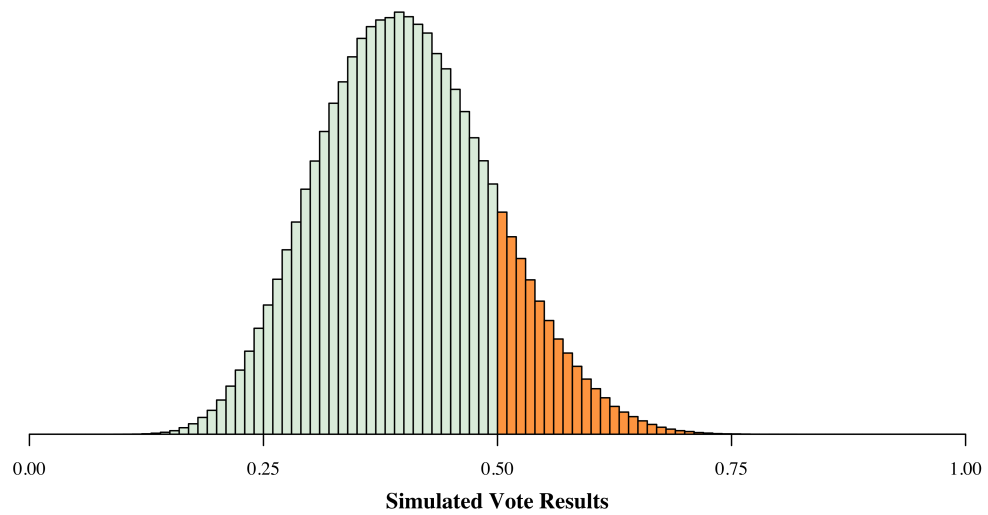


Figure 7.4: Histogram of the results of the Monte Carlo experiment described in the text. Note that the distribution has a slight right-skew as a result of the transformation process. Also note that there are no predicted vote outcomes less than 0 or greater than 1, as compared to the original untransformed model of Section 6.3.8. In fact, the lowest prediction is 9.0%, while the largest is 81.6%.

If we were into betting, we could also conclude that this model predicts that the odds of this ballot measure passing is $\frac{1-p}{p} = \frac{1-0.15}{0.15}$, 5.67-to-1 *against*. Thus, a ‘fair’ bet would pay \$5.67 for every \$1.00 bet in favor of the ballot measure and $\$1/5.67 = \0.176 for every dollar bet against the ballot measure passing.

Regardless, since the probability of the measure passing is 15%, a pass would not be wholly unexpected. Its passing is more likely than flipping a fair coin three times and having it come up heads all three times (15% vs. 12.5%) — definitely not unheard of.

The 95% prediction interval for the Děčín referendum outcome, according to our model, is from 23.5% to 59.0%. The observed value of 53% is well-within that interval.

Note: From this past discussion, we were able to estimate success probabilities and fair betting odds. This is yet another use of statistical modeling.

Note that we are estimating the probability of an event. Unless that probability is 0 or 1, there is always a chance the event will (or will not) happen. Thus, the passing of the Děčín referendum in 2009 does not directly detract from our model. There was a 15% chance it would pass, according to our model.

Warning: *Stay aware of what statistical model says and does not say — the choice is humility or humiliation.*



logarithm

7.1.2 DATA BOUNDED BELOW BY 0 When the dependent variable represents a proportion (bounded by 0 and 1), we can use the logit function to transform it into an unbounded variable, perform the usual analysis, and back-transform those results into level units (the previous section). However, not all bounded variables fit this bounding, e.g., age, height, income. These variables are bounded below by 0 and have no theoretical upper bound. For such variables, we may want to use the logarithm transform.¹

The logarithm function transforms variables bounded below by 0 into unbounded variables; in symbols,

$$\log : (0, \infty) \mapsto \mathbb{R}$$

Its inverse is the exponential function, $\exp : \mathbb{R} \mapsto (0, \infty)$. Both functions are bijections and strictly increasing and so are appropriate functions for transforming our variables.

Note that values of 0 are problematic for the logarithm in much the same way that values of 0 and 1 were problematic for the logistic function. Solutions are similar (Section 7.1.1, page 195).

Example 2

The gross domestic product (GDP) per capita is one of many measures of average wealth in countries. If extant theory is correct, then the wealth in the country is directly affected by the level of honesty in the government — countries with high levels of honesty (low levels of corruption) should be wealthier than those with low levels of honesty (high levels of corruption). Furthermore, if theory is correct, the level of democracy in a country should *also* influence the country's level of wealth — countries with higher levels of democracy should be wealthier than countries with lower levels of democracy.

Let us determine if reality (in the form of the data in the `gdp` data file) supports the current theory or if current theory needs to explain the severe discrepancies. Furthermore, let us estimate the GDP per capita for Ruritania and provide a 95% confidence interval for that estimate.

Solution: For this section, recall that the level of honesty in government for Ruritania is 5.1 and the level of democracy is -7. With that information, I leave it as an exercise for you to model the data *without* transforming the dependent variable and discovering the predicted GDP per capita for Ruritania is \$26,795.64. This

¹By “theoretical upper bound,” we mean there exists a limit (a single value) such that the variable can get sufficiently close to *that* limit, but no greater.

seems awesome for Ruritania. The 95% prediction interval is from \$5232 to \$48,360. That's rather wide. It is a function of the high level of variation in the data.

However, to see a problem with the model, let us estimate the GDP per capita for Papua New Guinea (democracy=10, hig=2.1). According to the model, the predicted GDP per capita is -\$2337, which is not physically possible. If nothing else, this prediction should suggest to you that the data needs transformation before being modeled.

The process to estimate the GDP per capita in Ruritania using a transformed model is formulaic for us by now: transform the dependent variable by applying the logarithm function, model the transformed variable, estimate in the transformed units, back-transformed into level units — here, dollars.

One feature of R that is shared by few other statistical packages is that you do not have to actually create a new variable; you can perform the transformation within the modeling command; *e.g.*,

```
|| modLog = lm(log(gdpcap) ~ democracy + hig)
```

The results table for this model is provided in Table 7.2. Again, as we have transformed the dependent variable, the coefficients are not in units of dollars. As such, their magnitudes cannot be directly compared to those in the untransformed model. Their *directions*, however, can be compared because the transformation we used was strictly increasing. Thus, this model tells us that higher levels of honesty in government correspond to countries with higher GDPs per capita (in this sample). Additionally, countries with higher democracy scores correspond to countries with *lower* GDPs per capita (in this *sample*).

The first finding is so strong in this sample that we can conclude that there is evidence of this relationship *in the population*. This second finding, which conflicts with current theory, is not statistically significant at the usual $\alpha = 0.05$ level. Thus, we cannot conclude that the effect in the population is negative, positive, or null (zero). All we can conclude is that we did not detect an effect *with this data*. Whether this is due to a lack of effect in the population, the sample selected, the sample size, no one can tell.

With this model, we can estimate the GDP per capita in Ruritania using the standard method, but remembering that we must back-transform the final estimate. That is, if we used the commands

	Estimate	Std. Error	t-value	p-value
Constant term	6.9333	0.1479	46.89	$\ll 0.0001$
Level of Democracy	-0.0028	0.0113	-0.25	0.8055
Honesty in Government	0.4702	0.0359	13.11	$\ll 0.0001$

Table 7.2: Results table for the GDP per capita modeling exercise. As the model is a transformed model, these effects estimates are not in units of dollars.

```
|| RUR = data.frame(hig=5.1, democracy=-7)
|| estLog = predict(modLog, newdata=RUR)
```

then we would report Ruritania's GDP per capita as an estimated value of \$11,508 (using `exp(estLog)`). ♦

Note: From your mathematics course, you may recall that $\log(1 + x) \approx x$ for small values of x . This means we can interpret the coefficients in the log-model as percent increases/decreases. For instance, the coefficient for the level of democracy in the country is -0.0028. We can interpret this as “one increase in the level of democracy decreases the GDP per capita by 0.28%, on average.” The coefficient of the level of honesty in government is 0.4702. We could interpret this as “one increase in the level of honesty in the government increases the GDP per capita by approximately 47%, on average.”

However, what do we mean by “small values of x ”? Anything less than 0.2 is usually fine. Our interpretation of the honesty-in-government coefficient probably should not have been done. A log-coefficient value of 0.4702 really corresponds to a percent increase of only 38.5%. It is more accurate, but less spiffy.

Here is my code to explore the relationship $\log(1+x) \approx x$:

```
|| x = seq(0,1, length=1e4)
|| y = log(1+x)
|| plot(x,y, col="blue1")
|| abline(0,1, col="orange")
```



The question asked us to calculate the estimate, but to also provide a 95% confidence interval. One way of doing this is to use Monte Carlo methods. The steps are all the same, with the additional step of back-transforming the estimates (last line).

Here is the code for parametric bootstrapping:

```
|| b.int = 6.933298
|| b.dem = -0.002776
|| b.hig = 0.470225
||
|| s.int = 0.147873
|| s.dem = 0.011253
|| s.hig = 0.035855
||
|| e.int = rnorm(trials, m=b.int, s=s.int)
|| e.dem = rnorm(trials, m=b.dem, s=s.dem)
|| e.hig = rnorm(trials, m=b.hig, s=s.hig)
||
|| outcome = e.int + e.dem*-7 + e.hig*5.1
|| est = exp(outcome)
```

The assignments in the second and third group are the coefficient estimates and standard errors from the model (Table 7.2). The histogram of these results are provided in Figure 7.5. To calculate a 95% confidence interval, we merely find the values of `est` for which 2.5% and 97.5% of the data are less.

```
|| quantile(est, c(0.025,0.975))
```

From this, we can conclude that our model estimates the GDP per capita for Ruritania is \$11,508, with a 95% confidence interval being from \$7075 to \$18,733. It is interesting to note that the actual GDP per capita in Ruritania is \$55,000, which is well above our confidence interval. Thus, our question is this: Is our model that weak, or is Ruritania doing that well?

Note: Here, I use the original estimate as the point estimate for the GDP per capita of Ruritania (\$11,508). It would have also been appropriate to use the mean of the Monte Carlo trials (\$11,870) *or* the median of the Monte Carlo

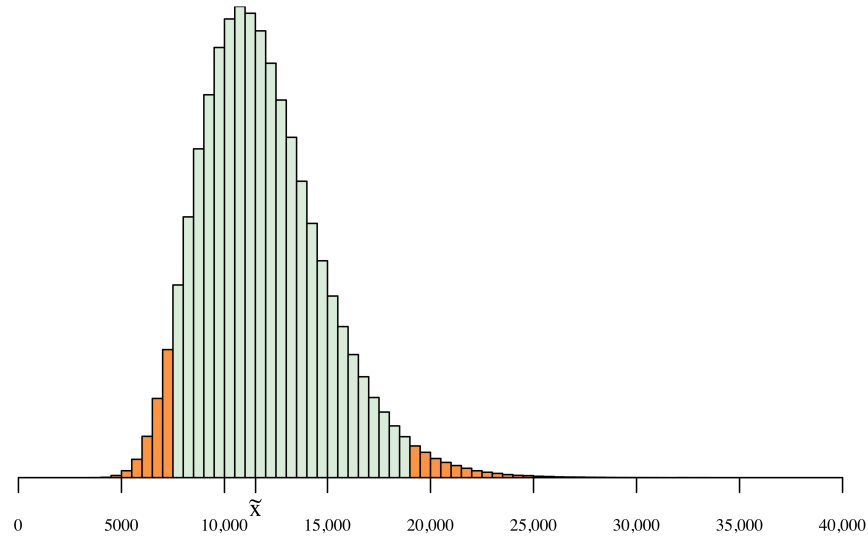


Figure 7.5: Results of the Monte Carlo experiment estimating the GDP per capita for Ruritania and its 95% confidence interval. Note that 5% of the estimates fall in the rejection (*tan*) region, 2.5% above and 2.5% below. The median of this distribution is designated by \bar{x} .

trials (\$11,510). All three are acceptable measures of the center. It is usual, however, to use the original prediction.

Here is an interesting question. In the previous example, we estimated a confidence interval. How could we estimate a prediction interval?

To answer this, we need to remember the only difference between confidence and prediction intervals. In a confidence interval, we are estimating an expected value. In a prediction interval, we are predicting a new outcome. That new outcome is a combination of the expected value *and* the σ^2 from the ε term.

And so, to get a prediction interval, we use the following. Check to see the difference between this and the previous script.

```

|| b.int   = 6.933298
|| b.dem   = -0.002776
|| b.hig   = 0.470225
|| b.err   = 0
||
|| s.int   = 0.147873
|| s.dem   = 0.011253

```

```
s.hig = 0.035855
s.err = 0.8841

e.int = rnorm(trials, m=b.int, s=s.int)
e.dem = rnorm(trials, m=b.dem, s=s.dem)
e.hig = rnorm(trials, m=b.hig, s=s.hig)
e.err = rnorm(trials, m=b.err, s=s.err)

outcome = e.int + e.dem*-7 + e.hig*5.1 + e.err
est      = exp(outcome)
```

From this, the 95% prediction interval is from \$1907 to \$69,345. Note that it is *much* wider than the confidence interval. Also note that this should not surprise us at all. Prediction intervals are always wider than the corresponding confidence interval.

7.1.3 ADDITIONAL BOUNDS Thus far, we have looked at transformation of a dependent variable when it is bounded above and below by 0 and 1 (two bounds), and when it is only bounded below by 0 (one bound). Other bounds are possible.² In this section, we figure out how to handle all types of bounds. The basic steps are to determine if the variable is bounded on one side or two. If one, then perform an *algebraic* transformation so that the new variable is bounded below by 0, then use the log transform. If two, then perform an algebraic transformation so that the new variable is bounded by 0 and 1, then use the logit transform. In either case, you will need to remember to back-transform the predictions with this algebraic transformation.

Note: The only bounds I frequently come across in my own research are those bounded by 0 and 1, bounded by 0 and 100 (percentages), bounded by 0 and 4 (GPAs), and bounded below by 0. The quick solution for percentages is to divide them by 100 to make them proportions, then multiply the predictions by 100 to turn the predictions back into percentages.

BOUNDED BY L AND U : What if our data has a theoretic lower bound L and a theoretic upper bound U ? As it is bounded above *and* below, we will change it into a proportion and using the logit transform as in Section 7.1.1, remembering to back-transform with the additional transformation. The algebraic transformation is

$$a(y) = p = \frac{y - L}{U - L} \quad (7.8)$$

The back-transform is

$$a^{-1}(p) = y = p(U - L) + L \quad (7.9)$$

Example 3

The scores on the quantitative portion of the Graduate Record Examination (GRE) range from $L = 200$ to $U = 800$. If we wished to properly model a person's GRE quantitative score, we would first subtract 200 from each score, then divide by $800 - 200 = 600$. The new variable would range from 0 to 1, a proportion.

²While other bounds are possible, the number of bounds can only be 0, 1, or 2. This makes this section so important.

Example 4

The grade point averages (GPAs) are bounded below by $L = 0$ and above by $U = 4$. To appropriately model GPAs, we would have to subtract 0, then divide by 4. This new variable would now be a proportion.

BOUNDED BELOW BY L : It may be that your dependent variables is bounded below by a specific value, L , but not bounded above. As it is bounded on only *one* side, we will transform it into a variable bounded below by 0 and then apply the logarithm transform as in Section 7.1.2, remembering to back-transform with the additional transformation. The algebraic transformation is

$$a(y) = p = y - L \quad (7.10)$$

The back-transform is

$$a^{-1}(p) = y = p + L \quad (7.11)$$

Example 5

Hourly workers make at least \$7.25 per hour. To model excess hourly wage, we would subtract off $L = 7.25$ from each hourly wage. This new variable is bounded below by 0, so we can apply the log transformation to it.

BOUNDED ABOVE BY U : It may be that your dependent variable is theoretically bounded *above* by U . As there is only *one* bound, we will perform an algebraic transformation so that it is bounded below by 0 and then apply the log transform as in Section 7.1.2, remembering to back-transform with the additional transformation. The algebraic transformation is

$$a(y) = p = U - y \quad (7.12)$$

The back-transform is

$$a^{-1}(p) = y = U - p \quad (7.13)$$

Example 6

In the ocean, different species live at different depths. In fact, we can predict the depth based solely on the species observed. Ocean depth is bounded above by 0 and has no theoretic lower bound (although it certainly has a genuine lower bound at the Challenger Deep in the Mariana Trench, which

has a depth of -35,994 ft). To transform the depths into a variable upon which we can perform a log transform, we subtract each value from $U = 0$. After we predict, we will have to back-transform by again subtracting each prediction from $U = 0$.

Of course, the transformation in this last example is equivalent to measuring depth in terms of 'distance below the surface', which is a positive number requiring no additional transformation.

7.2: Full Example: The South Sudanese Referendum

Free and fair elections are one of the requirements for a legitimate democratic system; furthermore, being a legitimate democratic State is necessary for some forms of external assistance. As such, many not-so-democratic States wish to appear democratic. They hold elections, but the elections are either fraudulent or the electoral system (rules governing the elections) is unfair.

There are many definitions for fairness in an election, but they all contain the same requirement that a person's vote has the same probability of being counted as anyone else's. In other words, the probability of a vote being invalidated is independent of the characteristics of the person casting the vote — including who the vote was for. This aspect of fairness can actually be tested in elections where the number of invalidated votes is counted: If the proportion of the vote for a specific candidate or position is not independent of the proportion of the vote invalidated in the electoral division, then there is evidence against this assumption of fairness.



And so, with this background, ...

Question

Does the 2011 independence referendum in southern Sudan indicate an issue with fairness?

Narrative Solution: As one of the conditions to the 2005 Naivasha Agreement, which ended the civil war in Sudan, the South was allowed to vote on independence from the North. That referendum was held January 9–15, 2011. Official results stated that 98.83% of the South Sudanese voted against unity and in favor of independence.

The `xsd2011referendum` data contains the number of votes in favor of independence (`Secession`), the number of votes declared invalid (`Invalid`), and the total number of votes cast (`Votes`). Load it and save it into the `xsd` variable without attaching the data. Because we need to determine if there is a relationship between the proportion of the vote for a specific side and the proportion of the vote invalidated in the electoral division, and because we just have vote counts, we need to create those proportions. The proportion of the vote for the candidate is the number of votes for the candidate divided by the number of valid votes. The

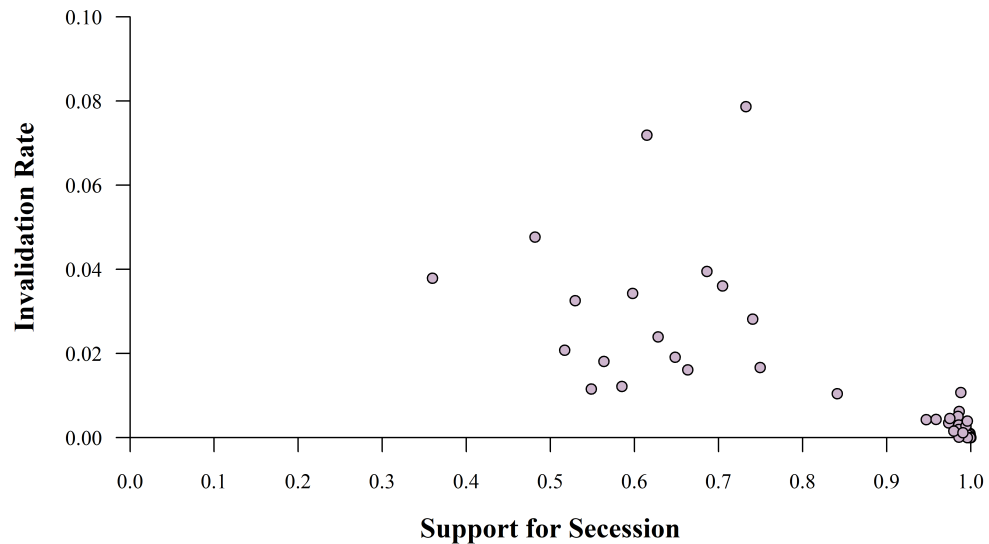


Figure 7.6: A scatterplot of the results of the 2011 referendum on independence for South Sudan. Note the apparent presence of a relationship between these two variables. As such, there appears to be evidence that the election was not fair for those voting against independence.

invalidation rate is the number of invalid ballots divided by the number of cast ballots (recall Section 7.1.3).

Once that is done, we need to transform these proportions using the logit transformation (why?), perform linear regression, and check for a (linear) relationship. If a relationship exists in the transformed variables, then a relationship exists in the untransformed variables. First, however, it is always a good idea to plot the variables to see if there is an obvious answer to the question. Figure 7.6 is the plot of proportion of the vote invalidated against the proportion of the vote in favor of independence.

Suggested by the plot, there appears to be a strong relationship between the two variables, evidence of an election that is not fair. Because of the direction of the slope, it appears as though those areas voting most strongly in favor of independence had a much lower probability of having their votes rejected.

Note: As we are using the logit transform, we must drop any electoral division (here, county) which has zero invalid votes or zero votes in favor of secession. We need to do this because the domain of the logit function is $p \in (0, 1)$.

	Estimate	Std. Error	t-value	p-value
Constant term	1.8978	0.7690	2.468	0.0155
Proportion of Vote for Independence	-9.3991	0.8287	-11.342	<< 0.0001

Table 7.3: Results table for the South Sudan referendum. The results are in logit units. Note the high level of statistical significance in the effect of the proportion of the vote in favor of independence. This is very indicative of a lack of fairness in the election.

Question

How will removing these counties affect the conclusions drawn?

To easily do this in R, we can use the `which` function, which returns which entries have the provided condition. Thus,

```
|| dr = which(xsd$Invalid==0)
```

returns a vector of values $\{15, 19, 23, 24, 28, 46, 47, 49, 50, 57, 72, 73\}$. These numbers correspond to the counties that had zero invalid votes cast. Storing this vector in the variable `dr` allows us to remove those counties from any subsequent calculations. As such, our proportion calculations are:

```
|| p.ind = xsd$Secession[-dr]/xsd$Votes[-dr]
|| p.inv = xsd$Invalid[-dr]/xsd$Votes[-dr]
```

The negative signs tells R to return values in the vector *other than* these entries.

And so, the two lines to transform the dependent variable and fit the OLS model are

```
|| l.inv = logit(p.inv)
|| model.xsd = lm(l.inv ~ p.ind)
```

The results of the linear regression on the transformed dependent variable are given in Table 7.3. There is a very strong relationship between the proportion of the vote invalidated in the county and the proportion of the vote in favor of secession: Those counties with a greater proportion of people voting for independence also had a lower proportion of the vote invalidated. That there is a strong relationship between these two variables is troubling.

To make this relationship more obvious, and to make our point stronger, we can plot the data, the prediction curve, and the 95% Working-Hotelling confidence bands on the same plot.

Note: What confidence intervals are to univariate data, confidence bands are to bivariate data. We briefly saw the Working-Hotelling confidence bands in Section 4.4.

7.2.1 THE GRAPHING PHILOSOPHY OF R In R, the philosophy behind graphing using base graphics is to start with a fresh plot and paint successive layers on top of it. This allows us to create graphs that tell the story and to do so easily. To make the graph described above, we need to

1. Plot the points (displayed in proportion units),
2. Plot the prediction curve (displayed in proportion units, but calculated in logit units),
3. Plot the 95% confidence bands (displayed in proportion units, but calculated in logit units).

The first step has been done already (Figure 7.6).

The second step requires the repeated use of the `predict` function. First, to make things easier, let us define `newX` as a series of “proportion of vote in favor of independence” values for which we want to make predictions: `newX = seq(0, 1, length=1e4)`. This creates a vector containing 10,000 values equally spaced between 0 and 1.

With this, our predict statement will be

```
l.pred = predict(model.xsd,
  newdata=data.frame(p.ind=newX),
  se.fit=TRUE)
```

Note: The `se.fit=TRUE` parameter, which calculates the standard error of the fit at that x-value, will be important for calculating the confidence bands. This is just a courtesy from R, as we know how to calculate this value from Theorem 4.3.

Remember that these predictions are in logit units. To get them into level units, we just apply the logistic function to these point predictions:

```
|| p.pred = logistic(l.pred$fit)
```

Note: The `$fit` selects only the fitted predictions from the `l.pred` variable. This is necessary as we are also using the `se.fit=TRUE` parameter.

Now that we have the predictions in the original units, we merely paint it on the current plot (from Step 1):

```
|| lines(newX, p.pred)
```

The third step requires us to calculate the 95% confidence bands and paint them on the plot as well. For want of better estimates, let us use the Working-Hotelling bands (Section 4.4). The formula to calculate the upper 95% confidence bands is

```
|| ucb.l = l.pred$fit+W*l.pred$se.fit
```

the lower,

```
|| lcb.l = l.pred$fit-W*l.pred$se.fit
```

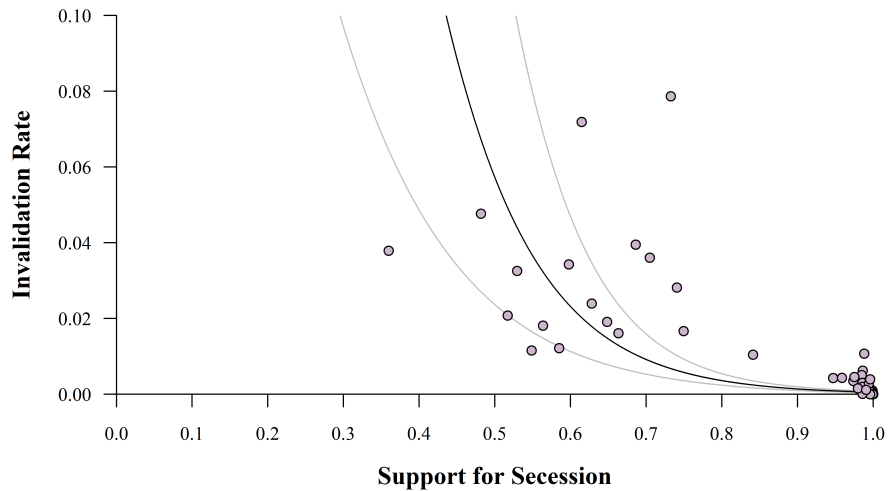


Figure 7.7: A plot of the results of the South Sudan referendum. Included are the prediction line (in black) and the 95% confidence bands (in grey). Note that a horizontal line cannot fit between the confidence bands. This indicates a statistically significant relationship between the proportion of the votes invalidated and the proportion of the votes in favor of independence. This, in turn, supports the conclusion of an unfair election.

Here, $W = \sqrt{2 F(1 - \alpha, 2, n - 2)}$, which translates to

```
|| W = sqrt( 2 * qf(1-0.05, 2, n-2) )
```

Note: The form of these formulas should look vaguely familiar. They are of the same form as when we calculated the upper and lower limits for Normal confidence intervals,

$$u = \bar{x} + 1.96s_x \quad \text{and} \quad l = \bar{x} - 1.96s_x$$

The W distributional multiplier comes from Working-Hotelling (1929) and its Scheffé extension (1959).

Once again, we must back-transform these two variables using the logistic function. So, our final confidence bands are

```
|| ucb = logistic(ucb.l)
|| lcb = logistic(lcb.l)
```

Finally, we paint this on the current plot with


```
|| lines(newX, ucb, col="grey")  
|| lines(newX, lcb, col="grey")
```

Putting all this together gives us Figure 7.7. Note that the predictions are curved in these units; they are straight in logit units. Also note the confidence bands are wider where the value of x is farther from \bar{x} (Theorem 4.3). Lastly, note that no horizontal line can fit between the two confidence bands. This illustrates that there is a statistically significant relationship between the two variables at the $\alpha = 0.05$ level.

Question

This illustrates that there is a statistically significant relationship between the two variables at the $\alpha = 0.05$ level. (Why?)

Note that the figure gives the same information as Table 7.3. The difference is that the graphic tells a clear story. Graphs usually make the points more manifest.

7.3: Heteroskedastic Adjustments

The above transformations also work well on fixing problems with heteroskedasticity and non-Normality. Unfortunately, if you perform an appropriate transformation to fix the problem with model fit, further transformations to fix heteroskedasticity may end up creating a new problem with model fit.

Thus, it may happen that you cannot find a way of fixing the heteroskedasticity without breaking something else. In such cases, we can adjust the standard errors using a technique introduced by White in 1980.

Recall from ordinary least squares estimation (page 48) that our estimator for \mathbf{B} is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (7.14)$$

From this we showed in Theorem 3.1.3 that this estimator was unbiased; that is, $\mathbb{E}[\mathbf{b}] = \mathbf{B}$. In Theorem 3.1.3, we also calculated the variance of the estimator as

$$\mathbb{V}[\mathbf{b}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \quad (7.15)$$

That result, however, required that $\mathbb{V}[\mathbf{Y}] = \sigma^2 \mathbf{I}$. This is the assumption of homoskedasticity (Section 3.1.2). Under heteroskedasticity, $\mathbb{V}[\mathbf{Y}]$ cannot be reduced. This leaves the variance of our OLS estimators as

$$\mathbb{V}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbb{V}[\mathbf{Y} | \mathbf{X}] \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \quad (7.16)$$

So, to better estimate $\mathbb{V}[\mathbf{b}]$, we need to estimate $\mathbb{V}[\mathbf{Y} | \mathbf{X}]$ from the data. (Everything else in Equation 7.16 is known.) How do we estimate $\mathbb{V}[\mathbf{Y} | \mathbf{X}]$ from the data? We recall that $\mathbb{V}[\mathbf{Y} | \mathbf{X}] = \mathbb{V}[\mathbf{E}]$. Thus, we estimate $\mathbb{V}[\mathbf{Y}]$ from the residuals, specifically from how large each residual is. If the i^{th} residual is large, then the value of $\mathbb{V}[Y_i]$ will be large; if e_i is small, then $\mathbb{V}[Y_i]$ will be small.

Note: The Huber-White Sandwich Estimator makes adjustments for heteroskedasticity. These adjustments, however, are based on the data, which are random. Including this new source of randomness affects all confidence intervals and needs to be acknowledged.

It seems to me that it would be easier to create an entirely new estimation method than to just pile on adjustment after adjustment on OLS. This is, in fact, where we are headed in this book.

	Estimate	Std. Error	t-value	p-value
Constant term	1.8978	0.6704	2.831	0.0045
Proportion of Vote for Independence	-9.3991	0.7420	-12.667	$\ll 0.0001$

Table 7.4: Results table for the South Sudan referendum using White (heteroskedastic-consistent) standard errors. Compare this to Table 7.3. The results are in logit units. Note the high level of statistical significance in the effect of the proportion of the vote in favor of independence. This is very indicative of a lack of fairness in the election.

7.3.1 HAVING R DO THIS FOR US Instead of performing the above calculations by hand, we can have R do the adjustments for us. That helps with the accuracy and precision. The `summaryHCE` function in the `RFS` package provides the adjustment and presents it in the form of our usual regression table.

Example 7

To illustrate the operation of the `summaryHCE` function, let us calculate the White-adjusted standard errors for the South Sudanese model above.

Solution: We have already calculated the regression table for the logit model (Table 7.3). From looking at the graphic, it seems as though there may be heteroskedasticity. There appears to be a lot more variation in the invalidation rate for smaller values of secession support than for larger values of secession support.

Running the following adjusts the standard errors to reflect the observed heteroskedasticity.

```
|| summaryHCE(model.xsd)
```

The heteroskedasticity-adjusted regression table is given in Table 7.4. Note that the estimates remain the same. That is because heteroskedasticity does not affect the estimates. The only changes are in the standard errors (and the test statistics and the p-values). ♦

Notice that adjusting the standard errors is rather easy using R. It is just a single line. Also notice that we did not model the heteroskedasticity, we merely adjusted for it.

At some level, it is unsettling to adjust for model weaknesses. It is a strong model that does not need fixes. Thus, if you can avoid using these Huber-White

standard errors, I recommend it strongly. Heteroskedasticity is an important part of the data/model. It seems sinful to ignore it.

7.4: Conclusion

In this chapter, we focused on transforming bounded variables so that they did not violate the Normality assumptions as strongly as they did without the transformation. To accomplish this, we noted that there are three basic types of continuous variables: unbounded, bounded on one side, and bounded on two sides. If the dependent variable is unbounded, we do not necessarily need to transform it (although some transforms may reduce the non-Normality of the residuals). If the variable is bounded on one side, we performed an algebraic transformation so that it is bounded below by zero, then applied a log transformation. If the variable is bounded on two sides, we performed an algebraic transformation so that it was bounded by 0 and 1, then applied a logit transformation.

In either case, we needed to ensure that we back-transformed to the original units, first using an exponential or a logistic back-transform, then the inverse of our algebraic transform — order matters.

While this chapter does not exactly mark the end of continuous dependent variables, it does end our view of them in terms of the Classical Linear Model (CLM). This chapter already shows why the CLM needs to be replaced. Here, we were able to stay within the framework, but we had to perform variable transformations to make it work. Once we stray from continuous data, the CLM cannot work; there is no way of transforming a discrete dependent variable into a Normally distributed random variable. As such, we need a new paradigm — Generalized Linear Models (GLMs). The next chapter introduces GLMs, while still using a continuous dependent variable. This is done to show that GLMs can do anything CLMs can do. In fact, if you had used the `glm` function in this and the previous chapter, in lieu of the `lm` function, the results would be *exactly* the same, only the table layout would be different.

7.5: End-of-Chapter Materials

7.5.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

STATISTICS:

lm(formula) This function performs linear modeling on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable, to retrieve the information through the `summary` and `names` functions.

predict(model) The `predict` function calculates the value of the dependent variable in the `model` given the independent variables used to create the model. If new predictions are required, the `newdata=` parameter must be used. This parameter takes a new set of data as its argument. Make sure that all independent variables used in the model are defined in the `newdata=` parameter. If not, an error message will result. Finally, the `se.fit=TRUE` parameter calculates the standard error at each prediction point.

summaryHCE(model) This function, a part of the `RFS` package, allows us to easily calculate the heteroskedastic-consistent standard errors (White 1980).

PROBABILITY:

pnorm(x) This function is the cumulative distribution function (CDF) for the Normal distribution. It returns a *probability* that a Normally-distributed variable will be less than or equal to `x`. This function has two additional parameters that remove the requirement that `x` has undergone the *z*-transformation, `m` and `s`.

rnorm(n, m, s) This function returns `n` draws from a Normal distribution centered at `m` and with a standard deviation `s`. This function is the cornerstone of much Monte Carlo analysis.

GRAPHICS:

lines(x,y) This is an extremely handy line-generating function, painting a line on the current plot (or returns an error if no plot exists). It first invisibly plots the pairs of points (x,y) then connects the points with drawn line segments.

If the `col` parameter is not set, then the line will be black. Otherwise, the line will be the color specified. There are three ways of stating the color: using the Windows 1-16 values, using names, and using the rgb values. The following all refer to 'red': `col=2`, `col="red"`, and `col="#ff0000"`.

plot() This function produces a scatterplot of the two-dimensional data. The call can be either `plot(x,y)` or `plot(y~x)`; both give identical results. This function can produce graphs that are very customized. The R help file for `par` is invaluable. Some important parameters include `xlab=""` (label for the x-axis), `ylab=""` (label for the y-axis), `xlim=c(min,max)` and `ylim=c(min,max)` (axis limits, min and max, for the x- and y-axis), and `las=1` (makes axis values painted horizontal).

MATHEMATICS:

log(x, b) This returns the logarithm of x , with a base of b . If you omit the b , this function returns the natural logarithm of x . To calculate the common logarithm, set $b=10$. The logarithm function is used to transform variables bounded on one side into variables bounded on neither side.

exp(x) This function returns the exponential of the argument, x ; that is, it returns e^x . The exponential function is the inverse of the logarithm function.

logit(x) This function returns the logit of the provided number. This number must be between 0 and 1, not including either 0 or 1. The logit function is frequently used to transform proportions into unbounded data. It is available through the [RFS](#) package.

logistic(x) This function returns the logistic of a given number. The range of the logistic function is 0 to 1, exclusive. it is the inverse of the logit function. As such, it is often used to transform predictions from logit units to proportion units. It is available through the [RFS](#) package.

cloglog(x) The complementary log-log function is a second appropriate transformation for proportion data. It is, however, not a symmetric function. It is available through the [RFS](#) package.

cloglog.inv(x) This function is the inverse of the complementary log-log function. It is available through the [RFS](#) package.

PROGRAMMING:

which(condition) This function returns a vector of indices corresponding to the original vector's values meeting the criteria. Thus, `which(x==4)` returns the indices of all elements in vector `x` that equal 4. Note that equality is checked with a *double* equals, `==`. Other comparisons include: `>`, `<`, `>=`, `<=`, `!=`, `&`, `|`, and `!`. The last four are 'not equal to', 'and', 'or', and 'not'.

7.5.2 EXERCISES This section offers suggestions on things you can practice from this chapter.

1. Predict the Venkovský 1994 cow ballot measure vote using the transformed vote model. Is *this* prediction physically possible?
2. Determine a 95% confidence interval, with the *untransformed* cow vote model, for predicting Děčín's vote. Is the actual outcome within the 95% confidence interval?
3. Determine a 95% confidence interval, with the *transformed* cow vote model, for predicting Děčín's vote. Is the actual outcome within the 95% confidence interval?
4. Determine if the assumptions of OLS are violated in the transformed cow vote model.
5. The actual vote share for Děčín was 52.8%. Explain why both models failed in predicting the actual vote outcome. How bad was the error? What can be done to improve the predictions?
6. The logit transformation is not the only possible choice. There is also the asymmetric complementary log-log transformation (`cloglog` in the `RFS` package). Use this function as the transformation to predict Děčín's vote, its 95% confidence interval, and the probability of the cow ballot measure passing. The inverse of the complementary log-log transform has no name, but the R function is `cloglog.inv`, also in the `RFS` package.
7. Estimate the GDP per capita for Papua New Guinea using the *untransformed* model, as well as the 95% confidence interval. How close is this estimate to the real answer, and is the real answer within the predicted confidence interval?

7.5.3 APPLIED READINGS

- James M. Avery. (2009) “Political Mistrust among African Americans and Support for the Political System.” *Political Research Quarterly* 62(1): 132–45.
- Mark Andreas Kayser. (2009) “Partisan Waves: International Business Cycles and Electoral Choice.” *American Journal of Political Science* 53(4): 950–70.
- Pamela A. Morris. (2008) “Welfare Program Implementation and Parents’ Depression.” *The Social Service Review* 82(4): 579–614.
- Kar Tean Tan, Christopher C. White, and Donald L. Hunston. (2011) “An adhesion test method for spray-applied fire-resistive materials.” *Fire and Materials* 35(4): 245–59.

7.5.4 THEORY READINGS

- George Casella and Roger L. Berger. (2001) *Statistical Inference*. New York: Duxbury Press.
- Annette J. Dobson and Adrian Barnett. (2008) *An Introduction to Generalized Linear Models*, Third Edition. New York: Chapman & Hall.
- Friedhelm Eicker. (1967) “Limit Theorems for Regression with Unequal and Dependent Errors.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*: 59–82.
- Julian J. Faraway. (2004) *Linear Models with R*. New York: Chapman & Hall.
- Julian J. Faraway. (2005) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. New York: Chapman & Hall.
- Peter J. Huber. (1967) “The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*: 221–233.
- John A. Nelder and Robert W. M. Wedderburn. (1972) “Generalized Linear Models.” *Journal of the Royal Statistical Society. Series A (General)* 135(3): 370–84.
- Henry Scheffé. (1959) *The Analysis of Variance*. New York: Wiley.
- Shayle R. Searle. (1997) *Linear Models*. New York: Wiley-Interscience.
- James H. Stapleton. (2009) *Linear Statistical Models*. New York: John Wiley and Sons.
- Robert S. Stritchartz. (2000) *The Way of Analysis*, Revised Edition. Boston: Jones and Bartlett Mathematics.
- Halbert White. (1980) “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*. 48(4): 817–838.
- Simon N. Wood. (2006) *Generalized Additive Models: An Introduction with R*. New York: Chapman & Hall.
- Holbrook Working and Harold Hotelling. (1929) “Applications of the Theory of Error to the Interpretation of Trends.” *Journal of the American Statistical Association* 24(1): 73–85.

Part II

Beyond the Ordinary

8	Other Least Squares	227
9	Quantile Regression	261
10	Maximizing the Likelihood	281

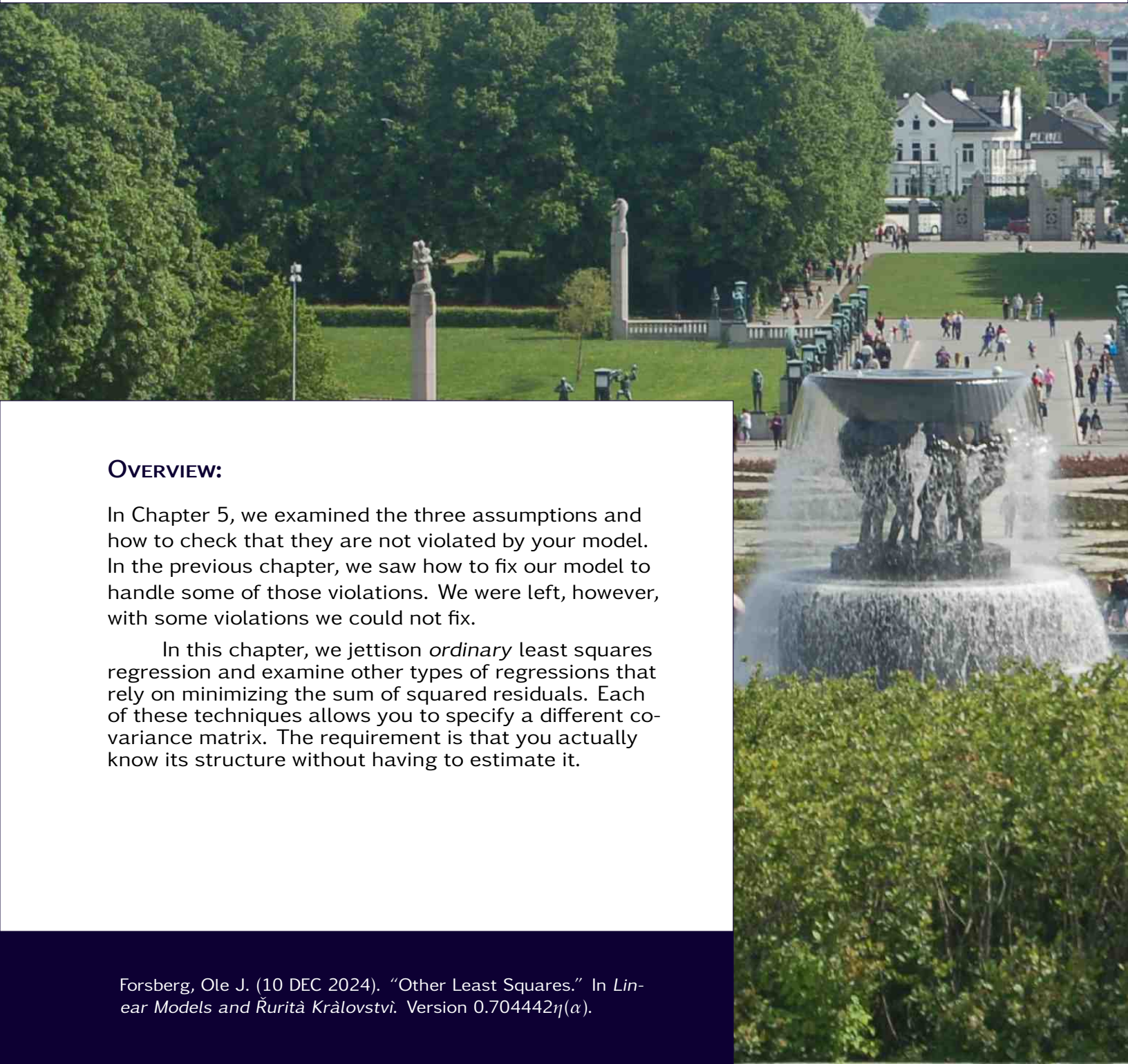
CHAPTER 8:

OTHER LEAST SQUARES

OVERVIEW:

In Chapter 5, we examined the three assumptions and how to check that they are not violated by your model. In the previous chapter, we saw how to fix our model to handle some of those violations. We were left, however, with some violations we could not fix.

In this chapter, we jettison *ordinary* least squares regression and examine other types of regressions that rely on minimizing the sum of squared residuals. Each of these techniques allows you to specify a different covariance matrix. The requirement is that you actually know its structure without having to estimate it.



Chapter Contents

8.1	Ordinary Least Squares	229
8.2	Weighted Least Squares	230
8.3	Generalized Least Squares	240
8.4	Full Example: May the (Strong) Force be with You.	246
8.5	Full Example: Elections in Ruritania.	250
8.6	Conclusion	255
8.7	End-of-Chapter Materials	256



In the past several chapters, we have examined the classical linear model (CLM) and how to estimate the parameters using ordinary least squares (OLS). That introduction came in Chapters 2, 3 and 4. In Chapter 5, we discovered how to check the requirements (assumptions) of the ordinary least squares method. Chapter 7 gave us some options for dealing with violations of the requirements.

However, it may be that those fixes do not fully succeed — or *cannot* fully succeed. This chapter provides two estimation methods that offer advantages over ordinary least squares, as long as you have sufficient knowledge (science) of the structure of the problem — also known as the data-generation process.

This chapter reintroduces ordinary least squares. It then focuses on the covariance matrix of the residuals. As we reduce requirements on that matrix, we move from ordinary least squares to weighted least squares to generalized least squares.

8.1: Ordinary Least Squares

First, let us review ordinary least squares (OLS). When formulating OLS estimation of the classical linear model (CLM), we made the assumption that the residuals are independent and identically distributed Normal with constant zero expected value and variance.

In symbols, this is written as either

$$\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0; \sigma^2) \quad (8.1)$$

or as

$$\mathbf{E} \sim \mathcal{N}_n(\mathbf{0}; \sigma^2 \mathbf{I}) \quad (8.2)$$

The two statements are different ways of saying the exact same thing.

Note that the covariance matrix of \mathbf{E} is $\sigma^2 \mathbf{I}$:

$$\mathbf{V}[\mathbf{E}] = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} \quad (8.3)$$

The values along the diagonal represent the variances of each residual *in the population*. That they are the same value, σ^2 , indicates that the variance of the residuals is constant.

The values off the diagonal represent the covariance between the residuals. For instance, the value at position 1,2 is the covariance between ε_1 and ε_2 , which we symbolized as $\sigma_{1,2}$ in Appendix S. Since that value is 0, we are specifying that the two are linearly uncorrelated (a.k.a. independent).

Thus, the covariance matrix above specifies that the variances of the residuals are constant and that the residuals are independent of each other. If this requirement is met, then we should use ordinary least squares regression. However, not always is this requirement met.

homoskedastic

independent

foreshadowing

8.2: Weighted Least Squares

It may be that the residuals are independent, but that their variance is known to not be constant. That is, we may have a model that leads to this assumption:

$$\varepsilon_i \stackrel{\text{ind}}{\sim} \mathcal{N}(0; \sigma_i^2) \quad (8.4)$$

or as

$$\mathbf{E} \sim \mathcal{N}_n(\mathbf{0}; \sigma^2 \mathbf{D}) \quad (8.5)$$

Here, \mathbf{D} is a diagonal matrix. Again, the two statements are different ways of saying the same thing.

Note that the covariance matrix of *this* \mathbf{E} is

$$\mathbb{V}[\mathbf{E}] = \sigma^2 \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ 0 & 0 & d_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \quad (8.6)$$

The values along the diagonal represent the variances of each residual (in the population). That they are not necessarily the same value indicates that the variance of the residuals can differ from observation to observation.

The values off the diagonal represent the covariance between the values of the residuals. So, the value at position 1,2 is the covariance between ε_1 and ε_2 , which we symbolized as $\sigma_{1,2}$ in Appendix S (note that these are population values). Since $\sigma_{1,2} = 0$, we are specifying that the two residuals are linearly uncorrelated in the population.

heteroskedastic

independent

Note: Remember that Greek letters refer to the population, while Latin refer to the sample (usually).

Thus, the covariance matrix above specifies that the variances of the residuals are allowed to be different and that the residuals are independent of each other.

This assumption is the only difference between *weighted* least squares and *ordinary* least squares. But, it is a rather significant difference.

Note: Remember that the value of σ^2 can indicate the variance of the population residuals *or* our uncertainty in the value of that residual.

To use weighted least squares (WLS), we need to know the structure of the \mathbf{D} matrix. We do not need to know the exact values, but we need to know them up to a constant multiplier. That is, we need to know the *structure* of that heteroskedasticity. This usually comes from understanding the data-generating process.

Frequently, this is not known (thus, WLS should probably *not* be used). However, there are some cases when we *would* know this structure. For instance, if we are working with a response variable that is a proportion arising from a binomially-distributed variable, we know that the variance is

$$\sigma_i^2 = \frac{\pi(1-\pi)}{n_i} = \pi(1-\pi)\frac{1}{n_i} \quad (8.7)$$

Thus, the diagonal elements will be $d_i = 1/n_i$ and the multiplier (constant part) will be $\pi(1-\pi)$.

8.2.1 FITTING WLS: THE MATHEMATICS Assuming we know the structure of the \mathbf{D} matrix, we can determine all we need to about the WLS estimators and estimates. We just reduce this problem to a previous problem.

To clarify the similarities and differences between ordinary and weighted least squares, here is the classical linear model (CLM) for ordinary least squares:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (8.8)$$

and for weighted least squares:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (8.9)$$

CLM

Those are the same, whether one does OLS or WLS, because both come from the fact you are using the classical linear model. The difference comes in the assumption. Here is the assumption for OLS:

$$\mathbf{E} \sim \mathcal{N}_n(0; \sigma^2 \mathbf{I}) \tag{8.10}$$

Here is the assumption for WLS:

$$\mathbf{E} \sim \mathcal{N}_n(0; \sigma^2 \mathbf{D}) \tag{8.11}$$

Remember that \mathbf{D} is a diagonal matrix.

THE TRANSFORMATION: There is a joke about how a mathematician solved the problem of a hose connected to a fire hydrant:

A mathematician and a physicist were asked the following question:

Suppose you walked by a burning house and saw a hydrant and a hose not connected to the hydrant. What would you do?

P: I would attach the hose to the hydrant, turn on the water, and put out the fire.

M: I would attach the hose to the hydrant, turn on the water, and put out the fire.

Then they were asked this question:

Suppose you walked by a house and saw a hose connected to a hydrant. What would you do?

P: I would keep walking, as there is no problem to solve.

M: I would disconnect the hose from the hydrant and set the house on fire, reducing the problem to a previously solved form.

And so, in the spirit of mathematicians, let us reduce the weighted least squares problem to that of ordinary least squares. If we can do this via a bijective transformation, then we have our confidence intervals and test statistics.

If we define our weighting matrix $\mathbf{W} = \mathbf{D}^{-1/2}$, then our problem is solved, *sans* burning down the house.

Talking Heads

Theorem 8.2.1

Let $\mathbf{E} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{D})$. If we define $\mathbf{W} = \mathbf{D}^{-1/2}$, then

$$\mathbf{WE} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I})$$

Proof. Since \mathbf{W} is a diagonal matrix and \mathbf{E} has a Normal distribution, \mathbf{WE} will also follow a Normal distribution. Thus, we need to calculate $\mathbb{E}[\mathbf{WE}]$ and $\mathbb{V}[\mathbf{WE}]$. In doing this, note that \mathbf{W} is not a random matrix; it is known.

The expected value of \mathbf{WE} is

$$\mathbb{E}[\mathbf{WE}] = \mathbf{W} \mathbb{E}[\mathbf{E}] \tag{8.12}$$

$$= \mathbf{W} \mathbf{0} \tag{8.13}$$

$$= \mathbf{0} \tag{8.14}$$

For the variance we have

$$\mathbb{V}[\mathbf{WE}] = \mathbf{W} \mathbb{V}[\mathbf{E}] \mathbf{W}' \tag{8.15}$$

$$= \mathbf{W} \sigma^2 \mathbf{D} \mathbf{W}' \tag{8.16}$$

$$= \sigma^2 \mathbf{D} \mathbf{W} \mathbf{W}' \tag{8.17}$$

$$= \sigma^2 \mathbf{D} \mathbf{D}^{-1/2} (\mathbf{D}^{-1/2})' \tag{8.18}$$

$$= \sigma^2 \mathbf{D} \mathbf{D}^{-1/2} \mathbf{D}^{-1/2} \tag{8.19}$$

$$= \sigma^2 \mathbf{D} \mathbf{D}^{-1} \tag{8.20}$$

$$= \sigma^2 \mathbf{I} \tag{8.21}$$

In these steps, remember that matrix multiplication is commutative *if* the matrices are diagonal (Theorem M.3.3).

Thus, putting these three parts together gives our conclusion. \square

How do we use this theorem? We pre-multiply the model equation by the matrix \mathbf{W} to obtain the following:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (8.22)$$

$$\mathbf{W}\mathbf{Y} = \mathbf{W}\mathbf{X}\mathbf{B} + \mathbf{W}\mathbf{E} \quad (8.23)$$

Now, redefine the parts to see how useful this result is

$$\mathbf{Y}^* = \mathbf{X}^*\mathbf{B} + \mathbf{E}^* \quad (8.24)$$

with $\mathbf{E}^* \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$ from Theorem 8.2.1. Thus, we can apply all of our OLS results to WLS, as long as we speak to the transformed response variable $\mathbf{W}\mathbf{Y}$ and the transformed independent variable(s) $\mathbf{W}\mathbf{X}$.

This quickly leads to our weighted least squares estimators of \mathbf{B} .

To prove this, we could proceed as we did back in Section 3.1 (page 46). Or, since we have reduced the WLS problem to an OLS problem, we can just write out the results and simplify:

$$\mathbf{b}_{WLS} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^* \quad (8.25)$$

$$= ((\mathbf{W}\mathbf{X})'\mathbf{W}\mathbf{X})^{-1}\mathbf{W}\mathbf{X}'\mathbf{W}\mathbf{Y} \quad (8.26)$$

$$= (\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{Y} \quad (8.27)$$

$$= (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y} \quad (8.28)$$

It also quickly leads to showing that the WLS estimator is unbiased for \mathbf{B} :

Theorem 8.2.2

Under the assumptions of weighted least squares, the WLS estimator for \mathbf{B} is unbiased.

Proof. I am tempted to give this to you as an exercise, but let's see how to prove it.

$$\mathbb{E}[\mathbf{b}_{WLS}] = \mathbb{E}\left[(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}\right] \quad (8.29)$$

Remember that the \mathbf{D} matrix is known, is *not* a random variable.

$$= (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}^{-1}\mathbb{E}[\mathbf{Y}] \quad (8.30)$$

Since $\mathbf{WY} = \mathbf{WXB} + \mathbf{WE}$, and since \mathbf{W} and \mathbf{D} are invertible (why?), we have

$$= (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}^{-1}\mathbf{XB} \quad (8.31)$$

$$= \mathbf{B} \quad (8.32)$$

Thus, the WLS estimator is unbiased if \mathbf{D} is invertible and if \mathbf{D} is known (non-stochastic). I will leave it as an exercise for you to prove this theorem if \mathbf{D} is a random variable independent of \mathbf{X} . \square

exercise

Theorem 8.2.3

Under the assumptions of weighted least squares, the variance of the WLS estimator for \mathbf{B} is

$$\mathbb{V}[\mathbf{b}_{WLS}] = \sigma^2 (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} \quad (8.33)$$

Proof. There should be no surprises with this proof. All you have to do is figure out what is a random variable and what is not. As such, I leave it as an exercise for you.

exercise

So very generous of me. =) \square

Note that the WLS estimator of \mathbf{B} is a linear combination of independent Normal random variables. With that final observation, we have the distribution of the WLS estimator of \mathbf{B} :

$$\mathbf{b}_{WLS} \sim \mathcal{N}\left(\mathbf{B}, \sigma^2 (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\right) \quad (8.34)$$

exercises

Note: We again note that the individual estimators are not independent of each other under typical circumstances. We also note that the confidence intervals for the estimators, estimates of γ , etc. can easily be determined in the WLS realm. Nothing new is here, only the mathematics is a bit more involved.

8.2.2 THE REAL QUESTION Weighted least squares takes care of the problem of heteroskedasticity in our data without introducing any major change in our modeling process or understanding. It just requires that we determine \mathbf{W} and transform our dependent and independent variables by premultiplying by that weighting matrix.

Question

How do we obtain that weighting matrix?

The best way of obtaining it is through **theory**. The second best way is to utilize the hat matrix, \mathbf{H} .

THEORY: Frequently, knowledge of the problem suggests the weighting matrix. Recall that the $\mathbb{V}[\mathbf{E}]$ covariance matrix measures our uncertainty in the residuals. If that uncertainty is known by the way the experiment is constructed, then \mathbf{W} can be determined.

For instance, if the dependent variable is the result of a Binomial experiment, perhaps it is the number of successes out of a given number of trials (which may change), then the weighting matrix is just a diagonal matrix of the square root of trial sizes.

Why? Recall that the variance of a binomially-distributed random variable is $\sigma^2 = \frac{\pi(1-\pi)}{N}$. The π are the unknown (constant) population proportion. The N_i is the (known) size within group i . The population parameter is assumed constant. The sample size is measurable.

This leads to the \mathbf{D} matrix being of the form

$$\mathbf{D} = \begin{bmatrix} 1/N_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/N_2 & 0 & \cdots & 0 \\ 0 & 0 & 1/N_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/N_n \end{bmatrix} \quad (8.35)$$

THE HAT MATRIX: When we do not have the theory to know the structure of the \mathbf{D} matrix, one may want to use the hat matrix to give us a hint about its structure.

Warning: *Make no mistake. This process is not mathematically correct and perfect... but what statistics procedure is? Statistics stands astride the real and the ideal, trying to get as much information about the real while acknowledging its limitations.*



Remembering Chapter 5, not all violations affect inferences the same. Perhaps a good thing for you to do is to use the processes of Chapter 5 to see how much using the hat matrix in lieu of a theoretically-driven \mathbf{D} matrix affects the estimates, confidence intervals, and p-values.

Let us use the symbol \mathbf{e} to represent the *observed* residuals. Until this point, we have only been working with the *theoretical* residuals, \mathbf{E} . The conceptual difference between the two is really just the difference between the population (theoretical) and the sample (observations). In effect, the difference is in terms of the variances.

Question

What is the variance of \mathbf{e} ?

Theorem 8.2.4

The variance of \mathbf{e} is $\mathbb{V}[\mathbf{e}] = \sigma^2(\mathbf{I} - \mathbf{H})$.

Proof.

$$\mathbb{V}[\mathbf{e}] = \mathbb{V}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] \quad (8.36)$$

$$= (\mathbf{I} - \mathbf{H})\mathbb{V}[\mathbf{Y}](\mathbf{I} - \mathbf{H})' \quad (8.37)$$

$$= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H}) \quad (8.38)$$

$$= \sigma^2(\mathbf{I} - \mathbf{H}) \quad (8.39)$$

□

So, that was totes cool. What was its purpose? What does it mean?

Remember how we can interpret that variance. It is either the variance of a gazillion observed residuals, *or* we can see it as the uncertainty inherent in the measured residual.

uncertainty

For example, the inherent uncertainty in the first residual can be estimated as

$$s_{1,1} = \text{MSE}(1 - h_{1,1}) \quad (8.40)$$

Here, $h_{1,1}$ is the first element of the diagonal of the hat matrix.

That means, those diagonal elements of $\mathbf{I} - \mathbf{H}$ indicate (or are estimates of) the precision of the y estimate for a given value of x . An estimate of the structure of the \mathbf{D} matrix is just the diagonal of the $\mathbf{I} - \mathbf{H}$ matrix.

precision

Note: The problem is that weighted least squares requires us to *know* the \mathbf{D} matrix, not that we estimate it from the data. This explains why the hat matrix technique is used only until something better comes along.

It does work nicely, but we statisticians like to “see the math” — sometimes. Also, if we are trying to draw important conclusions, using approximate methods tends to undercut the conclusions for many, especially for those who do not really understand statistics.

Question

What do the off-diagonal elements of $\mathbf{I} - \mathbf{H}$ estimate?

8.3: Generalized Least Squares

Both ordinary least squares and weighted least squares requires the errors be independent. Reality does not always meet this requirement. If the dependent variable consists of repeated measures on one unit over time, such as in modeling stock prices, it is quite likely that the residuals will be correlated. Also, if the dependent variable is measured on geographic structure, such as states in a country or trees in a forest, it is also likely that errors of near units are correlated.

In such examples, the covariance matrix of \mathbf{E} will *not* be diagonal. Thankfully, it is a covariance matrix, and therefore positive definite under the usual assumption of no multi-collinearity (Appendix M, Section M.5.1). Since it is positive definite, it is invertible. Thus, we can do a trick not unlike what we did for weighted least squares.

For a reminder, here are the model equations for ordinary, weighted, and general least squares:

Ordinary Least Squares	$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$
Weighted Least Squares	$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$
General Least Squares	$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$

They sure do look similar. That's because this is the classical linear model (CLM). The requirements on the residuals differs, however:

Ordinary Least Squares	$\mathbf{E} \sim \mathcal{N}(\mathbf{0}; \sigma^2\mathbf{I})$
Weighted Least Squares	$\mathbf{E} \sim \mathcal{N}(\mathbf{0}; \sigma^2\mathbf{D})$
General Least Squares	$\mathbf{E} \sim \mathcal{N}(\mathbf{0}; \mathbf{\Sigma})$

For ordinary least squares, the covariance matrix of the residuals is a constant multiple of the identity matrix, \mathbf{I} . This indicates the residuals are independent and have the same variance (uncertainty). For weighted least squares, the covariance matrix of the residuals is a constant multiple of a diagonal matrix, \mathbf{D} . This indicates the residuals are independent, but possibly with unequal variances.

For generalized least squares (GLS), the covariance matrix is (a constant multiple of) a symmetric, positive definite matrix, $\mathbf{\Sigma}$. This indicates the residuals are possibly correlated and with possibly unequal variances.

As with weighted least squares, you *do* need to know the structure of the covariance matrix. This requirement is sometimes met by the structure of the problem. The following are two examples showing how one *can* determine the $\mathbf{\Sigma}$ matrix.

An understanding of these examples is not needed. They are here only to illustrate that there are times Σ can be determined from the problem.

8.3.1 TIME SERIES ISSUES When data are collected on a single unit over time, the measurements will tend to be correlated. For instance, the unemployment rate in Ruritania over the past 20 years is 11.35, 11.41, 11.12, 11.08, 10.93, 10.86, 10.96, 11.05, 11.10, 10.87, 10.79, 10.76, 10.94, 10.94, 10.92, 11.01, 11.04, 11.16, 11.13, and 11.14.

Solution: Let us fit this using ordinary least squares regression, then examine the residuals for autocorrelation (correlation between subsequent values).

```
unemp = c(11.35, 11.41, 11.12, 11.08, 10.93, 10.86, 10.96,
          11.05, 11.10, 10.87, 10.79, 10.76, 10.94, 10.94, 10.92,
          11.01, 11.04, 11.16, 11.13, 11.14)
year = 1:20

mod = lm(unemp ~ year)
E = residuals(mod)

autocor.test(E)
```

Note the sample autocorrelation is 0.719 with a p-value of 0.0005 and a 95% confidence interval from 0.393 to 0.884. The p-value indicates the autocorrelation is not 0. The confidence interval indicates that the residuals are moderately-to-highly correlated.

In other words, adjacent observations are not independent, as both ordinary and weighted least squares require. Really, this makes sense because next year's unemployment rate will be heavily influenced by this year's rate.

There are many ways of modeling such a situation. One is called “Autoregressive-1” or AR(1) or ARIMA(1,0,0). This model assumes that the primary correlation is only directly between adjacent years. The covariance matrix, Σ , would have this structure if the correlation between those adjacent years is $\rho = 0.500$:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 & \dots \\ 0.5 & 1 & 0.5 & 0.25 & \dots \\ 0.25 & 0.5 & 1 & 0.5 & \dots \\ 0.125 & 0.25 & 0.5 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ & & & & & 1 \end{bmatrix} \quad (8.41)$$

You can get this particular matrix using this R code:

```
||| Sigma = diag(20)
||| Sigma = 0.5^abs(row(Sigma)-col(Sigma))
```

Note that the matrix has 1s along the diagonal and higher powers of 0.5 farther from the diagonal. The zeroes arise from the fact that the matrix is 20×20 ; that is, e.g., the entry in cell (1,20) is actually $0.50^{19} \approx 0$. ♦

Note: Again, this was just an example to show that the structure of Σ can be determined from some problems. There are entire sub-disciplines of statistics that examine such serial correlation. This sub-discipline is called “time series.”



Figure 8.1: A map showing the administrative divisions (*kraj*) of the Kingdom of Ruritania. For this example, note that no *kraj* abuts all other *kraj*.

8.3.2 GEOGRAPHIC ISSUES When data are collected from geographical units, such as neighborhoods, counties, or states, the residuals may be spatially correlated. This is a violation of the independence assumption of ordinary least squares.

How that geographic correlation is modeled is up to the expert (researcher). The subject of spatial modeling is extensive and quite interesting... and important. It can, with appropriate matrices, be extended to modeling three-dimensional spatial correlation over time. If you have the opportunity, I suggest studying this topic (Bivand, Pebesma, and Gómez-Rubio 2013, Blangiardo and Cameletti 2013, Sen 2016). If nothing else, it leads to fun maps!

Figure 8.1 is a map of Ruritania showing the nine *Kraj*. Note that some *kraj* abut some *kraj* but not others. For instance, region CS does not touch region CC.

If we are trying to model the spread of something (disease, unemployment, wealth), we *may* decide to take into consideration the fact that some units neighbor others. Thus, from the map above, we know there is a first-level transmission between CS and CD but not between CS and CF.

Example 1

Geographical Data Let us determine a matrix describing the adjacencies for the nine *kraj*.

Solution: Check that the following is the adjacency matrix for Ruritania¹

$$\Sigma = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad (8.42)$$

It is important to ensure the kraj order for the columns is the same as for the rows. The kraj ordering is: CA, CB, CC, CD, CE, CI, CM, CS, CSM. \blacklozenge

symmetry

Note that the adjacency matrix is symmetric. Why will this matrix always be symmetric, regardless of the map? Now, that we know Σ is symmetric, we can use the discussion following Lemma M.10 to conclude that Σ is positive definite, which indicates it is allowable as a covariance matrix.

Note: We do not know the constant multiplier, σ^2 . No probs. We only need to know the structure of the covariance matrix. We use the data to estimate the constant multiplier σ^2 .

Also, note that the analysis based on this covariance structure is only as good as our assumption that the contagion spreads through touch. If it spreads based on some sort of distance, then the Σ is not correct and we will need to create an appropriate covariance matrix *given our scientific understanding...* if such exists.

Finally, let me reiterate a point I made above. The purpose of this example is *only* to illustrate that these covariance structures *can* be determined from the problem without resorting to estimating them from the data.

Note: However, there is a lesson for all of us here. If we do not know the correct structure of the correlation matrix, then we should use several and see how sensitive our estimates, confidence intervals, and p-values are to that matrix. The results may be very sensitive, which is not a good position to be in, especially if we do not know the right mode of transmission.

¹One area of **geographical analysis** tries to decide what adjacency rules are appropriate for a given research question. This example uses a simple 0-1 scheme. Other schemes include distances (measured in some manner).

If the estimates, etc. are not sensitive to our choice of covariance matrix, then we need not be as concerned.

The rule is to explore all models that make sense and see how important our assumptions are to our results.

8.3.3 ONE MORE NOTE Above, we have only focused on being able to determine the structure of the covariance matrix (at least to a scalar multiple). There is one more thing that we need to pay attention to: What is the square root of Σ ?

When we introduced $\mathbf{D}^{1/2}$ in Section 8.2, we knew we could calculate it. After all, one square root for a diagonal matrix would be

$$[\mathbf{D}^{1/2}]_{i,j} = d_{i,j}^{1/2} \quad (8.43)$$

That is, the elements of the square root matrix are the square root of the entries of the matrix. This shortcut works because \mathbf{D} is diagonal.

In general, Σ does not have to have a well-defined square root. Some do, but some do not. Without $\Sigma^{1/2}$, calculating the GLS estimates is not possible using this method.

Sadness abounds.

8.4: Full Example: May the (Strong) Force be with You

Ruritania is a patron of high-energy physics — and of Star Wars enthusiasts. King Rudolph donated several million crowns to Switzerland to aid in researching the strong force.

That money was used at CERN (the *Conseil Européen pour la Recherche Nucléaire*) for several experiments. Each experiment consisted of a beam of protons crashing into a target. That beam had a constant energy level. What changed were the target sizes and the energy level of the proton after the collision. Many experiments were run at each energy level, and the standard deviation of the energies was measured.

In a theory proposed by Ruritanian scientists that is not entirely clear to His Majesty (or to your author), there should be a linear relationship between the cross sectional area and the inverse of the energy. The data are given in Table 8.1.

The first column is the value of the independent variable. The second column is the mean of the energy level of the photon after the collision. The third column is the standard deviation in those energy levels. Note that the variability at each cross-section differs. This is based on both the number of experiments *and* the inherent variability at that area.

Cross Section [b]	Energy [MeV]	St. Dev. [MeV]
1	848.9	7.8
2	476.9	9.2
3	350.9	9.4
4	289.2	10.2
5	251.7	7.4
6	225.8	9.3
7	209.7	7.2
8	193.9	5.3

Table 8.1: Data for the example regarding the strong nuclear force. Units are given in brackets.

8.4.1 ORDINARY LEAST SQUARES Let us ignore the different uncertainties in each energy level (the standard deviations). That is, let us just fit this as an OLS model.

Here is the code:

```
barns = c(1,2,3,4,5,6,7,8)
energy = c(848.9,476.9,350.9,289.2,251.7,225.8,209.7,193.9)
Ibarns = 1/barns

modOLS = lm(energy ~ Ibarns)
summary(modOLS)
confint(modOLS)
```

The output suggests that the relationship between the cross sectional area of the target and the inverse of the resulting energy of the photon is statistically significant.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  101.9148    0.5464   186.5 1.60e-12 ***
Ibarns       747.5306    1.2505   597.8 1.48e-15 ***
---
Residual standard error: 0.9719 on 6 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
F-statistic: 3.574e+05 on 1 and 6 DF, p-value: 1.479e-15
```

A 95% confidence interval for the relationship is from 744.5 to 750.6 (with units of MeV·barn).

8.4.2 WEIGHTED LEAST SQUARES Note, however, that the uncertainty in the measurements varies. We are more uncertain with some of our estimated energy than with others. If we do not take this uncertainty into consideration, we may be biasing our results. To include this information, we can use weighted least squares regression.

The code to fit with weighted least regression is as follows:

```
barns = c(1, 2, 3, 4, 5, 6, 7, 8)
energy = c(848.9, 476.9, 350.9, 289.2, 251.7, 225.8, 209.7, 193.9)
stdev = c(7.8, 9.2, 9.4, 10.2, 7.4, 9.3, 7.2, 5.3)
v = stdev^2

Ibarns = 1/barns

modWLS = lm(energy ~ Ibarns, weight=1/v)
summary(modWLS)
confint(modWLS)
```

Note that we are 95% confident the effect of the target's cross section on the resulting energy is from 744.5 to 751.2 MeV·barn.

Note: In R, as in many statistical programs, the weights you provide in the function call are inversely proportional to the variances. This is why we used `weight=1/v` in the function call.

With that being said, **always check the documentation** to make doubly sure. Frequently, this information is difficult to find and an error will not be thrown to let you know.

How do we know this?

Or, an even better question:

Question

How do we check that the weights really are proportional to the inverse of the variance?

The short answer is to check your work “by hand.” As we once did OLS by hand using the matrix functions in R, we can also do WLS by hand.

As always, make sure you know what each line does and why they are put together as they are:

```
## The data
barns = c(1,2,3,4,5,6,7,8)
energy = c(848.9,476.9,350.9,289.2,251.7,225.8,209.7,193.9)
stdev = c(7.8,9.2,9.4,10.2,7.4,9.3,7.2,5.3)

## A few minor calculations
v = stdev^2
Ibarns = 1/barns
n = length(Ibarns)

## The needed matrices
Y = matrix(energy, ncol=1)
X = matrix( c(rep(1,n), Ibarns), ncol=2 )
colnames(X) = c("b0", "b1")
D = diag(v)

## The estimate vector
solve(t(X) %*% solve(D) %*% X) %*% t(X) %*% solve(D) %*% Y
```

When I run this, I get the following output

```
      [,1]
b0 101.6074
b1 747.8364
```

These are the matrix calculations according to Equation 8.28 on page 234. Thus, this is the *correct* answer. Double-check that this known correct answer matches the answer given in `modWLS`. If it does not, then you will need to change the expression in the `weights` part of the function call. In R, it will match. In other pieces of software, it may not. **Be aware.**

Note: The difference between the effects estimated from using ordinary least squares and using weighted least squares is rather minor in this example. It need not be, as the next example shows.

8.5: Full Example: Elections in Ruritania

Even though it is an absolute monarchy, national elections are held in Ruritania to elect members of the Ruritanian parliament, the *Národní Shromáždění* (National Assembly). There are many parties represented in the parliament, but the party that consistently receives a majority of the seats and votes is the monarchist *Pohyb pro Ruritánii* (PR; Movement for Ruritania).

The main opposition party is the *Demokratické Hnutí* (DH; Democratic Movement) party, but votes are also usually received by the *Socialistická* (SP; Socialist Party), *Křstanská Demokratická* (KD, Christian Democratic), and *Republikánská* (RS, Republican) parties.

It is fortuitous that Ruritania does not use computerized ballots. They use ballot papers for the parliamentary election that consist of the party names, symbols, and abbreviations. . . and a box for the voter to place their inked fingerprint next to the party. After voting, the ballot is placed in a ballot box to await counting.

At the end of the evening, the ballot papers at each precinct are securely transported to the division headquarters, where they are counted by electoral officials. Each ballot is checked by that official to ensure that it was lawfully cast and that the “will of the voter” can be discerned.

When the division is finished counting the ballots, the totals are then telephoned to the Independent Electoral Commission (*Nezávislá Volební Komise*, NVK) in the capital. With much pomp and circumstance, and not a little fanfare, the division totals are added and reported to the people.

After the last election, the exiles in Denmark claimed that the ballot boxes were stuffed. That is, the ballot boxes had votes for the PR party in them before voting began. Because guarantees of the secret ballot are built into the Ruritanian Constitution, the ballot boxes are opaque.

In other words, direct evidence of ballot box stuffing does not exist, only claims by those who live in exile in another country (Denmark). However, if ballot box stuffing existed in this election to any great extent, it would leave evidence. Why/how?

Question

What do stuffing ballots have that naturally cast ballots do not?

Easy: The stuffing ballots are all for the ruling PR party *and* they are all completed (filled in) correctly. The naturally cast ballots will consist of votes for all parties and will include ballots not filled-in correctly.

And so, in the presence of systematic and significant ballot box stuffing, there will be a relationship between the invalidation rate and the level of support for the ruling party.

That is the theory. The exiles are paying for this analysis. We like the money, so we need to be confident — and clear — in our conclusions. The NVK is providing the official counts in the `rur2013parl` data file, so we need to ensure that the statistical analysis is clean. That is, it is up to us to do the analysis correctly, neither concluding too much nor too little.

And, as always, being clear in our reasoning.

8.5.1 ORDINARY LEAST SQUARES The first analysis we will do is ordinary least squares. The dependent variable is the invalidation rate; the independent variable is the support for the Movement for Ruritania (PR) party. Why them? Since they were in charge before the election, they were in position to stuff the ballot boxes.

Here is the code to load the data, create the variables, fit the model, and determine if a relationship between the invalidation rate and PR support rate can be detected.

```
votes=read.csv("http://rur.kvasaheim.com/data/rur2013parl.csv")
attach(votes)

Valid = Total-Invalid
pPR   = PR/Valid
pInv  = Invalid/Total

modOLS = lm(pInv ~ pPR)
summary(modOLS)
```

These results indicate that we did not detect a relationship. . . at the $\alpha = 0.05$ level ($p = 0.0668$). Thus, ordinary least squares *did not detect* unfairness in the vote.

Note: It is important to emphasize here that the correct terminology is that we did not detect unfairness. We cannot say there was no fairness. We can only say we didn't detect it.

Remember to check the assumptions. This point cannot be over-emphasized. If the assumptions are not met, then the model is not correct. Well, not *perfectly* correct. See Chapter 5 for a discussion of this point.

8.5.2 WEIGHTED LEAST SQUARES Note that ordinary least squares is *not* be the best option here. The invalidation rate has greater inherent variability in smaller divisions than in larger. We know this because of the distribution of the invalidation rate. Invalidation counts follow something akin to a Binomial distribution. Its two parameters are sample size (number of votes cast) and success probability (invalidation rate). The variance of a Binomial random variable is $n\pi(1 - \pi)$.

Dividing the invalidation count by the number of votes cast gives the invalidation rate. The distribution of the invalidation rate can be approximated with a Normal distribution (see the Central Limit Theorem, Section S.6.4). The expected value of the observed invalidation rate is π , the inherent invalidation rate. The variance is $\pi(1 - \pi)/n \propto 1/n$.

Because the data are heteroskedastic in nature, and because the structure of the heteroskedasticity is known, weighted least squares will be more appropriate here.

Here is the code. Compare it to the ordinary least squares code from above.

```
Valid = Total-Invalid
pPR   = PR/Valid
pInv  = Invalid/Total

modWLS = lm(pInv ~ pPR, weights=Total)
summary(modWLS)
```

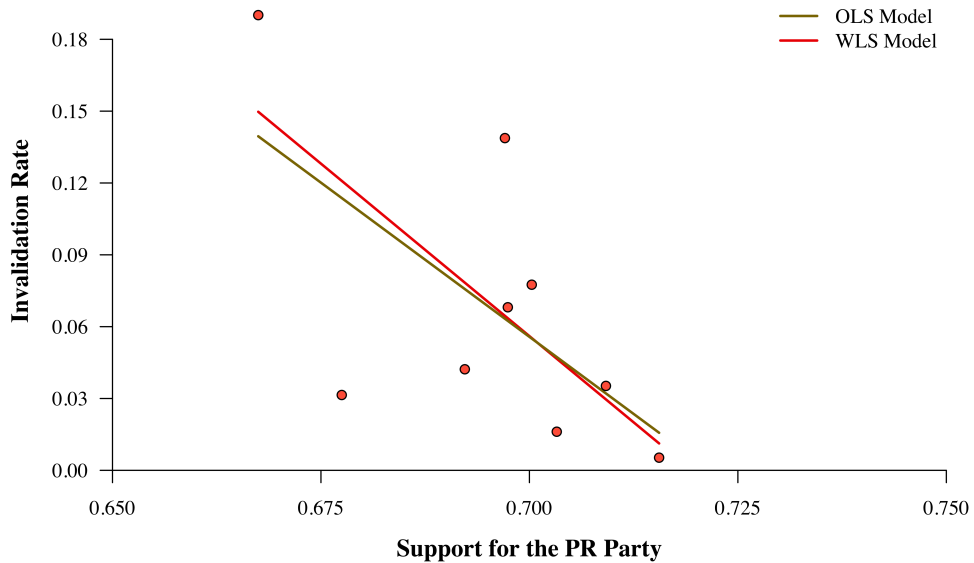


Figure 8.2: An invalidation plot for the 2013 Ruritania parliamentary election. The lines of best fit are provided. The OLS fit is in brown and the WLS is in red.

This produces the following (abbreviated) output:

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.0726	0.7109	2.915	0.0225 *
pPR	-2.8808	1.0124	-2.845	0.0249 *

Note that this method *did* detect a relationship between the invalidation rate and the PR party support rate ($p\text{-value} < 0.05 = \alpha$). Furthermore, that differential invalidation helped the ruling party. The negative coefficient indicates that those *kraj* with higher PR support also tended to count more of the votes (reject fewer). Thus, we can conclude that the data are consistent with the exile claim of ballot box stuffing. Figure 8.2 illustrates this.

Question

How much is the differential invalidation?

Well, from the regression output, we know that when the support for PR increases by 10 percentage points (from 65 to 75%, for instance), the invalidation drops by an average of 27pp.

That seems rather substantial to me.

proof

Note: Nothing in statistics ever constitutes *proof*. Nothing. Ever. Period. End of thought. *Žádné další!*

Statistics only provides evidence in favor of — or against — the null hypothesis. In this case, the p-value is 0.0249. If the null hypothesis is correct, then we would observe results this extreme or more so 2.49% of the time. This is not too rare, especially when you realize you are claiming the government cheated. Cheating is a more serious claim than just that someone was mistaken.

It is always better to report the results, interpret the results, be explicit that there is no proof and that the null hypothesis has a non-zero probability of being reality.

Do not live your statistical life ruled by $\alpha = 0.05$. Realize — and accept — that the p-value is a measure of how well the data support the null hypothesis, the hypothesis of no relationship/difference/effect/evidence.

8.6: Conclusion

In the previous chapters, we focused on ordinary least squares. This method required that the residuals were independent and identically distributed. From that assumption, we were able to generate a series of rich conclusions.

However, it is not true that residuals are always identically distributed or independent. While we did find a way of “fixing” the problem of heteroskedasticity, it is frequently better to use a modeling scheme that uses that heteroskedasticity instead of merely finding a way of ignoring it. This is what weighted least squares does. If you have theory behind how the variances should vary for each record, you can use this method. If not, then you are reduced to the “fixes” of Chapter 7.

Similarly, if your data are not independent, but you understand the structure of that dependence, you can use generalized least squares to model the relationship better... as long as that covariance matrix has an inverse square root matrix. And, there is no guarantee that it does.

8.7: End-of-Chapter Materials

8.7.1 R FUNCTIONS In this chapter, we were introduced to a few R functions that will be useful in the future. These are listed here.

PACKAGES:

nlme This package gives R the functionality to fit generalized least squares using the `gls` function. It actually has many other useful functions that allow us to fit non-linear models and random-effects models. Those are beyond the scope of this book, however.

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

STATISTICS:

autocor.test(e) This function calculates the auto-correlation, which is just the correlation between sequential values in the vector. It is a part of the `RFS` package.

gls(formula) This function performs generalized least squares regression. It even allows you to specify the correlation structure via the `correlation` parameter.

lm(formula) This function performs linear regression on the data, with the supplied formula. If you specify the `weights`, then they are applied and you are fitting the model using weighted least squares. As there is much information contained in this function, you will want to save the results in a variable.

residuals(mod) This calculates the simple residuals in a model, the observed values minus the predicted values.

MATHEMATICS:

%*% This multiplies two matrices in \mathbb{R} . Thus, running the command **A**B** will return the matrix product **AB**.

abs(x) This returns the absolute value of the real number x , a.k.a. $|x|$.

column(A) This returns the column number of the matrix **A**.

diag(n) If n is an integer, then this returns the \mathbf{I}_n identity matrix.

diag(v) This returns a diagonal matrix with the elements of the vector **v** along the diagonal.

diag(A) This returns the diagonal entries of the matrix **A**.

rep(n, x) This returns a vector of the number x repeated n times.

row(A) This returns the row number of the matrix **A**.

solve(A) This returns the inverse of the matrix **A**.

t(A) This returns the transpose of the matrix **A**.

8.7.2 EXERCISES

1. Let $\mathbf{E} \sim \mathcal{N}(\mathbf{0}; \sigma^2 \mathbf{D})$ be the residuals. Prove that if \mathbf{D} is a diagonal covariance matrix, then it is invertible.
2. Let $\mathbf{E} \sim \mathcal{N}(\mathbf{0}; \sigma^2 \mathbf{D})$ be the residuals. Here, \mathbf{D} is a diagonal covariance matrix. Determine a matrix \mathbf{W} such that $\mathbf{W}\mathbf{W} = \mathbf{D}$.
3. Prove Theorem 8.2.2.
4. Under the assumptions of weighted least squares, determine the formula for a confidence interval for β_1 .
5. What is the difference between \mathbf{e} and \mathbf{E} ?
6. Under the assumptions of generalized least squares, determine the formula for the estimator of \mathbf{B} .
7. Under the assumptions of generalized least squares, determine the formula for a confidence interval for \mathbf{b} .
8. Determine if Theorem 8.2.1 holds if the weights matrix \mathbf{D} is a random matrix independent of \mathbf{X} . If it does not, what is the distribution of $\mathbf{W}\mathbf{E}$?
9. Prove Theorem 8.2.2 if \mathbf{D} is independent of \mathbf{X} .
10. Theorem 8.2.3 requires \mathbf{D} is non-random. Determine the variance of \mathbf{b}_{wls} if \mathbf{D} is random, but independent of \mathbf{X} .
11. In Example 8.3.2, I state that the adjacency matrix is symmetric. Explain why this is so.

8.7.3 THEORY READINGS

- Adrian Baddeley, Ege Rubak, and Rolf Turner. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton, FL: Chapman & Hall/CRC.
- Roger S. Bivand, Edzer Pebesma, and Virgilio Gómez-Rubio. (2013) *Applied Spatial Data Analysis with R*. New York: Springer-Verlag.
- Marta Blangiardo and Michela Cameletti. (2015) *Spatial and Spatio-temporal Bayesian Models with R*. Hoboken, NJ: John Wiley & Sons.
- Chris Brunsdon and Lex Comber. (2015) *An Introduction to R for Spatial Analysis and Mapping*. Thousand Oaks, CA: SAGE Publications.
- Robert Haining. (2003) *Spatial Data Analysis: Theory and Practice*. Cambridge, UK: Cambridge University Press.
- Tonny J. Oyana and Florence Margai. (2013) *Spatial Analysis: Statistics, Visualization, and Computational Methods*. Boca Raton, FL: Chapman & Hall/CRC.
- Zekai Sen. (2016) *Spatial Modeling Principles in Earth Sciences*. New York: Springer-Verlag.
- Thorsten Wiegand and Kirk A. Moloney. (2013) *Handbook of Spatial Point-Pattern Analysis in Ecology*. Boca Raton, FL: Chapman & Hall/CRC.

CHAPTER 9:

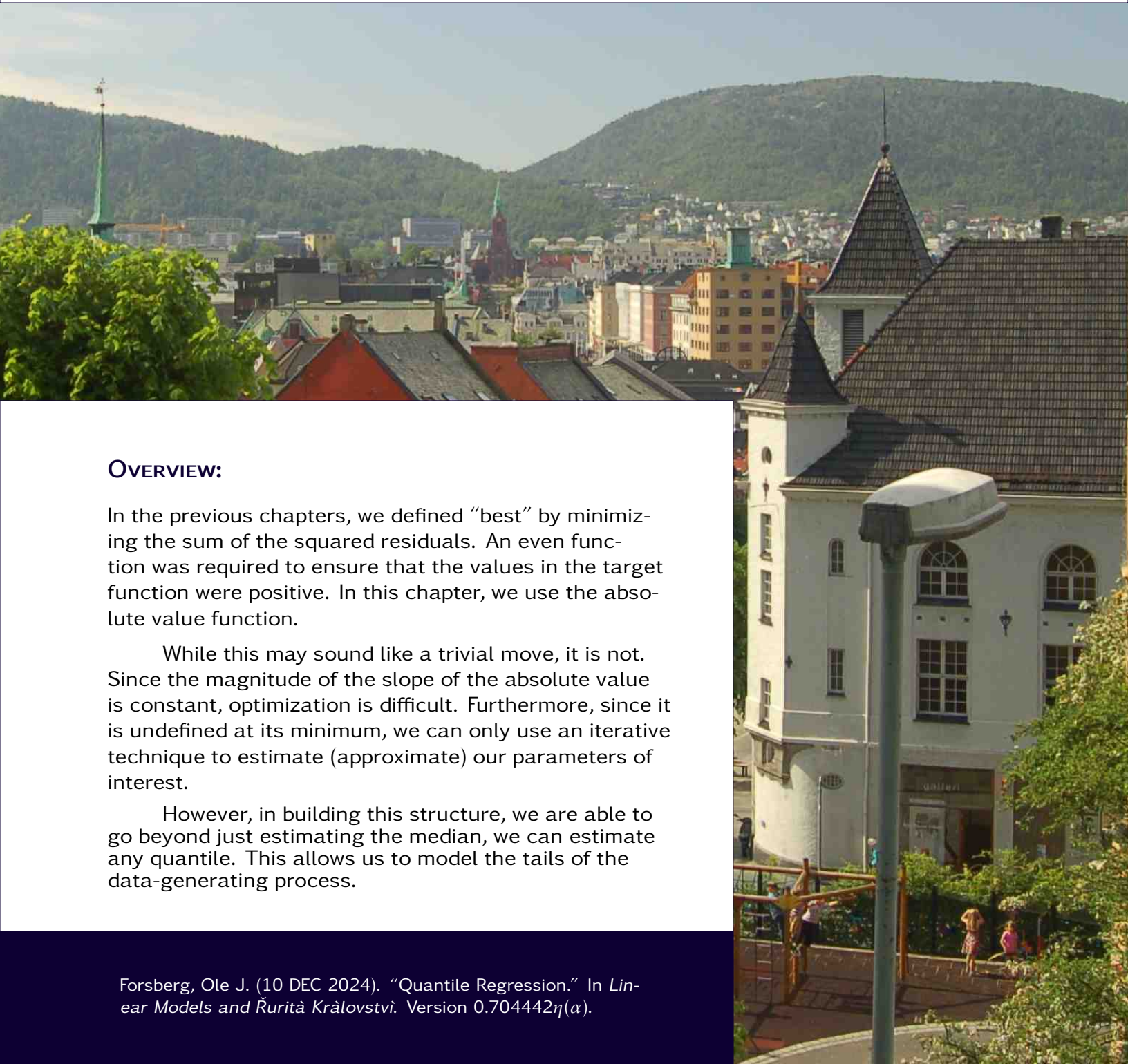
QUANTILE REGRESSION

OVERVIEW:

In the previous chapters, we defined “best” by minimizing the sum of the squared residuals. An even function was required to ensure that the values in the target function were positive. In this chapter, we use the absolute value function.

While this may sound like a trivial move, it is not. Since the magnitude of the slope of the absolute value is constant, optimization is difficult. Furthermore, since it is undefined at its minimum, we can only use an iterative technique to estimate (approximate) our parameters of interest.

However, in building this structure, we are able to go beyond just estimating the median, we can estimate any quantile. This allows us to model the tails of the data-generating process.



Chapter Contents

9.1	Parameter Estimation	263
9.2	Quantile Regression	269
9.3	Conclusion	275
9.4	End-of-Chapter Materials	276



least squares

In the previous sections we examined three types of least squares regressions — ordinary, weighted, and general. These three estimation methods have one thing in common: The estimates were obtained by minimizing the sum of squared residuals (properly weighted). We used the squaring function for two reasons. First, it is everywhere differentiable, especially at its minimum. Second, squaring the residuals ensures that you are adding non-negative values. All even functions attain the second goal. The class of functions that meet the first requirement is more restrictive.

The higher the even power, the more outliers affect the estimates; that is, the outliers will tend to have an increased effect on the estimator when the power is larger. One option to reduce the effect of these outliers is to use a different even function. The absolute value function has been used quite successfully in the past.

Unfortunately, the absolute value function is not everywhere differentiable. Even worse: it is not differentiable at its minimum — the point of interest. This means we cannot obtain a simple set of equations for our estimators. We can still, however, obtain estimators to an arbitrary degree of precision by using a set of equations that get us closer and closer to the true value of the estimate.

9.1: Parameter Estimation

Let us first think about how we *could* do this by hand:

In least squares, we just did calculus to get equations for the estimators. Here, since such solutions do not exist and since we need to use an iterative technique, I think working through a toy example may help understanding. And so, let us start with the data in the left two columns of Table 9.1.

Remember that we want to minimize the sum of the absolute values of the residuals.¹ Thus, the first step is to obtain residuals. This means we need to somehow obtain our first estimated regression line. Any will work as a starting point. So, let's say our first estimate of the line-of-best-fit is $\ell_1 : y = 3$, which is just the horizontal line at the median.

The next step is to calculate the residuals. This is the e_1 column in Table 9.1. The the target function is

$$Q_1 = \sum_{i=1}^n |e_i| \quad (9.1)$$

$$= \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9.2)$$

$$= \sum_{i=1}^n |y_i - 3| \quad (9.3)$$

$$= |y_1 - 3| + |y_2 - 3| + |y_3 - 3| + |y_4 - 3| \quad (9.4)$$

$$= |1 - 3| + |3 - 3| + |3 - 3| + |9 - 3| \quad (9.5)$$

$$= 2 + 0 + 0 + 6 \quad (9.6)$$

Thus, for line ℓ_1 , the value of the target function is 8.

L_1 Norm

¹In other words, we want to minimize the L_1 distance between the n -dimensional data vector and the p -dimensional parameter space. Recall Figure 3.2 where we illustrated this with least squares.

x	y	\hat{y}_1	e_1	\hat{y}_2	e_2	\hat{y}_3	e_3
0	1	3	-2.0	1.5	-0.5	0.0	1.0
1	3	3	0.0	2.5	0.5	2.0	1.0
2	3	3	0.0	3.5	-0.5	4.0	-1.0
3	9	3	6.0	4.5	3.5	6.0	3.0

Table 9.1: Raw data and a few columns of the median regression estimation process. This is more heuristic than actual. The actual fitting method depends on the program used.

The next step is to change the regression line. How? Well, that is the important question. Different methods may ultimately lead to slightly different answers. As this section only seeks to illustrate a method — and not even a good method — let’s use logic to see what would be next. Note that the lower values have estimates that are too low, and the higher values have estimates that are too high. So, it makes sense to increase the slope. So, let us increase the slope to 1. If we force the line to pass through the dimension-wise median $(\bar{x}, \bar{y}) = (1.5, 3.0)$, the linear equation will be $\ell_2 : y = 1.5 + 1x$. This produces the estimates and residuals in the next two columns of Table 9.1.

The value of the target function is

$$Q_2 = \sum_{i=1}^n |e_i| = \sum_{i=1}^n |y_i - \hat{y}_i| = \sum_{i=1}^n |y_i - (1.5 + x_i)| \quad (9.7)$$

Note that this value is 5. As this is lower than the previous value, we headed in the right direction; we are closer to the estimates because we have reduced the sum of the absolute errors.

We got closer. Note that the error for higher x -values is greater than for lower x -values. This suggests we should increase the slope yet again. So, let us select our third line as $\ell_3 : y = 0 + 2x$. Again, we are forcing the line to pass through the dimension-wise median.² The last pair of columns in Table 9.1 provide the predictions and residuals for this third line.

The value of the target function for this third line is $Q_3 = 6$. This value is not lower than Q_2 . Thus, this line is a worse fit than line ℓ_2 . The next line, ℓ_4 , needs to take this into consideration.

This process would continue until the change in target function values is “small enough.” Usually, we define “small enough” as being less than some tolerance, like $\tau = 0.000001$.

²Do we need to do this? No. There are algorithms that do not force this restriction. Again, the actual mathematics cannot reasonably be done by hand. I write this part so that you can get a feel for what the computer is doing.

9.1.1 THE BIG QUESTION The big question is how we get from one line-of-best-fit to the next, from ℓ_i to ℓ_{i+1} . Unfortunately, there is no “best method” to minimize the L_1 norm when there are more points than dimensions. It is even worse: We were able to find closed-form solutions to the unique estimators for the L_2 norm (squaring). That cannot be done when using the L_1 norm (absolute values). There are multiple appropriate algorithms. The estimators may not be unique. Those are just a *few* problems working with the L_1 norm.

For those interested, here are some methods:

- Barrodale and Roberts (1974),
- Koenecker and Bassett (1978),
- Koenker and d’Orey (1987, 1994),
- Li and Arce (2004), and
- Shu-guang and Jian-wen (1992).

Note: These algorithms make use of different paradigms, different ways of seeing the problems. That is what makes studying statistics fun and interesting. Looking at a problem differently may be the key to its solution.

In \mathbb{R} , a function to perform median regression is `rq` in the package `quantreg`, which does not come with the default \mathbb{R} installation. Its use is very similar to what we are used to. While the `rq` function allows you to select different optimization methods, the default is the Barrodale and Roberts (1974) method.

From my experience the optimization algorithm matters little for real data. If the data are all integers, there may be issues with non-unique solutions or non-convergent algorithms. The cause in these cases is the non-uniqueness of the median.

problem

Example 1

Using median regression, what is the relationship between the violent crime rates in 2000 and 1990 in the `crime` data?

Solution: The following code estimates the median regression line for the relationship between the violent crime rates in 2000 and 1990 in the `crime` data:

```
library(quantreg)

dt = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(dt)

mod1 = rq(vcrime00 ~ vcrime90)
summary(mod1)
```

The following is the output:

```
Call: rq(formula = vcrime00 ~ vcrime90)

tau: [1] 0.50

Coefficients:
              coefficients lower bd  upper bd
(Intercept)  93.24525      72.83955 102.31731
vcrime90      0.57764      0.57518  0.62676
```

The output is the usual output. The value of `tau` is 0.50, because we are examining the regression line for the median, the 50th percentile.

The coefficients are the estimates for the intercept and slope. The lower and upper bounds are the 95% confidence interval for those parameters. There are no p-values, because the distribution of the estimators does not follow a nice test distribution. However, because we have a confidence interval, we have even more information than what a simple p-value would give. We are 95% confident that the relationship between the violent crime rate in 1990 and 2000 is between 0.575 and 0.627. Since this does not include the value 0, we can conclude that there is a significant relationship between the two variables. ♦

exercise

I leave it as an exercise for you to see that the OLS estimator for that effect is 0.581, with a 95% confidence interval from 0.518 to 0.643.

There is a difference between the two estimation methods. That difference is in how the method is affected by influential points like the District of Columbia. Median regression reduces the influence of DC, while ordinary least squares does not.

The absolute value function increases linearly as the residual increases. The squaring function increases quadratically as the residual increases. Thus, ordinary least squares will work harder to avoid making the DC residual too big. Median regression will not weight it as heavily.

Example 2

Here is another example of using median regression. What is the relationship between the property crime rates in 1990 and 2000?

Solution: The code is quite similar to that above:

```
|| mod3 = rq(pcrime00 ~ pcrime90)
|| summary(mod3)
```

The following is the partial output:

```
|| Coefficients:
|| coefficients lower bd upper bd
|| (Intercept) 730.46936 349.56585 1093.31979
|| pcrime90 0.60584 0.50893 0.77457
```

Again, the relationship is positive. A point estimate for that relationship is $\tilde{\beta}_1 = 0.606$, with a 95% confidence interval from 0.509 to 0.775. I again leave it as an exercise for you to show that the OLS estimator is 0.582 with a 95% confidence interval from 0.458 to 0.707. ♦

Note: When the data are “well behaved” without influential points, there tends to be little difference in the estimators. Figure 9.1 illustrates this.

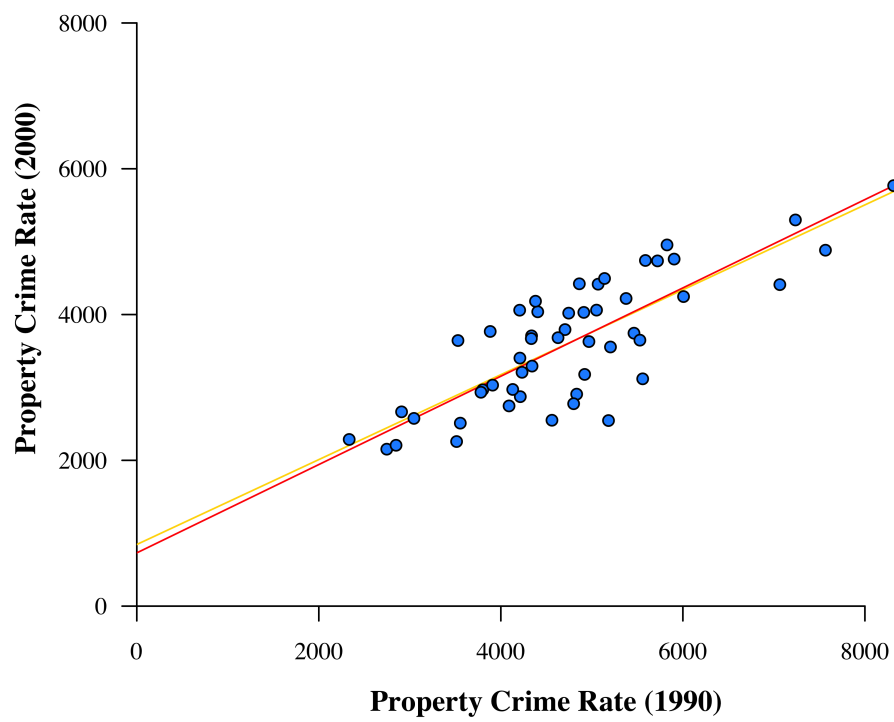


Figure 9.1: A graphic comparing the estimated lines from ordinary least squares (gold) and median (red) regression.

9.2: Quantile Regression

The previous section covered median regression. There, we motivated the method by focusing on minimizing the sum of the absolute value of the residuals. It turns out that this is equivalent to estimating the conditional median of the dependent variable (hence its name). In other words, the line of best fit is the line that best goes through the medians at each x -value.

conditional
median

Compare this to how we motivated ordinary least squares in Chapter 2: by minimizing the sum of the squared errors. This is equivalent to estimating the conditional mean of the dependent variable.

conditional
mean

In other words, OLS estimates $\mathbb{E}[Y | x]$ while median regression estimates $\text{Med}[Y | x]$, for want of better notation. (Perhaps $Q_2[Y | x]$ would be better notation?)

$P_{50}[Y | x]$?

There is absolutely no reason we need to focus *only* on the conditional median of the dependent variable (conditional on the independent variable). We may want to focus on other quantiles, like the 10th percentile. This happens a lot in sociology when studying poverty (10th percentile of income) or education (90th percentile of academic achievement).

The idea behind the fitting is the same (Koenker and Hallock 2001). The \mathbb{R} function is also the same. The only difference is that you need to specify the quantile. To see this, let us see a couple familiar examples.

Example 3

What is the relationship between the violent crime rates in 2000 and 1990 in the `crime` data at the 10th percentile?

Solution: Here is the code to perform this estimation:

```
|| mod5 = rq(vcrime00 ~ vcrime90, tau=0.10)
|| summary(mod5)
```

The following is the output:

```
Call: rq(formula = vcrime00 ~ vcrime90, tau = 0.1)

tau: [1] 0.1

Coefficients:
      coefficients lower bd  upper bd
(Intercept)  40.29964   -14.80397 100.38757
vcrime90      0.55616     0.38948   0.60422
```

Thus, for those states near the 10th percentile, the relationship between the 1990 and 2000 violent crime rate is between 0.389 and 0.604, with a point estimate of 0.556. This is only a little different from the median results, which suggests those states that are less crime-ridden (at the 10th percentile) still followed the same “rule” with respect to violent crime rate changes between 1990 and 2000. ♦

Example 4

What is the relationship between the property crime rates in 2000 and 1990 in the `crime` data at the 90th percentile?

Solution: Here is the code to perform this estimation:

```
mod6 = rq(pcrime00 ~ pcrime90, tau=0.90)
summary(mod6)
```

The following is the output:

```
Call: rq(formula = pcrime00 ~ pcrime90, tau = 0.9)

tau: [1] 0.9

Coefficients:
      coefficients lower bd  upper bd
(Intercept) 1761.72465   327.54503 2436.15997
pcrime90      0.53326     0.40262   0.84489
```

Thus, for those near the 90th percentile, the relationship between the 1990 and 2000 property crime rate is between 0.403 and 0.845, with a point estimate of 0.533. This differs a little from the median results (Example 9.1.1), which suggests those states

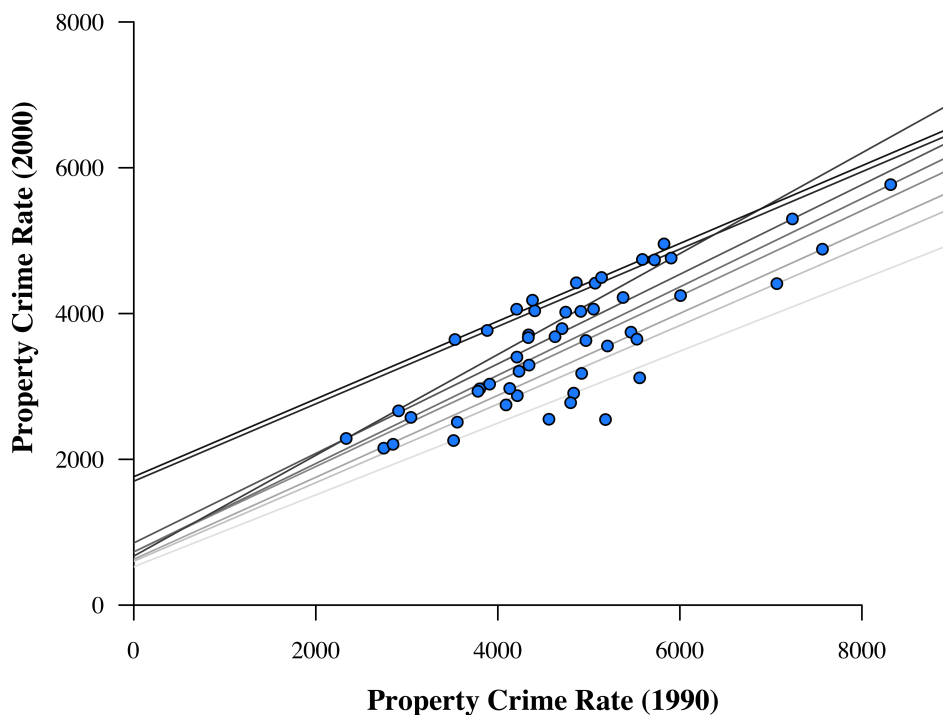


Figure 9.2: *Graphic illustrating the changing effect based on the quantile examined. The nine lines are regression lines for the deciles 0.10 through 0.90, with darker lines corresponding to higher quantiles.*

that are more (property) crime-ridden (at the 90th percentile) followed a similar “rule” with respect to violent crime rate changes between 1990 and 2000. Their rates dropped slightly more than did the typical (median) state. ◆

By the way, Figure 9.2 is a graphic of the deciles from 10 to 90% for the relationship between property crime rates in 1990 and 2000. Note that the effect does appear to change as one looks at middle-rate states. The highest levels, quantiles 80 and 90, are very similar in effect to the lower levels, quantiles 10 and 20. However, those states near quantile 50 seem to have a greater slope. If we had only looked at the median, we would have only reported these steeper effects. This may have overstated the effect.

Example 5

What is the relationship between the state's wealth in 1990 and the property crime rate in 2000? Show the effects at the first, second, and third quartiles.

proxy

Solution: We will use the GSP per capita as a proxy measure of wealth in the state. I leave the coding to you. Here is the appropriate output for the median:

```
Call: rq(formula = pcrime00 ~ gspcap90)

tau: [1] 0.5

Coefficients:
      coefficients lower bd  upper bd
(Intercept) 3061.50674  2383.13082 4989.55600
gspcap90      0.02403    -0.08109   0.04996
```

Interpreting the table indicates that there is no significant evidence that there is a relationship between the average wealth in 1990 and the property crime rate in 2000 for the median state (the confidence interval contains 0).

For the first and third quartile, the conclusions will be the same. As both confidence interval contain both positive and negative numbers, we are unsure of the relationship between these two variables. ♦

I leave it as an exercise for you to show that ordinary least squares indicates a statistical significant relationship (p -value = 0.0475). It also provides a point estimate of that relationship of $b_1 = 0.03025$.

Figure 9.3 provides the results for all deciles. Note that the slopes also seem to vary according to the quantile examined. Thus, the effect of wealth on property crime rates seem to be a function of those property crime rates. The lowest quantiles suggest the steepest effect. However, performing the analysis shows that the relationship is not statistically significant at the $\alpha = 0.05$ level. In other words, we were unable to detect a relationship.

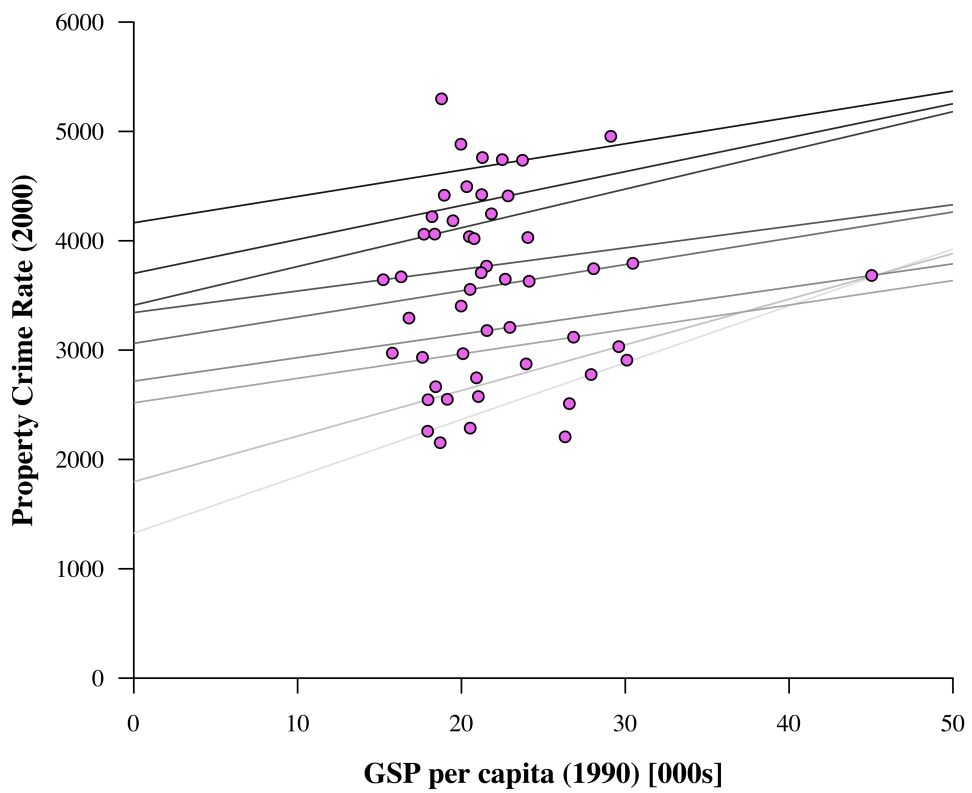


Figure 9.3: Graphic illustrating the changing effect based on the quantile examined. The nine lines are regression lines for the quantiles 0.10 through 0.90, with darker lines corresponding to higher quantiles.

9.2.1 THE ULTIMATE QUESTION So, is there a relationship between average wealth in 1990 and the property crime rate in 2000? One thing we know is that if there is a relationship, then it is minor.

It is not surprising that median regression does not detect a relationship while ordinary least squares does. Median regression, like all statistics based on the median (ranks), has a lower power than ordinary least squares (these statistics require Normality).

Note: So, the answer to the ultimate question is “I’m not sure.” This is unsatisfying. It is also reality. By using both OLS and median regression, we have a better understanding of the relationship between average wealth and property crime rates. That is the goal of statistics, not coming up with binary answers.

9.3: Conclusion

In this chapter, we covered quantile regression. We initially motivated the topic by modifying our definition of “best fit” to focus on the absolute value of the residuals in lieu of the square of the residuals. This led to an iterative process that allowed us to obtain estimates to any desired accuracy — at the cost of time and computing power.

This chapter then noted that median regression was just a specific instance of quantile regression, one in which the quantile was 0.500. This set the stage to introduce the results of quantile regression, in general. One may see quantile regression in research that focuses on better understanding the “wings” of the distribution instead of its middle.

Quantile regression uses **the entire data set**. It does not look at only the data corresponding to the q^{th} quantile. Such data may not actually exist. What states are at the 10th quantile of the property crime rates in 1990 and 2000? That’s not enough data to obtain any meaningful estimates.

Quantile regression estimates the q^{th} quantile of the response variable given the value of the independent variable.

9.4: End-of-Chapter Materials

9.4.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in quantile regression. These are listed here.

PACKAGES:

quantreg This package contains many functions associated with quantile regression. This chapter just skimmed the surface of what can be done and what should be checked. As this package is not a part of the base R, you will need to install it before loading it with `library(quantreg)`.

STATISTICS:

rq(formula) This is the function that performs quantile regression. The formula is required. By default, the quantile examined is $\tau = 0.50$, but that can be changed by specifying the value of that τ .

summary(model) This is the familiar command that allows us to see the regression table produced by the regression method. Here, it provides the effect estimates (coefficients) and the central 95% confidence interval for that effect.

9.4.2 EXERCISES

1. In the setting of Example 9.1.1, perform ordinary least squares regression to calculate the effect estimate and its 95% confidence interval.

9.4.3 APPLIED READINGS

- Lingxin Hao and Daniel Q. Naiman (2008). *Assessing Inequality*. Sage Publishing (Quantitative Applications in the Social Sciences; 166).

9.4.4 THEORY READINGS

- Ian Barrodale and F. D. K. Roberts (1974) “Solution of an overdetermined system of equations.” *Communications of the ACM* 17(6), 319–320. doi: 10.1145/355616.361024
- Lingxin Hao and Daniel Q. Naiman (2007). *Quantile Regression*. Sage Publishing (Quantitative Applications in the Social Sciences; 149).
- Roger W. Koenker (2005). *Quantile Regression*. Cambridge University Press.
- Roger W. Koenker and Gilbert W. Bassett, Jr. (1978). “Regression quantiles,” *Econometrica*, 46, 33–50.
- Roger W. Koenker and Kevin F. Hallock (2001). “Quantile Regression,” *The Journal of Economic Perspectives*, 15: 143–156.
- Roger W. Koenker and Vasco D’Orey (1987). “Algorithm 229: Computing regression quantiles.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36, 383–393.
- Roger W. Koenker and Vasco D’Orey (1994). “Remark AS R92: A Remark on Algorithm AS 229: Computing Dual Regression Quantiles and Regression Rank Scores.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43, 410–414.
- Yinbo Li and Gonzalo R. Arce (2004). “A Maximum Likelihood Approach to Least Absolute Deviation Regression.” *EURASIP Journal on Applied Signal Processing*, 12: 1762–1769.
- Yang Shu-guang and Liao Jian-wen (1992). “Solution of an Overdetermined System of Linear Equations in L_2, L_∞, L_p Norm Using L.S. Techniques.” *Journal of Computational Mathematics*, 10: 29–38.

CHAPTER 10:

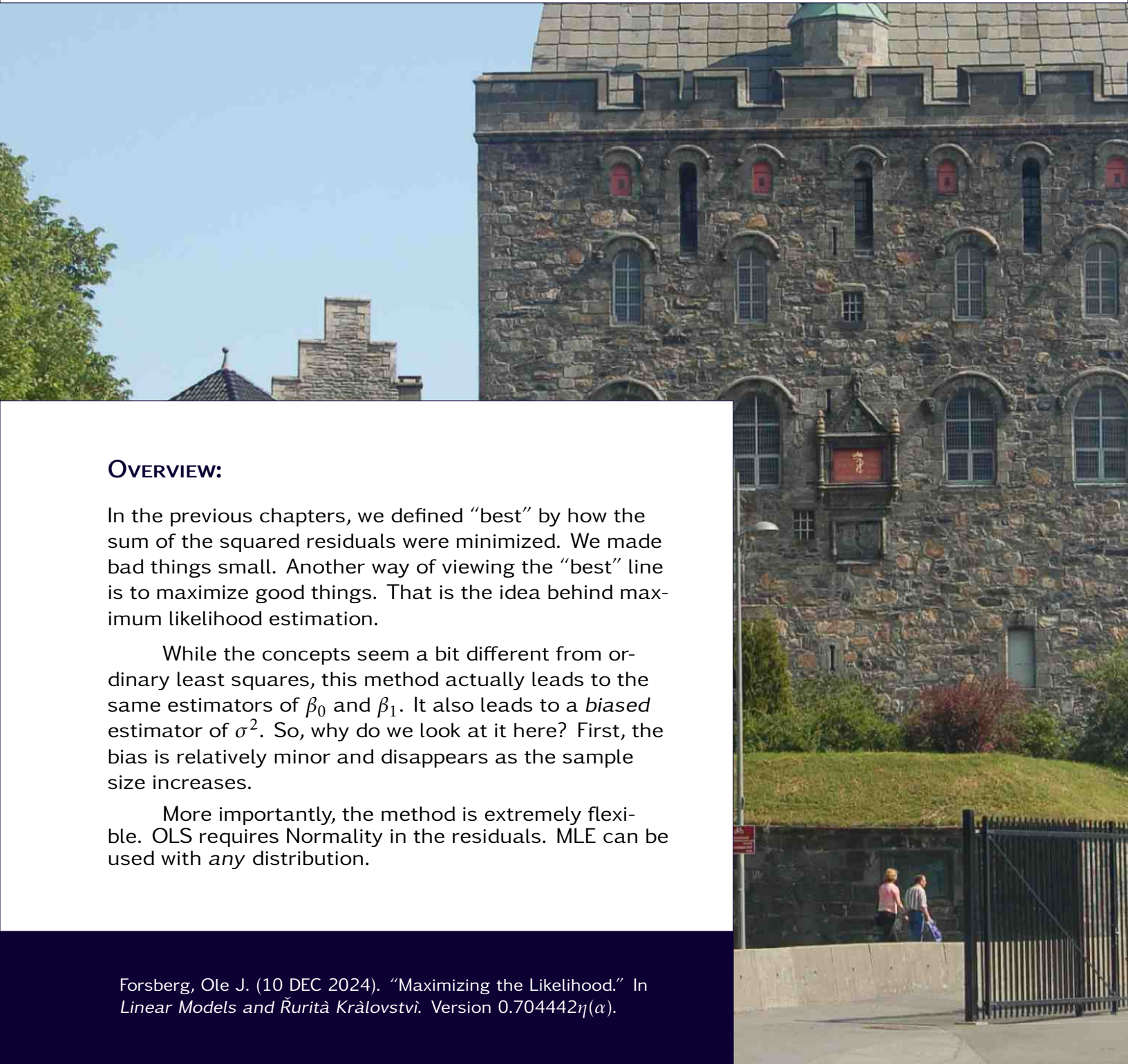
MAXIMIZING THE LIKELIHOOD

OVERVIEW:

In the previous chapters, we defined “best” by how the sum of the squared residuals were minimized. We made bad things small. Another way of viewing the “best” line is to maximize good things. That is the idea behind maximum likelihood estimation.

While the concepts seem a bit different from ordinary least squares, this method actually leads to the same estimators of β_0 and β_1 . It also leads to a *biased* estimator of σ^2 . So, why do we look at it here? First, the bias is relatively minor and disappears as the sample size increases.

More importantly, the method is extremely flexible. OLS requires Normality in the residuals. MLE can be used with *any* distribution.



Chapter Contents

10.1 The Likelihood	283
10.2 The MLE and the CLM	294
10.3 Conclusion	302
10.4 End-of-Chapter Materials	303



In the previous chapters, we have progressed from our desire to minimize some function of the residuals. This led to several related techniques:

- ordinary least squares
- weighted least squares
- generalized least squares
- ordinary least absolutes

All of these techniques sought to make the ‘bad’ things as small as possible, to produce a model that minimizes these residuals. However, our first definition of “best” from page 18 was based on making good things as large as possible, where that “good thing” is the likelihood of observing this particular data.¹

The theory is that the estimate most likely to have produced the observed data is the “best” estimate. Note that this differs from previous estimation methods in both the objective function and the size we desire. Bigger is better... bigger in terms of the “likelihood.”

The fundamental purpose of this chapter is to introduce you to the likelihood and the methods to maximize it. In trying to accomplish this, we will start with the simple and proceed to the less-simple.

¹Technically, it is the “likelihood of the data given our parameter estimates.”

10.1: The Likelihood

From a theoretical standpoint, the likelihood is just a generalization of probability. Where probability is bounded by both 0 and 1, the likelihood is only bounded below by 0. Values with higher probabilities are more likely to be observed. The same is true of likelihood: Values with higher likelihoods are more likely to be observed.

In the discrete case, the likelihood and the probability (mass) are the same. In the continuous case, the likelihood is the probability density. In other words, you really *have* come across the likelihood before. In your previous statistics course, the likelihood was called the “probability density” for continuous random variables and the “probability mass” for discrete random variables.

The difference between the likelihood and the probability mass or density is only one of emphasis. The probability mass (or density) is a function of observable values given the parameters of the distribution.

The likelihood is a function of the parameters, given the observed values (data). That difference is illustrated in the next two examples.

Example 1

Given that the success probability of a binomial random variable is $\pi = 0.25$, what is the probability of observing exactly one success out of two trials?

Solution: The probability mass function of the binomial distribution is

$$f(x; \pi, n) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad (10.1)$$

In this particular instance, the probability mass function is

$$f(x; \pi = 0.25, n = 2) = \binom{2}{x} 0.25^x (1 - 0.25)^{2-x} \quad (10.2)$$

And now calculating the probability gives

$$f(1; \pi = 0.25, n = 2) = \binom{2}{1} 0.25^1 (1 - 0.25)^{2-1} \quad (10.3)$$

$$= 2 \cdot 0.25^1 (0.75)^1 \quad (10.4)$$

$$= 2 (0.25) (0.75) \quad (10.5)$$

$$= 0.375 \quad (10.6)$$

Thus, the probability I observe exactly one success in two trials, given the success probability is 0.25 is just 0.375, which is a probability of 3 in 8. ♦

Example 2

Given that I observed exactly one success in two trials, what is the likelihood that the success probability is $\pi = 0.25$?

Notice that this question is very similar to the previous. The difference is subtle. The previous question asked about the probability of an observation. This one asks about the *likelihood of the parameter*.

Solution: The likelihood for a binomial random variable is

$$f(\pi; x, n) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad (10.7)$$

In this particular instance, the likelihood is

$$f(\pi; x = 1, n = 2) = \binom{2}{1} \pi^1 (1 - \pi)^{2-1} \quad (10.8)$$

Thus, the value of the likelihood for $\pi = 0.25$ is

$$f(0.25; x = 1, n = 2) = \binom{2}{1} 0.25^1 (1 - 0.25)^{2-1} \quad (10.9)$$

$$= 2 (0.25)^1 (0.75)^1 \quad (10.10)$$

$$= 2 (0.25) (0.75) \quad (10.11)$$

$$= 0.375 \quad (10.12)$$

Thus, we have calculated the likelihood that $\pi = 0.25$ is 0.375. Is this a lot? It actually depends heavily on the number of data points. In general, the larger your

sample size, the smaller the likelihood (of observing that particular data). Thus, the likelihood can only *meaningfully* be interpreted when in relation to other likelihoods based on the same data. ♦

same data

This last part deserves to be repeated.

Why? It is because we will come across statistics like the AIC and the BIC that can be used for model selection. However, since both are dependent on the likelihood, the values of the dependent variable need to be the same.

Warning: *The likelihood can only meaningfully be interpreted when in relation to other likelihoods based on the same data.*



The probability and the likelihood are *numerically* the same. The use and interpretation, however, are different. With probability, we are looking at a function of possible *outcomes*. With likelihood, we are looking at a function of possible values of the *parameters*. Thus, in the first case, we could ask questions about which value of x is most likely. In the second case, we would ask questions about which value of π is most likely.

Likelihood cries out to be maximized. Is $\pi = 0.25$ the maximum likelihood in the previous example? No. Calculate the likelihood of $\pi = 0.40$ to see that 0.25 is not the maximum (the value of π that produces the largest likelihood value). If you calculated $f(0.40; x = 1, n = 2) = 0.48$, then you did the calculations correctly.

Note that $f(0.40; x = 1, n = 2) > f(0.25; x = 1, n = 2)$. Thus, $\pi = 0.25$ is not the maximum likelihood estimate of π in this case. What is? Such optimization requires using calculus. From the above, you should be able to see that the objective function is

$$Q(\pi) = \binom{2}{1} \pi^1 (1 - \pi)^{2-1} = 2 \pi (1 - \pi) \quad (10.13)$$

This is a function of the parameter, since we are trying to determine the value of π that is *most* likely, given the data. The optimization proceeds as expected:

$$\frac{d}{d\pi} Q(\pi) = 2(1 - 2\pi) \quad (10.14)$$

$$0 \stackrel{\text{set}}{=} 2(1 - 2\hat{\pi}) \quad (10.15)$$

$$0 = 1 - 2\hat{\pi} \quad (10.16)$$

$$1 = 2\hat{\pi} \quad (10.17)$$

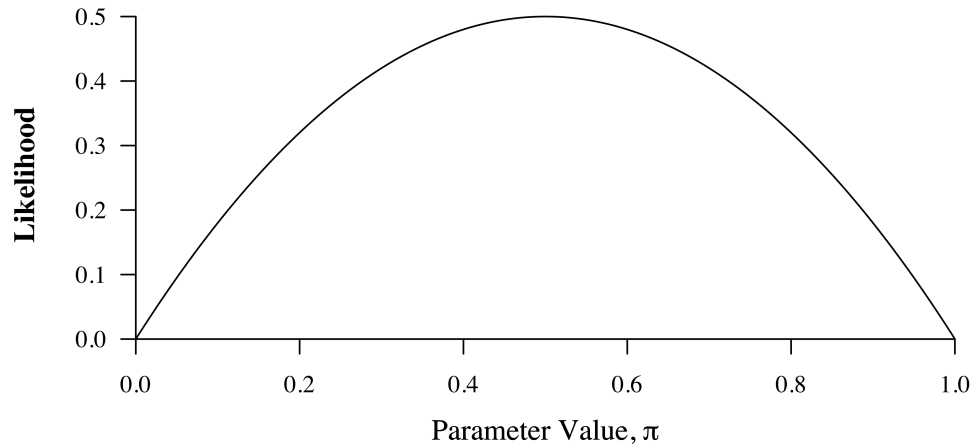


Figure 10.1: A graphic showing how the likelihood varies as the parameter changes. Unsurprisingly, the maximum likelihood occurs at $\hat{\pi} = 0.500$.

$$\frac{1}{2} = \hat{\pi} \tag{10.18}$$

Thus, given that we observed 1 success in 2 trials, the maximum likelihood estimator of π is $\hat{\pi} = 0.500$. For some reason, I am not surprised at this outcome. Are you?

Figure 10.1 shows how the value of the likelihood changes as the parameter value changes. It can be shown that the likelihood for this problem ranges from 0 to 0.5.

In general, one can show that the maximum likelihood estimator of π is $\hat{\pi} = x/n$, where x is the number of successes and n is the number of trials. I will leave that as an exercise.

exercise

To prove this, you would perform the same steps, but leave the x and n in the likelihood. If you do the calculations and the calculus correctly, you will end up with

$$\hat{\pi} = \frac{x}{n} \tag{10.19}$$

Example 3

Let Y be the number of Ruritarians walking through the door of the Valné Shromáždění, the general assembly building of Ruritania. The King would like to estimate the average number of people entering between noon and 1pm.

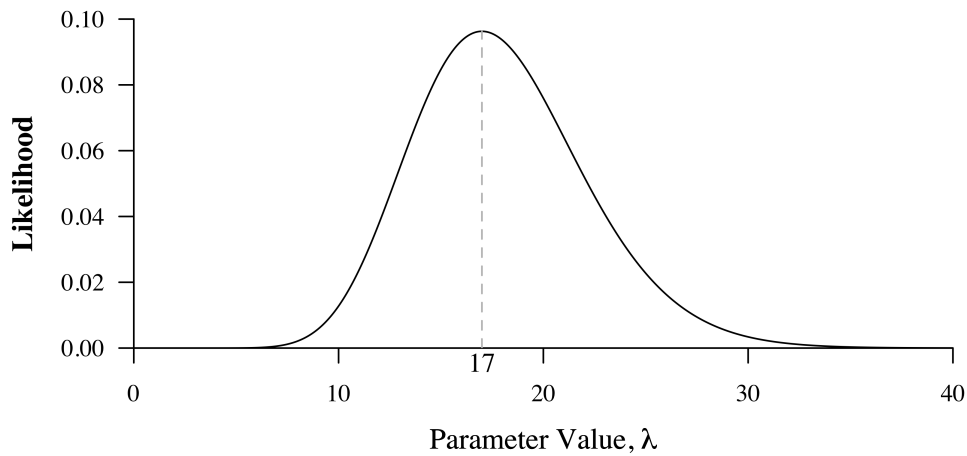


Figure 10.2: A graphic showing how the likelihood varies as the parameter changes.

To do this, the King had his Secretary of the Interior count the number of people entering the building during that hour on Monday.

On Monday, the Secretary counted $y = 17$. With this information, let us calculate the estimate of λ using maximum likelihood estimation.

A second distribution that you probably saw in your previous statistics class is the **Poisson distribution**. It has just one parameter, λ , the average rate. In this example, we will determine the maximum likelihood estimator of λ .

Solution: The likelihood for a discrete distribution, like the Poisson, is just the probability mass function:

$$\mathcal{L}(\lambda; y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (10.20)$$

That is the likelihood of each observation. Here, we only took one measurement (on Monday). Thus this is also the entire likelihood.

The next step is to maximize the likelihood with respect to the parameter, λ :

$$\frac{d}{d\lambda} \mathcal{L}(\lambda; y = 17) = \frac{d}{d\lambda} \left(\frac{e^{-\lambda} \lambda^{17}}{17!} \right) \quad (10.21)$$

$$= \frac{-e^{-\lambda} 17(\lambda^{17-1}) + e^{-\lambda} \lambda^{17}}{17!} \quad (10.22)$$

Now, set this equal to zero and solve for the estimator, $\hat{\lambda}$.

$$0 \stackrel{\text{set}}{=} -e^{-\hat{\lambda}} 17(\hat{\lambda}^{16}) + e^{-\hat{\lambda}} \hat{\lambda}^{17} \quad (10.23)$$

Since λ is constrained to the positive real numbers, we have the following simplification

$$0 = -17(\hat{\lambda}^{16}) + \hat{\lambda}^{17} \quad (10.24)$$

$$0 = -17 + \hat{\lambda} \quad (10.25)$$

Thus, the maximum likelihood estimate of λ is $\hat{\lambda} = 17$.



surprised?

And so, we report to His Majesty that our estimate of the average number of people passing through the doors of the Valné Shromáždění is 17 per hour.

By the way, a graphic of the likelihood function for varying values of λ is given in Figure 10.2. Note that the function achieves its maximum when $\lambda = 17$. Thus, the MLE for λ is $\hat{\lambda} = 17$.

Example 4

His Majesty liked the report, especially the font (he likes serifs). However, he asked an excellent question: “Bylo by lepší měřit více než jednou?”

To address his point, the Secretary of the Interior decided to take multiple measurements over several days. So, for the next week, he measured the number of people entering the Valné Shromáždění an hour at a time, randomly selecting the time of day each time. Here is that data: 15, 20, 23, 34, 23.

With that new data what is the maximum likelihood estimator of λ , given these $n = 5$ measurements?

Solution: From the previous example, we know that the likelihood of a single observation is

$$\mathcal{L}(\lambda; y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (10.26)$$

Thus, the likelihood of n independent observations is

$$\mathcal{L}(\lambda; y, n) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad (10.27)$$

How do we know this? Remember from your introductory statistics class the product rule for independent events.

Theorem 10.1.1

Let A and B be two independent events. The probability of both events happening is the product of the individual events. That is

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B] \quad (10.28)$$

Note: This theorem can easily be extended to any finite number of events. The requirement is that the events are independent. The result is that the probability of all occurring is the product of the probability of each occurring.

Since there is a product involved, it will be easier to maximize the logarithm of the likelihood,

$$l(\lambda; y, n) = \sum_{i=1}^n (-\lambda + y_i \log \lambda - \log y_i!) \quad (10.29)$$

And so, we maximize this function with respect to λ to obtain our estimator:

$$\frac{d}{d\lambda} l(\lambda; y, n) = \frac{d}{d\lambda} \sum_{i=1}^n (-\lambda + y_i \log \lambda - \log y_i!) \quad (10.30)$$

$$= \sum_{i=1}^n -1 + \sum_{i=1}^n \frac{y_i}{\lambda} \quad (10.31)$$

$$= -n + \frac{n\bar{y}}{\lambda} \quad (10.32)$$

Now, setting this equal to zero and solving for the estimator gives us

$$0 \stackrel{\text{set}}{=} -n + \frac{n\bar{y}}{\hat{\lambda}} \quad (10.33)$$

$$n = \frac{n\bar{y}}{\hat{\lambda}} \quad (10.34)$$

Thus, with multiple measurement, the maximum likelihood estimator of λ is

$$\hat{\lambda} = \bar{y} \quad (10.35)$$

Before moving on, think about the result to ensure that it makes sense. This is always an important step! ◆



Warning: *At the end of every result, you should think about its consequences. Make sure the results make sense. If they do not, then double-check your work or see the world in a more subtle light.*

Another important distribution is the exponential distribution. It is used to model the time until some event occurs. Actuaries may use it to model (estimate) the time until a person dies or gets into an automobile accident or gets sued or some other wonderful event.

It has a single parameter, λ , which is the rate.² This means that the average will be $1/\lambda$. Double-check that this actually makes sense.

The following examples deals with this distribution.

²If you are having *déjà vu* again, do not worry. There is an intimate connection between the Poisson and exponential distributions. If the time between arrivals follows an exponential distribution, then the number of arrivals follows a Poisson distribution.

Example 5

His Majesty has some additional work for us. He would like to estimate the average lifetime of Ruritarians.

Let us use maximum likelihood estimation to provide an estimator for λ , the average rate of a person dying (NOT the average time until death).

Solution: The probability density function for the exponential distribution, when parameterized on its rate, is

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad (10.36)$$

Thus, the likelihood function for a single observation is

$$\mathcal{L}(\lambda; x) = \lambda e^{-\lambda x} \quad (10.37)$$

And, the likelihood function for n independent observations is

$$\mathcal{L}(\lambda; x, n) = \prod_{i=1}^n \lambda e^{-\lambda x_i} \quad (10.38)$$

As this is a product, the log-likelihood will be easier to differentiate. It is

$$l(\lambda; x, n) = \sum_{i=1}^n (\log \lambda - \lambda x_i) \quad (10.39)$$

Now, we maximize it.

$$\frac{d}{d\lambda} l(\lambda; x, n) = \frac{d}{d\lambda} \sum_{i=1}^n (\log \lambda - \lambda x_i) \quad (10.40)$$

$$= \sum_{i=1}^n \frac{1}{\lambda} - \sum_{i=1}^n x_i \quad (10.41)$$

$$= \frac{n}{\lambda} - n\bar{x} \quad (10.42)$$

$$0 \stackrel{\text{set}}{=} \frac{n}{\hat{\lambda}} - n\bar{x} \quad (10.43)$$

$$0 = \frac{1}{\hat{\lambda}} - \bar{x} \quad (10.44)$$

$$\hat{\lambda} = \frac{1}{\bar{x}} \quad (10.45)$$

◆

From this, it can be shown that the maximum likelihood estimator of the *mean* of an exponential distribution is

$$\hat{\mu} = \bar{x} \quad (10.46)$$

All it takes is knowing that the expected value of an exponential distribution is $\mu = 1/\lambda$.



Since the original question dealt with the average age, we would want to calculate $\hat{\mu}$, not $\hat{\lambda}$. I leave it as an exercise to show that a maximum likelihood estimator of μ for the following parameterization of the exponential distribution

$$f(x; \mu) = \frac{1}{\mu} e^{-x/\mu} \quad (10.47)$$

is $\hat{\mu} = \bar{x}$.

Note: It should be noted that the maximum likelihood estimator is awesome in that functions “pass through.” In other words, it can be shown that

$$\widehat{f(x)}_{\text{MLE}} = f(\widehat{x}_{\text{MLE}}) \quad (10.48)$$

In words, the maximum likelihood estimator of a function of a parameter is that function of the maximum likelihood estimator of the parameter.

This is as good a time as any. There are two “drawbacks” to using maximum likelihood to estimate parameters. The first is that there is no guarantee that the estimator is unique. The second is that there is no guarantee that the estimator is unbiased.

While these seem bad, there is a nifty theorem that states the MLE is asymptotically unbiased; that is, as the sample size increases, its bias goes to zero.

Figure 10.3 shows the likelihood graph for λ of an Exponential distribution. From this graphic, you should be able to estimate the value of μ , the average age of death in Ruritania.

Note: In a future course, you may be dealing with maximum likelihood estimators frequently. Note that the graphic above tells a story beyond the estimate. It also gives insight into how precise the estimate is. The flatter a graphic around the estimate, the greater the uncertainty.

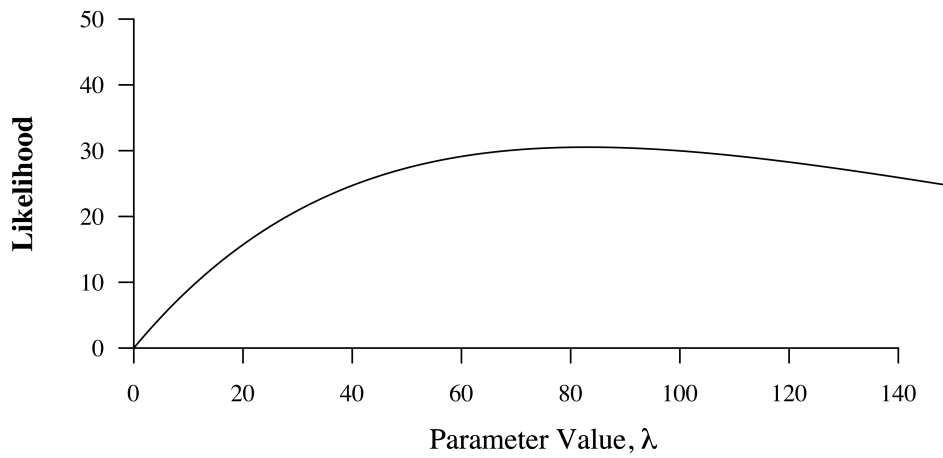


Figure 10.3: A graphic showing how the likelihood varies as the parameter changes.

To see this, note that you probably had more difficulty estimating the maximum in Figure 10.3 because of the curve's flatness. The likelihood in Figure 10.2 is more sharp, thus it is much easier to determine the maximum value. If you want to explore this, please check out **Fisher Information**, which is defined as

$$\mathcal{I}(\theta) = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f(x; \theta) \right)^2 \middle| \theta \right] \quad (10.49)$$

10.2: The MLE and the CLM

Recall that the classical linear model assumes

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \quad (10.50)$$

with $\mathbf{E} \sim \mathcal{N}_n(\mathbf{0}; \sigma^2 \mathbf{I})$. When we fit this model using ordinary least squares (OLS), we obtained the following estimators:

$$b_0 = \bar{y} - b_1 \bar{x} \quad (10.51)$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (10.52)$$

Let us see what we get when we fit this model using maximum likelihood methods.

Theorem 10.2.1

The maximum likelihood estimator of β_0 is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (10.53)$$

This is equivalent to the OLS estimator of the y-intercept.

Proof. The first step is to determine the likelihood function. The second step is to maximize that likelihood with respect to the parameter. As is usual, one maximizes the logarithm of the likelihood instead of the likelihood itself. It is generally easier.

Remember the conditional distribution of y . With that in mind, here is the likelihood for *one* observation:

$$\mathcal{L}(\mu, \sigma^2; x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y - \mu)^2}{\sigma^2}\right] \quad (10.54)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y - \hat{y})^2}{\sigma^2}\right] \quad (10.55)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y - (\beta_0 + \beta_1 x))^2}{\sigma^2}\right] \quad (10.56)$$

Jeeee-willikers! That is just the probability density function for the normal distribution, where μ (as always) represents an expected value.

That was for a single observation. However, we rarely deal with just one data point. We deal with n of them. We remember from our introductory statistics course that if the data are independent, then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. That means that the likelihood of observing all of our data is just the product of the individual likelihoods (also see Theorem 10.1.1).

IF

With that, we have

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2; \mathbf{x}, \mathbf{Y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2}\right] \quad (10.57)$$

Since you may not have seen the notation before, \prod is the product symbol just like \sum is the summation symbol.

The next step is to maximize this likelihood. From calculus, we recall the product formula for derivatives. Just try applying the product formula here. You will shortly go bald from pulling out your hair. There is no easy way to maximize this likelihood function directly. *¡Què lastima!*

However, if we apply an increasing bijection to this likelihood, then maximizing *that* function is equivalent to maximizing the original likelihood... equivalent in terms of the value that produces the maximum (the “argmax”).

**one-to-one
and onto**

Because the likelihood has a lot of products, and because it is easier to maximize a sum, we use the logarithm function. The log-likelihood function of the above function is just

$$l(\beta_0, \beta_1, \sigma^2; \mathbf{x}, \mathbf{Y}) = \sum_{i=1}^n \left(-\frac{1}{2} \log(-2\pi\sigma^2) - \frac{1}{2} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \right) \quad (10.58)$$

Taking the derivative of a summation is so much easier than taking the derivative of a product... so much easier!

And, now that we have a practically differentiable function, we use calculus to maximize it with respect to β_0 :

$$\begin{aligned} \frac{\partial}{\partial \beta_0} l(\beta_0, \beta_1, \sigma^2; \mathbf{x}, \mathbf{Y}) &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^n \left(-\frac{1}{2} \log(-2\pi\sigma^2) - \frac{1}{2} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \right) \\ &= \sum_{i=1}^n -\frac{1}{2} \frac{2(y_i - (\beta_0 + \beta_1 x_i))(-1)}{\sigma^2} \end{aligned} \quad (10.59)$$

$$= \sum_{i=1}^n \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma^2} \quad (10.60)$$

Now, set this to zero and solve for $\hat{\beta}_0$:

$$0 \stackrel{\text{set}}{=} \sum_{i=1}^n \frac{y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i}{\sigma^2} \quad (10.61)$$

$$0 = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i \quad (10.62)$$

$$\sum_{i=1}^n \hat{\beta}_0 = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_1 x_i \quad (10.63)$$

$$n\hat{\beta}_0 = n\bar{y} - n\hat{\beta}_1 \bar{x} \quad (10.64)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (10.65)$$

Thus, we have shown that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, as we desired. Note that this is also the OLS estimator of the y-intercept. Very interesting! \square

Theorem 10.2.2

The maximum likelihood estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (10.66)$$

Proof. From our proof of the estimator of β_0 , we have the following as our log-likelihood function:

$$l(\beta_0, \beta_1, \sigma^2; \mathbf{x}, \mathbf{Y}) = \sum_{i=1}^n \left(-\frac{1}{2} \log(-2\pi\sigma^2) - \frac{1}{2} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \right) \quad (10.67)$$

And so, the proof proceeds by taking the derivative with respect to β_1 and solving for $\hat{\beta}_1$.

$$\begin{aligned} \frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1, \sigma^2; \mathbf{x}, \mathbf{Y}) &= \frac{\partial}{\partial \beta_1} \sum_{i=1}^n \left(-\frac{1}{2} \log(-2\pi\sigma^2) - \frac{1}{2} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \right) \\ &= -\frac{1}{2} \sum_{i=1}^n \frac{2(y_i - (\beta_0 + \beta_1 x_i))(-x_i)}{\sigma^2} \end{aligned} \quad (10.68)$$

$$= \sum_{i=1}^n \frac{x_i y_i - \beta_0 x_i - \beta_1 x_i^2}{\sigma^2} \quad (10.69)$$

Setting this to zero and solving for the estimator, $\hat{\beta}_1$ gives

$$0 \stackrel{\text{set}}{=} \sum_{i=1}^n \frac{x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2}{\sigma^2} \quad (10.70)$$

$$= \sum_{i=1}^n x_i y_i - \sum_{i=1}^n \hat{\beta}_0 x_i - \sum_{i=1}^n \hat{\beta}_1 x_i^2 \quad (10.71)$$

$$= \sum_{i=1}^n x_i y_i - n \bar{x} \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i^2 \quad (10.72)$$

$$= \sum_{i=1}^n x_i y_i - n\bar{x}(\bar{y} - \hat{\beta}_1 \bar{x}) - \sum_{i=1}^n \hat{\beta}_1 x_i^2 \quad (10.73)$$

$$= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}_1 \bar{x}^2 - \sum_{i=1}^n \hat{\beta}_1 x_i^2 \quad (10.74)$$

Moving the $\hat{\beta}_1$ terms to the left side gives

$$\sum_{i=1}^n \hat{\beta}_1 x_i^2 - n\hat{\beta}_1 \bar{x}^2 = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \quad (10.75)$$

$$\hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \quad (10.76)$$

And finally,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad (10.77)$$

We have seen this before. It is the OLS estimator of the slope parameter. No surprise.

surprise!

To finish the proof, use algebra to show that the final equation above is equivalent to $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$. □

Theorem 10.2.3

The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (10.78)$$

Proof. It has been a while, so I will leave this as an exercise for you to prove this. I have already shown the log-likelihood function. All you have to do is differentiate with respect to σ^2 , solve for $\hat{\sigma}^2$, and use algebra to move things into the right form. \square

exercise

Note that the above formula (Eqn 10.78) is equivalent to

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2 \quad (10.79)$$

This *should* raise a red (maybe only light yellow or a nice fuschia) flag, as this is a biased estimator of σ^2 .

Why?

Note, however, that asymptotically (as $n \rightarrow \infty$), the estimator becomes unbiased. This can be proven — more easily than you may imagine.

10.2.1 CONSEQUENCES I leave it as an exercise to prove the following consequences:

1. $\hat{\beta}_0$ is unbiased for β_0 .
2. $\hat{\beta}_1$ is unbiased for β_1 .
3. $\hat{\sigma}^2$ is biased for σ^2 .

Because the maximum likelihood estimators are identical to the ordinary least square estimators, and because we have not altered the Normality assumption of the classical linear model, everything from Chapters 2, 2, and 4 hold.

Well, that is not entirely true. Remember that the MLE estimator of σ^2 is not the same as the OLS estimator. Thus, the test statistic and confidence interval will need to be altered a bit. However, the differences are minor for large samples.³

³And this is the problem that William Sealy Gosset had to deal with (see Section S.4.6). Things easily work for large samples. He had to deal with small samples in his work.

10.2.2 MULTIVARIATE DISTRIBUTIONS* There is one prerequisite to this textbook that would make things a little easier: an introduction to multivariate distribution. Thus far, I have “hand-waved” over the topic. Here, I will *briefly* discuss the topic. Feel free to treat this section only as a passing interest.

The following is a univariate distribution:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right] \quad (10.80)$$

This is the (in)famous probability density function for the normal distribution. It is a function of just one variable (value), x . This is what makes it univariate. The prefix *uni* stands for neither the University of Northern Iowa nor edible sea urchin gonads. It is a Latin combining form for “one.” Thus, “univariate” indicates “one variable.”

Ewww!

The following is one example of a *bivariate* distribution:

$$f(x, y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x}\right) \left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \quad (10.81)$$

This is a distribution where X and Y are distributed jointly normal, and where they have correlation ρ between them. There are a lot of symbols there because it is written in scalar form. Were we to write it in matrix form, we could generalize all of these “—variate” distributions into one form.

If the random vector \mathbf{Y} follows a multivariate normal distribution such that $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, then

$$f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right] \quad (10.82)$$

Here, n is the number of *variables* that are jointly normal. This means that each random variable follows a normal distribution, given the values of the others. The vector $\boldsymbol{\mu}$ is a column vector of expected values for each X_j . Finally, $\boldsymbol{\Sigma}$ is the correlation matrix between the n random variables. If the x values are independent, then $\boldsymbol{\Sigma} \in \mathcal{D}_n$ (diagonal). If the x values are independent and identically distributed, then $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$.

If $n = 1$, then the multivariate normal reduces to the univariate normal. If $n = 2$, then it reduces to the bivariate normal, where the off-diagonal entries in $\boldsymbol{\Sigma}$ are equal to $\rho\sigma_1\sigma_2$ and the diagonal entries are σ_1^2 and σ_2^2 .

That is, if $n = 2$, then

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (10.83)$$

Still, not much in this subsection is important — if the observations are independent. If the observations are independent, then the n -variate normal is just the product of the n univariate normals.

If the observations are *not* independent, however, then the joint distribution — the actual distribution we care about — is not so simple. In many cases, it has not been entirely formulated. For instance, what is the multivariate Binomial distribution? That is, what is the distribution of $\{y_1, y_2, y_3, \dots, y_n\}$, given correlation amongst those n measurements?

Even better: How could we measure such correlation?⁴

⁴In such cases, Dai, Ding, and Wahba (2013) may give you some insight into the difficulty of these questions — and their answers! This really makes you appreciate random sampling, where both independence and identical distribution hold.

10.3: Conclusion

From this chapter, we have discovered how to perform maximum likelihood estimation (MLE). The steps are as usual: formulate the objective function, use calculus to maximize it.

The maximum likelihood estimators for the two main parameters of the classical linear model are the same as the ordinary least squares estimators. Thus, they are both unbiased. The maximum likelihood estimator of the error variance, however, is biased. We know this because it does not equal the MSE, which is unbiased.

Thus, it appears as though maximum likelihood gives us nothing helpful. However, this is not true. First, there is a theorem (beyond the scope of this course) that proves *all* maximum likelihood estimators are asymptotically unbiased. In other words, if your estimator is a maximum likelihood estimator, you have nothing to prove with respect to asymptotic bias (Panchenko 2006, Thode et al. 2002). All other estimators (like OLS) require separate proofs. So, we gain there.

Second, ordinary least squares requires that the conditional distribution of the dependent variable is normal. Maximum likelihood does not have that as a requirement. This allows us to go beyond the classical linear model and the requirement of Normality. In fact, the next part of this class examines this feature of maximum likelihood estimation.

10.4: End-of-Chapter Materials

10.4.1 R FUNCTIONS This chapter had no R functions. It was all mathematics and concepts. Yay!!

10.4.2 EXERCISES

1. Prove that the maximum likelihood estimator of π is x/n in a Binomial experiment.
2. Prove Theorem 10.2.3.
3. Prove $\hat{\beta}_0$ is unbiased for β_0 .
4. Prove $\hat{\beta}_1$ is unbiased for β_1 .
5. Prove $\hat{\sigma}^2$ is biased for σ^2 and that the bias is exactly $\frac{n-1}{n}\sigma^2$.

10.4.3 THEORY READINGS

- Bin Dai, Shilin Ding, and Grace Wahba (2012). “Multivariate Bernoulli Distribution.” *Bernoulli* 19(4): 1465–1483. doi: 10.3150/12-BEJSP10
- Dmitry Panchenko (2006). “Lecture 3: Properties of MLE: consistency, asymptotic normality. Fisher information.” *Open Courseware/MIT*. URL = <https://ocw.mit.edu/courses/math.443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf>
- Aaron Thode, Michele Zanolin, Eran Naftali, Ian Ingram, Purnima Ratilal, and Nicholas C. Makris (2002). “Necessary conditions for a maximum likelihood estimate to become asymptotically unbiased and attain the Cramer–Rao lower bound. II. Range and depth localization of a sound source in an ocean waveguide.” *The Journal of the Acoustical Society of America*. 112(1890). doi: 10.1121/1.1496765

Part III

Beyond the Classical Model

11	Generalized Linear Models	307
12	Binary Dependent Variables	331
13	Binomial Dependent Variables	375
14	Count Dependent Variables	407
15	Nominal and Ordinal Dependent Variables	439



CHAPTER 11:

GENERALIZED LINEAR MODELS

OVERVIEW:

Until this point, we have been applying the classical linear model (CLM) to our problems of modeling a dependent variable. It is the model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, with Normal errors. While this model is quite prevalent in the literature, it does not always do a good job of approximating reality.

In this chapter, we introduce the generalized linear model (GLM) and start to show its versatility. We also repeat much of the previous chapter, but from a different perspective, one of paying attention to the data-generating process.

Chapter Contents

11.1	The CLM and the GLM	309
11.2	The Requirements for GLMs	310
11.3	Assumptions of GLMs	315
11.4	The Gaussian Distribution	316
11.5	Generalized Linear Models in \mathbb{R}	318
11.6	Conclusion	325
11.7	End-of-Chapter Materials	326



In our regression examples thus far, we have been dealing with continuous dependent variables. The classical linear model (CLM) requires this because the dependent variable needs to be conditionally distributed according to the Normal (a.k.a. Gaussian) distribution. Chapters 2, 3, and 4 discussed this in detail.

Chapter 7 examined how we can handle a couple of types of violations of these assumptions, focusing on the case where the dependent variable is bounded. When the dependent variable is bounded, it *cannot* be distributed Normal. (Why? What is the support of the Normal distribution?) As such, if your dependent variable *is* bounded, you will have to transform that variable into an unbounded analogue. Once this is done, one might be able to use the methods of the usual CLM paradigm.

We have, however, encountered some difficulties with this transformation method. In each of our examples from Chapter 7, the dependent variable was bounded — but was *never* equal to its bound. This was necessary. If the dependent variable ever is equal to its bound, then the transformation function you use will return an infinite value (either $-\infty$ or $+\infty$).

In this part of the book, we will extend the classical linear model (CLM) to be more general, and we will introduce a unifying framework allowing us to fit many different types of dependent variables — both continuous and discrete.

11.1: The CLM and the GLM

The Classical Linear Model (CLM) assumes that the relationship between the dependent and the independent variables is linear and that the response variable can take on all possible values; i.e., $Y \in \mathbb{R}$. Furthermore, to come to statistical conclusions, least squares methods assume that the errors are normally distributed.

However, not all relationships fit this model. Statisticians who realized this, modified the CLM to handle many different types of relationships, much in the same way we have (see, e.g., Chapters 5 through 7). Thus, if the dependent variable is continuous and bounded, we modify the dependent variable. If there is heteroskedasticity in the model, we pre- and post-multiply the variance-covariance matrix to better approximate the true standard errors.¹ If you need to weight the data based on some information (such as reliability), you multiply by the weight matrix. And so forth.

However, there are certain types of dependent variables that *cannot* be fit using this model (or fit optimally). These are the models with discrete dependent variables. If we want to hold on to the CLM paradigm, we will have to pretend such variables are continuous.² Often, this assumption is not a good one. When variables are binary, continuous approximations result in predictions that do not reflect reality. When variables are counts, the variances are functions of the expected value and are heteroskedastic. When the dependent variable is nominal, there is little we can do using the classical linear model.

The Classical Linear Model can *usually* be altered to create good predictions.³ However, the further your variable is from being continuous and unbounded, the more corrections you will have to make, and the more complex the process of estimation and prediction becomes — if even possible.

This chapter serves to bridge the gap between the classical linear model (CLM) and the generalized linear model (GLM). In this chapter, we will regenerate the results from the previous chapters, but use a different paradigm. This new paradigm will help us understand the assumptions underlying ordinary least squares regression. It will also serve as a basis for understanding the assumptions of this new modeling paradigm.

11.2: The Requirements for GLMs

¹These are called ‘sandwich estimators’ and were developed by Peter Huber (1967) and Halbert White (1980).

²This assumption may not be a bad one. If we are modeling house value, then the discrete variable is very close to its continuous approximation.

³While the predictions will frequently be fine, the confidence bounds will be based on assumptions not met by the data.

The *Generalized* Linear Model (GLM) is a paradigm that extends the CLM and many adjustments to it.⁴ To accomplish this feat, the model parts are named and examined. Those three parts are the linear predictor, the conditional distribution of the dependent variable, and the link function. While we have already mentioned all three of these concepts, let us explore them in greater detail before we derive the mathematical results.

11.2.1 THE LINEAR PREDICTOR Of the three knowledge requirements for using generalized linear models (GLMs), the linear predictor is the most familiar. It is merely the weighted sum of your chosen explanatory variables that you used throughout the classical linear model chapters:

$$\eta := \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (11.1)$$

$$= \mathbf{XB} \quad (11.2)$$

The only difference is that we are providing a name for the weighted sum (η , the Greek letter eta) and we are calling it the “linear predictor.” It is a “linear” predictor because the expression is linear in each of the coefficients (β_i). It is a predictor because it is used to predict the expected value of the dependent variable from the independent variables.

Note that the values produced by the linear predictor are unbounded. That is, note that $\eta \in \mathbb{R}$. This is very important to realize, especially when we get to the third requirement: the link function.

⁴There is a modeling paradigm termed *General Linear Models*, which merely allows for multiple independent variables to the CLM; technically, the CLM uses only one independent variable. General Linear Models are rarely discussed separately from the CLM, as such there is standardized no abbreviation for them. However, authors that do discuss General Linear Models frequently abbreviate them by GLM. These same authors will abbreviate Generalized Linear Models by GLZ. Upshot: When searching for information on GLMs, make sure you are reading about Generalized Linear Models and not General Linear Models.

Dependent variable is ...	Default Distribution	Canonical Link	Treated in Chapter
Continuous, unbounded	Gaussian	Identity	Chapter 11
Discrete, dichotomous	Bernoulli	Logit	Chapter 12
Discrete, bounded count	Binomial	Logit	Chapter 13
Discrete, <i>unbounded</i> count	Poisson	Log	Chapter 14
Discrete, very limited	Multinomial	Logit	Chapter 15

Table 11.1: A listing of several classes of dependent variables and appropriate distributions and links, and the chapter in which we discuss the variable class more closely.

11.2.2 THE CONDITIONAL DISTRIBUTION The first “new” addition is the conditional distribution of the dependent variable (its distribution, conditional on the values of the independent variable). Naming it is usually not as difficult as it may seem — a few rules of thumb are very helpful. The distribution chosen reflects your knowledge of the domain of the dependent variable. If the dependent variable can take on all Real values (as before), then an appropriate distribution is the Gaussian distribution (as before).⁵ If the dependent variable can take on only values of 0 and 1, then an appropriate distribution is the Bernoulli distribution. And so forth. Table 11.1 provides appropriate distributions for several different types of dependent variables (and the chapter in which we discuss them). This is not an exhaustive list, nor are the listed distributions always correct. They are just a good place to start.

dependent variable

Note: All of these distributions have something in common: They are members of the *exponential family of distributions* (or *exponential class* of distributions). Section 11.2.4 discusses why this family of distributions was selected and which distributions belong to it.

⁵The Gaussian distribution is the eponymous distribution named for Johann Carl Friedrich Gauss (1777–1855). We already know it as the normal distribution. That we are using the name Gaussian reflects standard terminology in GLMs and a desire to give credit where it is due. Well, in Francophone areas, the distribution is known as the Gauss-Laplace distribution to give appropriate credit to Pierre-Simon, Marquis de Laplace (1749–1827). However, Laplace also has his own distribution. Both the Gaussian and the Laplace distribution were created to describe errors in measurement.

expected value

The distribution is important in that its expected value automatically restricts the outcome to appropriate values of the dependent variable. Note that we are **explicitly modeling the expected value** of the dependent variable (given the values of the dependent variables). That we are modeling the expected value may sound odd, but we did this previously with the linear models: Our prediction line was a line of the expected value of the dependent variable, $\mathbb{E}[Y | x]$. The same is true for GLMs: The fitting routine predicts the expected value of the distribution, $\mathbb{E}[Y | x]$, not the observed value.

11.2.3 THE LINK FUNCTION The third aspect you need to know in order to use the GLM framework is the link function, which links the linear predictor and the expected value of the distribution. If we symbolize the expected value of the distribution as μ and the linear predictor as η , then the link function is $g(\cdot)$, such that $g(\mu) = \eta$.

mapping

The most important requirement for the link function is that it maps the bounded domain of the expected value of Y to the unbounded domain of the linear predictor η . An additional requirement is that it is a bijection; that is, the link and its inverse are both functions. It is also usual to make the link a strictly increasing function. This forces the direction of the effect of your variable to be in the same direction as the sign of the estimated coefficient: if the coefficient estimate is positive, then the variable has a positive effect on the dependent variable.

canonical link

Table 11.1 lists the canonical link functions for each of the provided distributions. One can use links that are not canonical — and often should — but the canonical link is the default link function used. In subsequent chapters, when an alternate link function is appropriate, we will discuss why.

11.2.4 THE MATHEMATICS* Nelder and Wedderburn (1972) formulated the GLM paradigm to unify modeling techniques for several different classes of problems, including logistic regression, count regression, and linear regression. Starting with a member of the exponential family of distributions, Nelder and Wedderburn created an estimation method called iteratively re-weighted least squares (IRLS). This method uses maximum likelihood estimation (MLE) to estimate the parameter effects using an iterative procedure. MLE remains the primary method of fitting GLMs, but other approaches are used, including maximum quasi-likelihood estimation, Bayesian estimation, and several variance stabilization methods.

Pong

Their choice of MLE was simply one of computing ease. Remember that the early 1970s were a time of loud polyester clothes, not of cheap computing power. However, even though MLE was chosen for ease, these estimates have some helpful properties. As such, this is still the most widely used method for fitting GLMs, just as OLS has been the preferred method for fitting CLMs for many decades.

EXPONENTIAL CLASS OF DISTRIBUTIONS: The one and only requirement on the distribution is that it belongs to the exponential class of distributions (Nelder and Wedderburn 1974; Wood 2006). Many of the distributions we experience belong to this class, so it is not an issue. Examples of distributions in this class are

- beta
- chi-squared
- exponential
- gamma
- geometric
- normal
- Poisson
- standard uniform

Specifically, to be a member of this family, the probability density function (or probability mass function, if discrete) must be expressible in the following form:

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \quad (11.3)$$

Let's look at a few features of this form to better understand what each of the parts indicates.

The Mean. The expected value of the distribution is just

mean

$$\mathbb{E}[Y] = b'(\theta) \quad (11.4)$$

Recall that the expected value is important, as it is what we actually model in GLMs.

variance

Variance. The variance is

$$\mathbb{V}[Y] = b''(\theta) \cdot a(\phi) \tag{11.5}$$

dispersion

The $a(\phi)$ is called the “dispersion parameter.” Infrequently, the chosen distribution forces this to be a specific value. Usually, however, this variable is free to reflect the data (be estimated from). For those distributions that force this to be a specific number, we either need to use quasi-likelihood to fit the model *or* we need to test this assumption.

Maka × Soul

Canonical Link. Next, the θ is the canonical link function, $g^{-1}(\cdot)$. It is a function of the parameters of the distribution selected. In the Gaussian case, the canonical link is the identity function, $\mu = \eta$. In the Bernoulli (and Binomial when n is known) case, the canonical link is the logit function, $\text{logit}(\mu) = \eta$, where the logit function is defined as

$$\text{logit}(\mu) := \log \left[\frac{\mu}{1 - \mu} \right] \tag{11.6}$$

Nuisance

Nuisance Parameters. Finally, $c(y, \phi)$ is a term that allows some flexibility to the exponential family of distributions. Without the c function, far fewer distributions would belong to this family. Further, note that the c function affects neither the expected value nor the variance.

11.3: Assumptions of GLMs

When we were creating ordinary least squares (OLS) regression, we made one assumption: $\varepsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. After learning the mathematics of fitting the models, we went back and figured out how to test these assumptions. The same will be true here.

When performing generalized linear modeling, you make at least three assumptions: you assume the linear predictor is correct; you assume the conditional distribution of the dependent variable is correct; and you assume the link function is correct. If these assumptions are not met by the data and model, then there is information in the data that you are ignoring.

Testing these is usually not as easy as in the case of OLS regression. The linear predictor and the link function, together, determine the functional form. It can sometimes be tested using a runs test. That is the easy part. Testing the correctness of the conditional distribution is much more involved. It requires that one understands the hypothesized distribution, especially in terms of range, expected values, and variances. Note that tests of heteroskedasticity may not be useful here; many distributions are heteroskedastic.

The testing must be done, however.

Note: As you read through this part of the book, always keep in mind what we are assuming. That will help you determine the requirements and how to test them.

assumptions

good news!

functional form

11.4: The Gaussian Distribution

normal distribution

To illustrate what we did in the previous sections, let us apply what we know to the Gaussian distribution, determining the canonical link, the expected value, and the variance. Hopefully, the results will not surprise us.

We start with the probability density function (pdf).

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \quad (11.7)$$

Now, to write this in standard form. This just takes algebra and some rules of logarithms.

$$= \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] \quad (11.8)$$

$$= \exp\left[-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] \quad (11.9)$$

$$= \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] \quad (11.10)$$

$$= \exp\left[\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{y^2}{2\sigma^2}\right] \quad (11.11)$$

standard form

Recall from Section 11.2.4 that the standard form is

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \quad (11.12)$$

Thus, we can see the correspondences. Thus, we have the following:

- $y = y$
- $\theta = \mu$
- $a(\phi) = \sigma^2$
- $b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2$
- $c(y, \phi) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{y^2}{2\sigma^2}$

Thus, the canonical link is $g(\mu) = \mu$, also known as the identity function. Note that the dispersion parameter is the variance, $a(\phi) = \sigma^2$. Also note that the expected

value is

$$\mathbb{E}[Y] = b'(\theta) \quad (11.13)$$

$$= \frac{d}{d\theta} \left(\frac{1}{2} \theta^2 \right) \quad (11.14)$$

$$= \theta \quad (11.15)$$

$$= \mu \quad (11.16)$$

Hopefully, this is as we expect. Finally, note that the variance is

$$\mathbb{V}[Y] = b''(\theta) a(\phi) \quad (11.17)$$

$$= \frac{d^2}{d\theta^2} \left(\frac{1}{2} \theta^2 \sigma^2 \right) \quad (11.18)$$

$$= \frac{d}{d\theta} (\theta \sigma^2) \quad (11.19)$$

$$= \sigma^2 \quad (11.20)$$

Also as we expect, hopefully.

OTHER LINK FUNCTIONS: While the canonical link is the identity function ($\eta = \mu$), it is *not* the only allowable link function. In Section 7.1.2, we transformed the continuous dependent variable because it was bounded below by (but never equaled) zero. In such a case, the logarithm is an appropriate link function: The dependent variable has a restricted range. The link function converts that range to an unbounded range. The same is true under the GLM framework. Similarly, the logit function is frequently an appropriate link function, as it was in Section 7.1.1.

not canon

With that, we start to see that for continuous dependent variables, what we did under the CLM paradigm we can do under the GLM paradigm. This is *always* true; the GLM paradigm extends the CLM paradigm to handle different classes of dependent variables.

11.5: Generalized Linear Models in R

In previous chapters, we performed linear modeling using the `lm` function. To perform *generalized* linear modeling, we use the `glm` function. When one uses the Gaussian distribution and its canonical link, results between the two methods will be *identical*. That is, we could have fit all of the `lms` with `glms` and not change a thing.

Note: If one uses the Gaussian distribution and a non-canonical link, the predictions will be very close, but not identical. The reason is that the transformation is performed on different quantities between the two methods... as the following shows.

To see this, we can look at two things. The first is focusing on what the alteration applies to (the residuals????). When transforming the dependent variable:

$$g^{-1}(\mathbf{Y}) = \mathbf{XB} + \mathbf{E} \quad (11.21)$$

$$\Rightarrow \quad \mathbf{Y} = g(\mathbf{XB} + \mathbf{E}) \quad (11.22)$$

When using the link function:

$$g^{-1}(\mathbb{E}[\mathbf{Y} | \mathbf{X}]) = \mathbf{XB} \quad (11.23)$$

$$\mathbb{E}[\mathbf{Y} | \mathbf{X}] = g(\mathbf{XB}) \quad (11.24)$$

$$\mathbf{Y} = g(\mathbf{XB}) + \mathbf{E} \quad (11.25)$$

$$(11.26)$$

So, the only difference is in whether the function applies to the residuals. In the CLM, it does; in the GLM, it does not. This is why there will be (usually slight) differences between the CLM transformed and the GLM with a link function.

Second, we can see this in an old example, use the GLM paradigm to find the answers, and see that the results are slightly different from when the model fit when transforming the dependent variable.

Example 1

Let us return to the `cows` data file. The voters of Děčín are being sent to the polls to vote on a constitutional referendum (ballot measure) that proposes to limit the number of cows that can be housed within the city limits. This was not the first time that Ruritarians were sent to the polls to vote on this or a closely related issue. Given the information from previous votes, what is the estimated proportion of voters who will vote in favor of the ballot measure in Děčín?

Solution: The example asks us to estimate the proportion of voters who will vote in favor of the ballot measure in Děčín. The dependent variable will be `propWin` and the independent variables will be `yearPassed`, `chickens`, and `religPct`. For now, let us assume a linear relationship between the independent variables and the dependent variable; that is, the equation we will use to fit the data is

$$\text{propWin} = \beta_0 + \beta_1(\text{yearPassed}) + \beta_2(\text{chickens}) + \beta_3(\text{religPct}) + \varepsilon \quad (11.27)$$

This is equivalent to

$$\mathbb{E}[\text{propWin}] = \beta_0 + \beta_1(\text{yearPassed}) + \beta_2(\text{chickens}) + \beta_3(\text{religPct}) \quad (11.28)$$

which is more clearly connected to the GLM paradigm than before.

Performing Generalized Linear Modeling in R is straight-forward (as it is in all modern statistical packages). The function to use is `glm` (for ‘Generalized Linear Modeling’):

```
|| glm(propWin ~ yearPassed + chickens + religPct)
```

As `glm` returns a lot of information, we should store its results in a variable, which I will call `mod1`. Once the computer computes the regression (and all associated information), we can summarize the results in the standard results table (Table 11.2) using the command

```
|| summary(mod1)
```

Notice that all three variables of interest are statistically significant at the $\alpha = 0.05$ level. Additionally, the model has a residual deviance of 0.063072 (as compared to the null deviance of 0.286802). This indicates that the model reduced the deviance by a factor of

$$1 - \frac{0.063072}{0.286802} = 0.7801 \quad (11.29)$$

pseudo- R^2

	Estimate	Std. Error	t-value	p-value
Constant Term	0.1512	0.0659	2.293	0.0295
Year Passed (post 2000)	-0.0201	0.0036	-5.618	≪ 0.0001
Banned Chickens	-0.0373	0.0200	-1.868	0.0723
Percent Religious	0.0095	0.0011	8.801	≪ 0.0001

Table 11.2: Results table for the regression of proportion support of a generic ballot limiting cows in Děčín against the three included variables. The residual deviance is 0.063072, on 28 degrees of freedom, and the AIC is -98.523. As the hypotheses were one-tailed hypotheses, all three explanatory variables are statistically significant at the standard level of significance ($\alpha = 0.05$).

And this agrees with the R^2 from Section 6.3.

Thus, the equation for the line of best fit (also known as the prediction line) is approximately

$$E[\text{propWin}] = 0.1512 - 0.0201(\text{yearPassed}) - 0.0373(\text{chickens}) + 0.0095(\text{religPct}) \quad (11.30)$$

According to this model, what is the expected vote in Děčín? To answer this, we need this information about the Děčín ballot measure: $\text{yearPassed} = 9$, $\text{chickens} = 0$, $\text{religPct} = 48$. With this information, and under the assumption that the model is correct, we have our prediction that 42% of the Děčín voters will vote in favor of this ballot measure. ♦

There is nothing in the previous paragraphs that differs from the analysis results from Section 6.3. This is because the Generalized Linear Model paradigm *extends* the Classical Linear Model paradigm and is equivalent to it when the dependent variable is Gaussian distributed and the link is the identity function. We can even use something like the goodness-of-fit measure we developed in Section 2.4, the R^2 measure. Here, however, we calculate it based on the null and residual deviances. The null deviance is the deviance inherent in the data (akin to the variance of the data, TSS). The residual deviance is the deviance in the data unexplained by the model (akin to the SSE).

If we wish to predict the results of a Venkovský ballot measure from 1994, which also restricted chickens, we would still get an impossible prediction — one that is outside logical limits. In Section 7.1.1, we corrected this error by transforming the data, modeling, then back-transforming the results. Instead of transforming the dependent variable, let us merely change the link function. Here is how that is done in R and with `glm`:

The command to use is

impossible

	Estimate	Std. Error	t-value	p-value
Constant Term	-1.8909	0.2898	-6.53	≪ 0.0001
Year Passed (post 2000)	-0.0886	0.0157	-5.63	≪ 0.0001
Banned Chickens	-0.2318	0.0878	-2.64	0.0134
Percent Religious	0.0475	0.0047	10.06	≪ 0.0001

Table 11.3: Results table of the results of ordinary least squares regression on the logit-transformed dependent variable. The residual deviance is 0.064987, the null deviance is 0.286802, the R^2 is 0.7734, and the AIC is -97.6 .

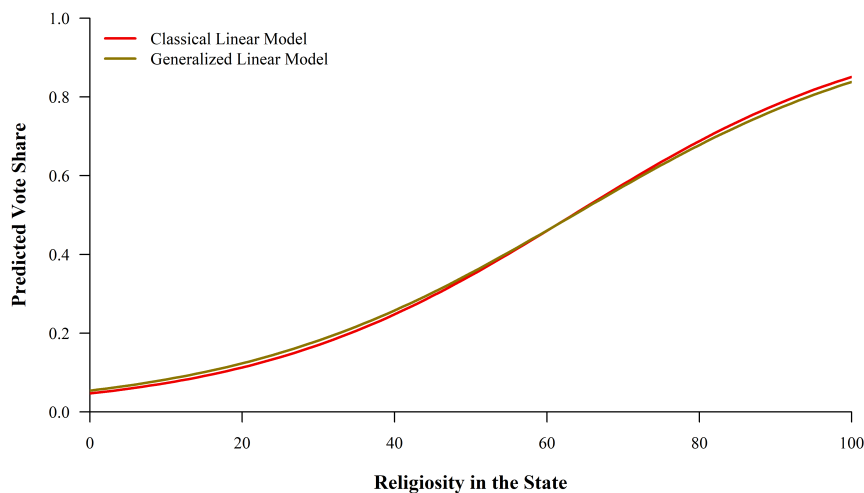


Figure 11.1: A plot of the predictions across various values of religiosity comparing the two models: CLM and GLM. Note that while the two results tables provided different results, the prediction plots are quite close together. The curves would have been equal only if we were to use the canonical link and the Gaussian distribution. For the predictions, the year was 2010 and the ballot measure also banned chickens.

```
||| mod3 = glm(propWin ~ yearPassed + chickens + religPct,
|||           family=gaussian(link=make.link("logit")))
```

Now, `summary(mod3)` provides many results. Note that all three independent variables are more statistically significant than in the non-transformed model. Also note that the effect directions are the same as before.

Finally, note that these parameter estimates are *not* the same as those where we used the Classical Linear Model with a logit transformation to fit the data in Chapter 7. However, if we make predictions, we see that the results are very close (Figure 11.1). CLMs and GLMs give identical results only with the Gaussian distribution and its canonical link. Here, we used the logit link.

identical

	Estimate	Std. Error	t-value	p-value
Constant Term	8.1595	0.1546	52.77	≪ 0.0001
Level of Democracy	-0.0452	0.0061	-7.44	≪ 0.0001
Honesty in Government	0.3335	0.0219	15.20	≪ 0.0001

Table 11.4: The results table from fitting the GDP data using Generalized Linear Models (cf. Table 7.2). Note that both independent variables are significant at the $\alpha = 0.05$ level here (highly significant).

Let us now re-examine Example 7.1.2 from Chapter 7. Recall that in that example, we were modeling a variable that was bounded below, but not above. This led us to transform the dependent variable using the logarithm function. Here, we fit the model with the Gaussian distribution and the non-canonical link.

Example 2

The gross domestic product (GDP) per capita is one of many measures of average wealth in countries. If extant theory is correct, then the wealth in the country is directly affected by the level of honesty in the government — countries with high levels of honesty (low levels of corruption) should be wealthier than those with low levels of honesty (high levels of corruption).

Furthermore, if theory is correct, the level of democracy in a country should *also* influence the country's level of wealth — countries with higher levels of democracy should be wealthier than countries with low levels of democracy.

Let us determine if reality (using the data in the `gdpcap` data file) supports the current theory or if current theory needs to explain the severe discrepancies.

Solution: The process of fitting this model with a GLM should be getting rote by now as it is so similar to fitting with a CLM. The R command is

```
|| m2 = glm(gdpcap ~ democracy + hig,
           family=gaussian(link = make.link("log")))
```

To see the results, we perform a `summary` call. The results of that call are provided in Table 11.4. Note that both independent variables are highly significant at the usual level of significance, $\alpha = 0.05$. Furthermore, the effect directions are the same as in the CLM model (Table 7.2 on page 201). ♦

Note: For some link functions, R allows you to skip the “`make.link`” portion. The `log` link is one of those for the Gaussian. Thus, the following command *would* also work:

```
|| glm(gdpcap ~ democracy + hig, family=gaussian(link="log"))
```

I recommend writing it out. That helps those who follow you to interpret what you are doing.

To predict the GDP per capita for Papua New Guinea, we repeat the same steps as when we were fitting CLMs: predict, then back-transform. Thus, a prediction statement will be

```
|| PNG = data.frame(hig=2.1, democracy=10)
|| exp(predict(m2, newdata=PNG))
```

The predicted GDP per capita for Papua New Guinea was \$2678 when fitted with the CLM. For this model, the prediction is \$4481. Thus, the prediction for Papua New Guinea is higher using GLMs than when using CLMs. Looking at the prediction graph (Figure 11.2), we see that GLM predictions are lower than CLM predictions for certain values of the dependent variable (and larger for others).

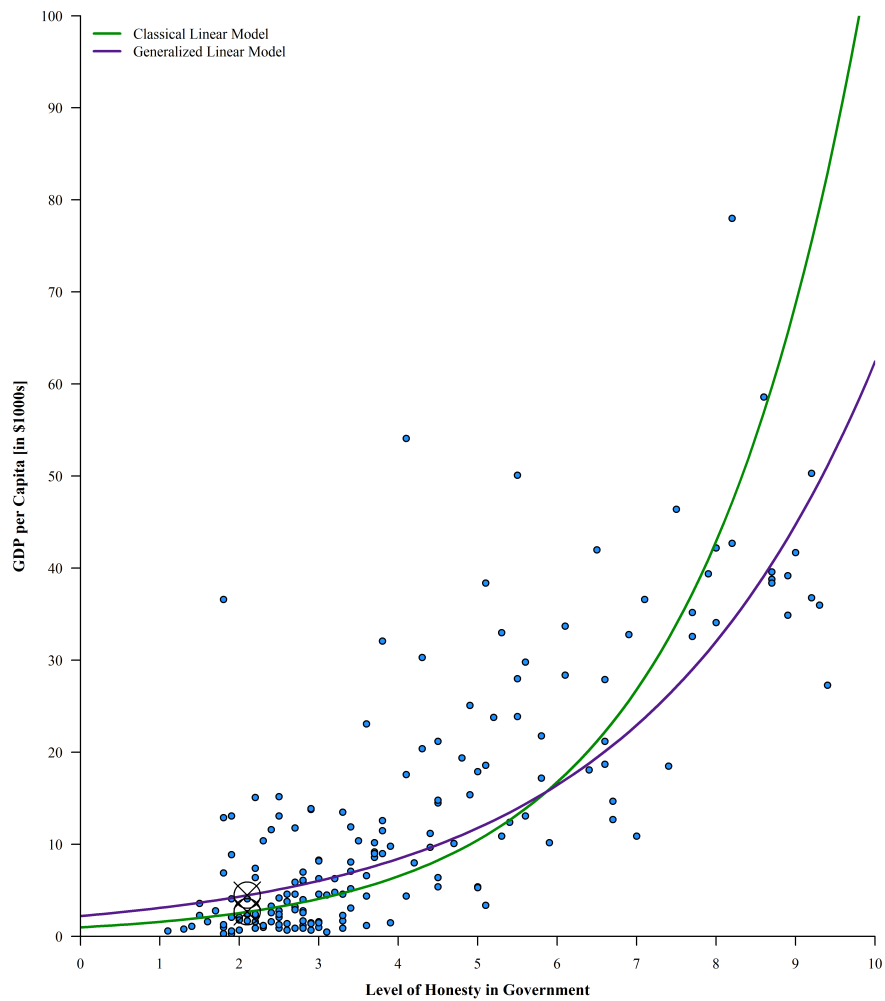


Figure 11.2: A plot of the two prediction curves, corresponding to the model fit using the Classical Linear Model and the Generalized Linear Model. Note that the two prediction curves are similar, but not really that close for large values of honesty in government. Estimates for Papua New Guinea are shown with the two \otimes symbols.

11.6: Conclusion

This chapter introduced the Generalized Linear Model paradigm, which is an extension of the Classical Linear Model paradigm from the previous two chapters. The advantage of the GLM paradigm is that more classes of dependent variables can be fit. The disadvantage (*if* we can call it that) is that we need to understand our data and model better. The three things we need to know are the linear predictor, the distribution of the dependent variable, and the function that links the expected value of the distribution with the linear predictor.

We tied this chapter to the previous chapters by showing that a GLM model using the Gaussian distribution (and the identity link) is *equivalent* to using the CLM. Three examples showed that the steps in modeling using the Generalized Linear Model paradigm are very similar to the steps used in modeling using the Classical Linear Model paradigm.

This chapter actually marked a major departure in how we see our data. Before, whenever a datum was different from our prediction, we viewed it as an error. Now, we realize that this variation is simply due to random fluctuations. We know this because we realize that our dependent variable is a random variable.

In the next chapters in this part of the book, we will examine more classes of dependent variables: binary, limited discrete (both nominal and ordinal), count, and non-negative continuous. As we examine these classes, pay attention to the selected distribution and the possible link functions. Table 11.1 provides several of the distributions and their canonical link functions.

Before you move on to the next chapter, ask yourself one thing: What requirements do we need to check for our models in this chapter? Make sure you can explain why they are requirements, too.

11.7: End-of-Chapter Materials

11.7.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

STATISTICS:

lm(formula) This function performs linear regression on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable.

glm(formula) This function performs generalized linear model estimation on the given formula. There are three additional parameters that can (and often should) be specified.

The `family` parameter specifies the distributional family of the dependent variable, options include `gaussian`, `binomial` (this chapter), `poisson` (next chapter), `quasibinomial`, `quasipoisson`, and `gamma`. If this parameter is not specified, R assumes `gaussian`.

The `link` parameter specifies the link function for the distribution. If none is specified, the canonical link is assumed.

Finally, the `data` parameter specifies the data from which the formula variables come. This is the same parameter as in the `lm` function.

predict(model, newdata) As with almost all statistical packages, R has a `predict` function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

11.7.2 EXERCISES This section offers suggestions on things you can practice from just the information in this chapter. As the purpose of this chapter was to introduce Generalized Linear Models and emphasize that everything we have done thus far can be done with GLMs, all of the extension questions are from previous chapters. For each of these, use the Generalized Linear Model paradigm (and the `glm` function).

SUMMARY:

1. What are the three aspects of your model that must be known before using generalized linear models?
2. When doing ordinary least squares regression, what were these three aspects?
3. How does the canonical link function differ from a link function?
4. What is $a(\phi)$ for the Gaussian distribution?

DATA:

5. Now, note that the value for Reka is 46% weekly church attendance. If, in the year 2012, the voters of Reka were faced with a ballot measure limiting the number of cows in the city limits, but not restricting chickens, what is the probability that it will pass?
6. Calculate a 95% confidence interval, with the *transformed* Cow Vote model, for predicting Děčín's vote. Is the actual outcome within the 95% confidence interval?
7. The logit transformation is not the only possible choice as a link for proportion data, there is also the asymmetric complementary loglog transformation (`cloglog` in the **RFS** package). Use this function as the link function to predict Děčín's vote, its 95% confidence interval, and the probability of the SSM ballot measure passing. The inverse of the complementary log-log transform has no name, but the R function is `cloglog.inv`, also in the **RFS** package.
8. Estimate the GDP per capita for Papua New Guinea. For this problem, use the *untransformed* model. Also, calculate a 95% confidence interval for this estimate. How close is this estimate to the real answer, and is the real answer within the predicted confidence interval?
9. Estimate the GDP per capita for Papua New Guinea. For this problem, use the *transformed* model. Also, calculate a 95% confidence interval for this estimate. How close is this estimate to the real answer, and is the real answer within the predicted confidence interval?
10. Compare and contrast the results of your Papua New Guinea estimates (Problems 8 and 9). Which model works best for Papua New Guinea? Which model works best overall?

11.7.3 APPLIED READINGS

- Denise Gammonley, Ning Jackie Zhang, Kathryn Frahm, and Seung Chun Paek. (2009) "Social Service Staffing in U.S. Nursing Homes." *Social Service Review* 83(4): 633–50.
- Katarina A. McDonnell and Neil J. Holbrook. (2004) "A Poisson Regression Model of Tropical Cyclogenesis for the Australian–Southwest Pacific Ocean Region." *Weather & Forecasting* 19(2): 440-55.
- Michael A. Neblo. (2009) "Meaning and Measurement: Reorienting the Race Politics Debate." *Political Research Quarterly* 62(3): 474–84.
- Weiren Wang and Felix Famoye. (1997) "Modeling Household Fertility Decisions with Generalized Poisson Regression." *Journal of Population Economics* 10(3): 273–83.

11.7.4 THEORY READINGS

- Hirotugu Akaike. (1974) “A New Look at Statistical Identification Model.” *IEEE Transactions on Automatic Control* 19(6): 716–23.
- Hirotugu Akaike. (1977) “On Entropy Maximization Principle.” In: P. R. Krishnaiah (Editor). *Applications of Statistics: Proceedings of the Symposium Held at Wright State University, Dayton, Ohio, 14-18 June 1976*. New York: North Holland Publishing, 27–41.
- George Casella and Roger L. Berger. (2002) *Statistical Inference*, Second edition. New York: Duxbury.
- Carl F. Gauss. (1809) “*Theoria motus corporum coelestium.*” *Werke*, 7, Göttingen: K. Gesellschaft Wissenschaft.
- Peter J. Huber. (1967) *The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233.
- Pierre-Simon Laplace. (1812) “*Théorie analytique des probabilités.*” Paris.
- Peter McCullagh and John A. Nelder. (1989) *Generalized Linear Models*. London: Chapman and Hall.
- John A. Nelder and Robert W. Wedderburn. (1972) “Generalized Linear Models.” *Journal of the Royal Statistical Society Series A (General)* 135(3): 370–84.
- Samuel S. Wilks. (1938) “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.” *The Annals of Mathematical Statistics* 9(1): 60–62.
- Simon N. Wood. (2006) *Generalized Additive Models: An introduction with R*. New York: Chapman & Hall.
- Halbert White. (1980) “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica* 48(4): 817–838.

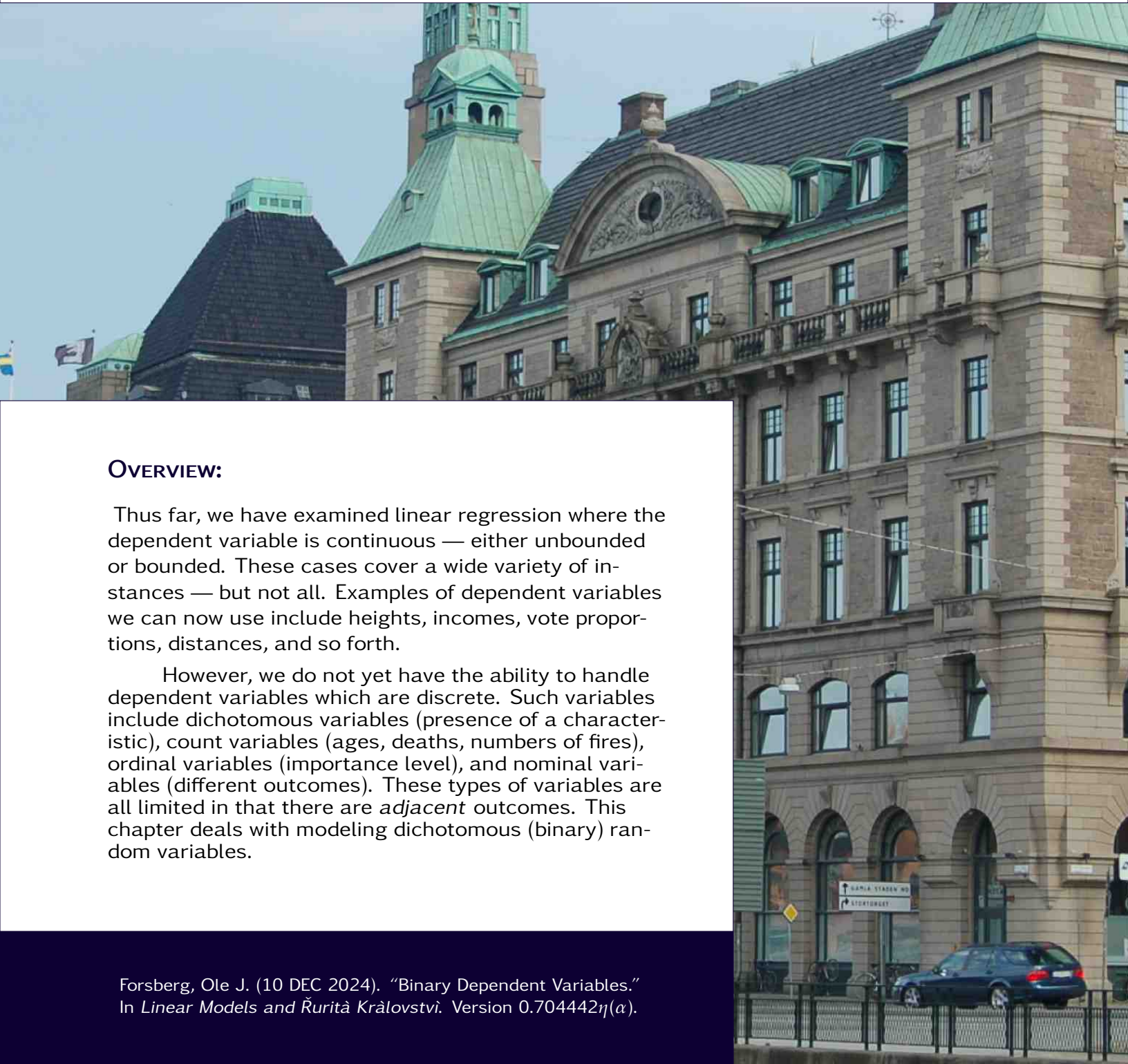
CHAPTER 12:

BINARY DEPENDENT VARIABLES

OVERVIEW:

Thus far, we have examined linear regression where the dependent variable is continuous — either unbounded or bounded. These cases cover a wide variety of instances — but not all. Examples of dependent variables we can now use include heights, incomes, vote proportions, distances, and so forth.

However, we do not yet have the ability to handle dependent variables which are discrete. Such variables include dichotomous variables (presence of a characteristic), count variables (ages, deaths, numbers of fires), ordinal variables (importance level), and nominal variables (different outcomes). These types of variables are all limited in that there are *adjacent* outcomes. This chapter deals with modeling dichotomous (binary) random variables.



Chapter Contents

12.1	Binary Dependent Variables.	333
12.2	Latent Variable Modeling.	336
12.3	The Mathematics.	339
12.4	Modeling with the Logit	347
12.5	Prediction Accuracy	351
12.6	Modeling with Other Links	359
12.7	Model Selection	362
12.8	Conclusion	368
12.9	End-of-Chapter Materials	369



The previous chapter introduced the generalized linear model paradigm (GLM). Modeling with GLMs requires that we specify three things:

1. the conditional distribution of the dependent variable (thus a formula for the expected value of the dependent variable), μ ;
2. the linear predictor, $\eta = \mathbf{XB}$; and
3. a (bijective) function linking the two, $g(\mu) = \eta$.

In that chapter, we showed that the Classical Linear Model is just a special case of the Generalized Linear Model. Specifically, the CLM is just a GLM using the Gaussian (Normal) distribution and the identity link. In this chapter, we cover the case of dichotomous (binary) dependent variables. In the following pages, we determine the appropriate distribution and the canonical link function.

12.1: Binary Dependent Variables

A dichotomous variable is one that can take one of two values: 1 or 0, True or False, Yes or No, success or failure. In research, these variables include the *occurrence* of terrorism, the *election* of a specific party to power, the *existence* of a fire, and the *failure* of a plane. In each of these cases, there are only two possible values, which we will refer to as success and failure. This is the hallmark of dichotomous variables. Before Nelder and Wedderburn (1972) created the GLM framework, statisticians created special models for binary dependent variable problems (with different transformations).

dichotomous

They did so because the classical linear model invariably makes predictions outside the logical range, demonstrates heteroskedasticity, and has residuals that are not Normally distributed — all violations of the OLS assumptions. To illustrate this, let us model the decision to purchase life insurance using age and income and the classical linear model (fit using OLS). The next example illustrates these issues.

Example 1

In Ruritania, the decision to buy life insurance is related to several variables, including age and income. Table 12.1 includes records of several individuals. Fit this data with a linear model using OLS:

$$\text{insurance} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{income} \quad (12.1)$$

Next, predict whether Václav will buy life insurance, given that his age is 65 and his income is \$125,000. Finally, determine if the assumptions of ordinary least squares are violated with this model and data.

Solution: Using our statistical program, we get the following as our linear regression equation

$$\text{insurance} = -0.4277 + 0.0130 \times \text{age} + 0.0088 \times \text{income} \quad (12.2)$$

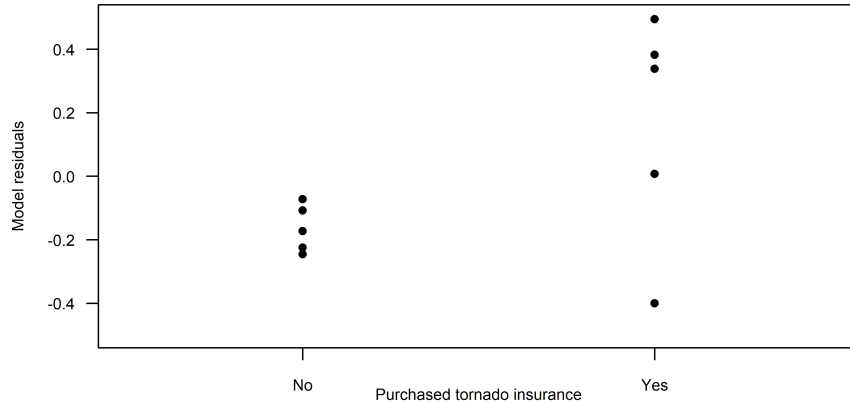


Figure 12.1: Scatter plot of the residuals against the values of the dependent variable. Note the different variances for the two groups. As such, the linear model is not appropriate in this case.

Using the provided information, we predict Václav will buy life insurance at (with?)

$$\text{insurance} = -0.4277 + 0.0130 \times \text{age} + 0.0088 \times \text{income} \quad (12.3)$$

$$= -0.4277 + 0.0130 \times 65 + 0.0088 \times 125 \quad (12.4)$$

$$= +1.5121 \quad (12.5)$$

What does this value of 1.5121 actually mean?

I don't know, either.

Individual	Insurance	Age	Income (\$000)
1	0	25	20
2	0	30	30
3	0	21	30
4	0	35	25
5	0	28	27
6	1	80	90
7	1	55	25
8	1	40	60
9	1	40	65
10	1	25	125

Table 12.1: Insurance pseudo data to accompany Example 12.1 in the text, in which we predict a person purchasing life insurance based on the person's age and income.

Next, to check the assumptions of OLS, let us merely check the assumption of homoskedasticity (constant variance). To do this, we plot the residuals against the values of the dependent variable. Figure 12.1 shows that the variation in the residuals significantly differs across the two groups in this model — a violation of our assumptions. In fact, calculations show that the variance for those who bought insurance is about 24 times higher than for those who did not (0.1325 vs. 0.0055). This is an example of non-constant variance. Performing the usual F-test for comparing two variances, we also see that this difference is statistically significant ($F = 0.0416, v_n = 4, v_d = 4, p = 0.0093$). Therefore, we conclude that our model is not appropriate for this data. ♦

There were two problems with this analysis. First, the model predicted an outcome that did not make sense. Second, the model violated at least one assumption of ordinary least squares (it actually violates all three). To solve the first problem, we *could* create a decision rule that any predicted value above the threshold $\tau = 0.500$ will be treated as a ‘Buy’ prediction, and any predicted value less than $\tau = 0.500$ will be treated as a ‘Not Buy’ prediction.

The second problem is more serious and not so easily solved, especially if we care about our estimate’s uncertainty (i.e., create confidence intervals). One may consider performing a transformation on the dependent variable to make it unbounded. A logit transformation would be a natural transformation for this; however, all of the dependent variables are either 1 or 0, which means the transformed values will be either $+\infty$ or $-\infty$. Furthermore, this transformation would not take care of the relationship between the residuals and the (transformed) dependent variables.

Note: There is a tendency to feel disappointed when our model violates assumptions, such as here. However, *instead* of seeing the existence of a relationship between the residuals and the dependent variable as a problem, let us realize such a relationship tells us that there is *more information* in the data than we are modeling at this point. As an interested researcher, we want to use that information to get more from our data. Thus, violations are not steps backwards; they are a path towards a deeper understanding of the data generating process.

12.2: Latent Variable Modeling

In Example 12.1, we discovered that Václav has a *something* of 1.5121 to buy life insurance. What is the *something*? Our gut really wants us to say that it is the probability that he buys life insurance. In fact, it would be *very* helpful if we could predict Václav’s probability of buying life insurance. Unfortunately, what we estimated cannot be a probability, as the value is greater than 1.

Notice, however, that we have just made an unconscious step in our minds: We are no longer thinking in terms of modeling the actual outcome (1 or 0); we are thinking in terms of modeling the expected value of the outcome, $\mathbb{E}[Y | x]$; here, that is the *probability* of a success, π .

In other words, we are now modeling a variable we cannot measure — a latent variable. Instead of modeling an actual outcome, we now think in terms of modeling the underlying probability that the person will purchase life insurance. This has the dual advantage of being a continuous variable and of being bounded by 0 and 1 — *exclusive*.

latent

As such, we can model it using previous techniques. Remember that the predicted value will be a *probability*, not an actual outcome we can measure. To predict the outcome, there is an additional step: selecting a threshold value, τ , above which we predict the individual bought insurance; below which, not. The traditional threshold value is $\tau = 0.500$; however, there is **no reason** we cannot alter it *to better fit the data* (Section 12.5.3).

Thus, our research model in the life insurance example becomes

$$\text{logit}(\mathbb{P}[\text{insurance}]) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{income} \quad (12.6)$$

We use the logit function for the same reason we used it before (Chapter 7): to transform the bounded variable into an unbounded variable. The right hand side of Equation 12.6 is η (eta), a linear function that can take on all real values — the linear predictor. Figure 12.2 shows a schematic of what we are actually modeling. The diagonal line in Figure 12.2, top, is the line of best fit for the linear predictor. The horizontal line is the threshold value we chose to distinguish between ‘Success’ predictions and ‘Failure’ predictions, which corresponds to $\text{logit}(\tau)$ in this top graph, τ in the bottom. The bottom figure is the linear predictor back-transformed into ‘probability’ units. The horizontal line is the actual τ chosen, here $\tau = 0.500$.

If we need to actually calculate the probability that Václav will purchase life insurance, we can calculate it from the linear predictor:

$$\text{logit}(\mathbb{P}[\text{insurance}]) = \eta \quad (12.7)$$

This is equivalent to

$$\mathbb{P}[\text{insurance}] = \text{logistic}(\eta) \quad (12.8)$$

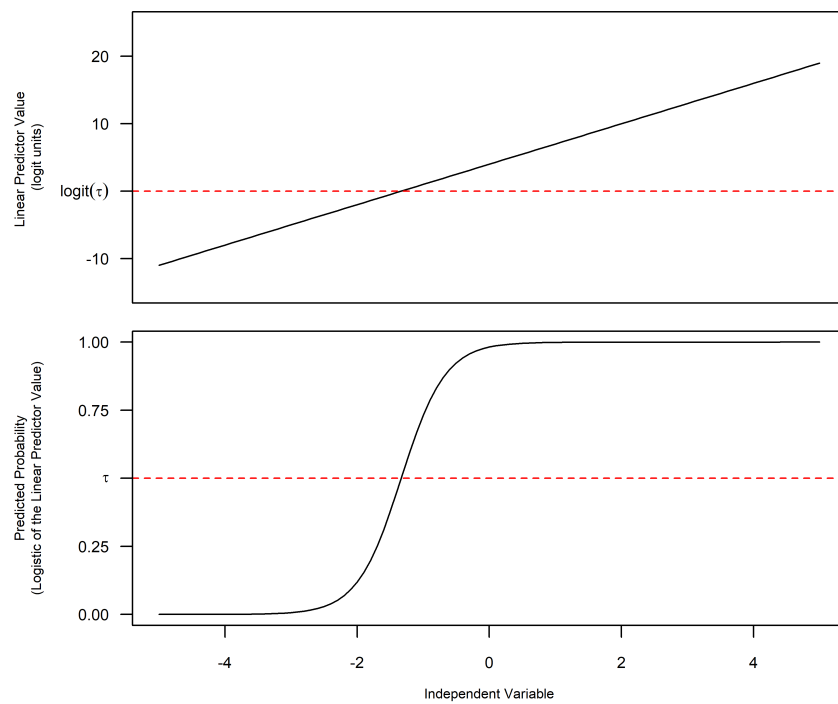


Figure 12.2: Plot of the linear predictor and a possible threshold for a typical latent binary dependent variable model. The logit of the Linear Predictor is in level units (proportion units).



This section examined the relationship between the line of best fit for the linear predictor, η , and the predicted probability of a success. However, we did not discuss *how* that line of best fit was determined. The next section does just that.

12.3: The Mathematics

When we model using the Classical Linear Model, we actually model/predict the *expected value* of the dependent variable, the mean. In the above insurance example, we modeled/predicted the probability of a person purchasing life insurance. What is the connection? It is that $\mathbb{E}[Y | x] = \pi$.

Remember from Chapter 11, performing GLM estimation requires that we know three things about our data and our model: the linear predictor, the conditional distribution of the dependent variable, and the function that links the two domains. The previous section discussed the linear predictor ($\eta = \beta_0 + \beta_1 \text{age} + \beta_2 \text{income}$) and a link function, $\text{logit}(\mu)$, for our example. That only leaves the conditional distribution of the dependent variable.

What are the possible values of the dependent variable? They are $\{0, 1\}$. What distribution has only these two outcomes? It is the Bernoulli distribution.¹ For the Bernoulli distribution, the probability of getting a ‘1’ (success) is π and the probability of getting a ‘0’ (failure) is $1 - \pi$. Mathematically, this means the full probability mass function (pmf) is

$$f(y) = \begin{cases} \pi^y (1 - \pi)^{1-y} & y \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases} \quad (12.9)$$

Strictly speaking, the probability mass function is not as important as the expected value of this distribution. Why? Remember that the Generalized Linear Model paradigm models the *expected value*, $\mathbb{E}[Y | x]$, of the distribution of the dependent variable.

Calculating the expected value of the Bernoulli distribution is easy using the definition of expected value:

$$\mathbb{E}[Y] := \sum_i y_i f(y_i) \quad (12.10)$$

$$= 0 f(0) + 1 f(1) \quad (12.11)$$

$$= 0 (1 - \pi) + 1 (\pi) \quad (12.12)$$

¹The Bernoulli distribution is a special case of the Binomial. It is equivalent to the Binomial distribution when $n = 1$; that is, if $Y \sim \text{Bern}(\pi)$, then $Y \sim \text{Bin}(1, \pi)$.

$$= \pi \tag{12.13}$$

Thus, the expected value of a Bernoulli random variable is π , the success probability.

This fact makes the results of modeling more apparent: As the GLM paradigm models the expected value, when we use the Bernoulli distribution, we end up modeling the probability of a success, which is what we want.

homoskedasticity

Note: Recall that one of the assumptions of Ordinary Least Squares is that the variance is constant with respect to the independent variable. When the outcomes are Bernoulli random variables, we can easily prove that the variance is *not* constant with respect to the expected values. If there is a relationship between the independent variable and the probabilities (which will be true if X affects Y), then a relationship between variance and expected value indicates heteroskedasticity.

To see this, let $Y \sim \text{Bern}(\pi)$. With this, and with the pmf above, we use the definition of variance to calculate $\mathbb{V}[Y]$:

$$\mathbb{V}[Y] := \sum_i (y_i - \mu)^2 f(y_i) \tag{12.14}$$

$$= (0 - \mu)^2 f(0) + (1 - \mu)^2 f(1) \tag{12.15}$$

$$= (0 - \pi)^2 f(0) + (1 - \pi)^2 f(1) \tag{12.16}$$

$$= \pi^2(1 - \pi) + (1 - \pi)^2 \pi \tag{12.17}$$

$$= \pi(1 - \pi) \left[\pi + (1 - \pi) \right] \tag{12.18}$$

This last line simplifies to $\mathbb{V}[Y] = \pi(1 - \pi)$, as $\pi + (1 - \pi) = 1$, which means $\mathbb{V}[Y]$ is a function of π , the expected value. It is not a constant with respect to the expected value, π . Binary dependent variables violate the assumption of homoskedasticity — **by definition**.

Note: The variance is a quadratic function of the success probability, $\mathbb{V}[Y] = \pi(1 - \pi)$. From this formula, we see that we are most unsure (the variance is highest) when the probability of a success is $\pi = 0.500$. Check that this makes sense: Which has a more uncertain outcome, a fair coin ($\pi = 0.500$) or a two-headed coin ($\pi = 1.000$)?

Now that we understand our choice of distribution a bit better, and the resulting expected value, let us examine the third facet: the link function. First, note that π is

bounded: $\pi \in (0, 1)$. Thus we need a function that takes a doubly-bounded variable and transforms it into an unbounded variable. We have already met a link function that can handle this — the logit function (see Chapter 7).²

And so, we have the three necessary components to use generalized linear models in this example:

- the linear predictor,

$$\eta = \beta_0 + \beta_1 \text{age} + \beta_2 \text{income} \quad (12.19)$$

- the distribution of the dependent variable,

$$\text{insurance} \sim \text{Bern}(\pi) \quad (12.20)$$

with the formula for the expected value $\mu = \pi$.

- and the link function,

$$\text{logit}(\mu) = \eta \quad (12.21)$$

²Here, I must mention that the logit is *not* the only appropriate link function. *Any* monotonic function that maps $(0, 1) \mapsto \mathbb{R}$ is appropriate. This includes the entire class of quantile functions, of which the probit is a member.

The choice of the link function often reduces to tradition within your field. However, social science theory is getting advanced enough to suggest link functions that are more appropriate than others.

Note: Here is what you need to take away from this section: The distribution must fit the possible outcomes. The link must translate the bounds on the parameter to the linear predictor. Both require you to know some distributions, which is why they are briefly covered in Appendix S.

12.3.1 DERIVING THE CANONICAL LINK* In Chapter 11, we mentioned that each distribution has a canonical link. Let us derive the canonical link for the Bernoulli distribution. As a side note, one does not have to understand this section to use Generalized Linear Models.

The steps to determine the canonical link are the same for the Binomial as it was for the Gaussian (Chapter 11):

1. Write the probability mass function (pmf).
2. Write the probability mass function in the required form.
3. Read off the canonical link.

For this distribution, this results in:

$$\text{pmf :} \quad \pi^y(1 - \pi)^{1-y} \quad (12.22)$$

$$= \exp\left[\log\left(\pi^y(1 - \pi)^{1-y}\right)\right] \quad (12.23)$$

$$= \exp\left[\log(\pi^y) + \log\left((1 - \pi)^{1-y}\right)\right] \quad (12.24)$$

$$= \exp\left[y \log(\pi) + (1 - y) \log(1 - \pi)\right] \quad (12.25)$$

$$= \exp\left[y \log(\pi) + \log(1 - \pi) - y \log(1 - \pi)\right] \quad (12.26)$$

$$= \exp\left[y(\log(\pi) - \log(1 - \pi)) + \log(1 - \pi)\right] \quad (12.27)$$

Link		Inverse Link	
Logit	$\log(\mu/(1-\mu))$	Logistic	$(1 + \exp(-\eta))^{-1}$
Probit	$\Phi^{-1}(\mu)$	Normal CDF	$\Phi(\eta)$
Cauchit	$\tan(\pi(\mu - \frac{1}{2}))$	Cauchy CDF	$\arctan(\eta)/\pi + \frac{1}{2}$
log-log	$-\log(-\log(\mu))$		$\exp(-\exp(-\eta))$
Complementary log-log	$\log(-\log(1-\mu))$		$1 - \exp(-\exp(\eta))$

Table 12.2: A list of several possible link functions (not all) to use for binary dependent variables. For the case of the Bernoulli distribution, remember that $\mu = p$.

$$= \exp\left[y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right] \quad (12.28)$$

$$= \exp\left[y \operatorname{logit}(\pi) + \log(1-\pi)\right] \quad (12.29)$$

$$= \exp\left[\frac{y \operatorname{logit}(\pi) + \log(1-\pi)}{1} + 0\right] \quad (12.30)$$

This is in the required form:

$$\exp\left[\frac{y \theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \quad (12.31)$$

Thus, reading off the standard form, we have the following:

- $y = y$
- $\theta = \operatorname{logit}(\pi)$
- $a(\phi) = 1$
- $b(\theta) = \log(1-\pi) = -\log(1 + e^\theta)$
- $c(y, \theta) = 0$

As such, the canonical link is the logit function, $g(\pi) = \operatorname{logit}(\pi)$.

probit

12.3.2 OTHER LINKS As mentioned in Chapter 11, we do not *have* to use the canonical link. Any monotonic, increasing function that maps the restricted domain to the unrestricted domain works. Thus, there are several options for the link function. Table 12.2 gives some options.³

The logit link is the canonical link. The probit link is frequently used in biostatistics. Its advantage is that it is based on the Normal distribution, with which we are intimately familiar. There is usually little difference between predictions made with the logit link and those made by the probit link. The coefficient estimates will usually differ by a factor of approximately 3.7, and the levels of significance will usually be close. The cauchit link is a symmetric link with heavy tails, as compared to the logit and the probit links (see Figure 12.3).

The log-log link and the complementary-log-log link are asymmetric links. The log-log link has a heavy right tail; the complementary-log-log link, a heavy left tail (see Figure 12.4). Most science theory is only now beginning to be able to state which of the three types of link functions will be most appropriate for the given model (symmetric, heavy left, heavy right).

Note: R has the built-in ability to model using the following link functions for the binomial (Bernoulli) distribution: logit, probit, cauchit, log, and complementary-log-log. The `RFS` package adds the log-log link function:

```
|| glm( y ~ x, family=binomial(link=make.link("loglog")) )
```



Again, the link choice is usually a matter of tradition, rarely of theory. Statistical significance of the variables should be similar across the several link functions. So, from a theory-testing standpoint, the link functions are rather interchangeable. With that said, *predictions* will vary depending on the link function chosen. Thus, if prediction is important then you will want to investigate the effect of different link functions on your predictions (and confidence bounds).

³Note that Table 12.2 is not an exhaustive list. Because we need an increasing function mapping (0,1) to the real numbers, any quantile function (inverse CDF) will work — any. However, the typical link functions for this type of problem are the logit and the probit.

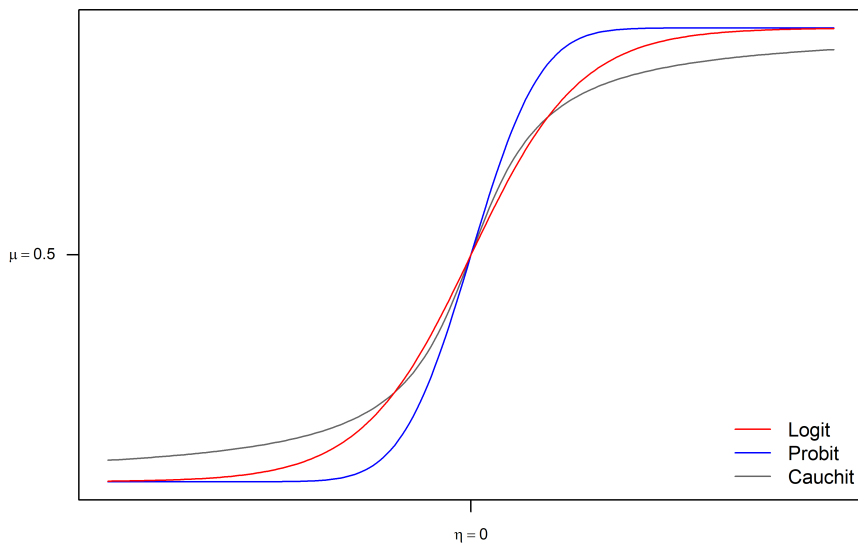


Figure 12.3: A graph of three symmetric links (logit, probit, and cauchit). Note that they all cross when the linear predictor $\eta = 0$ and that they cross at $\mu = 0.5$.

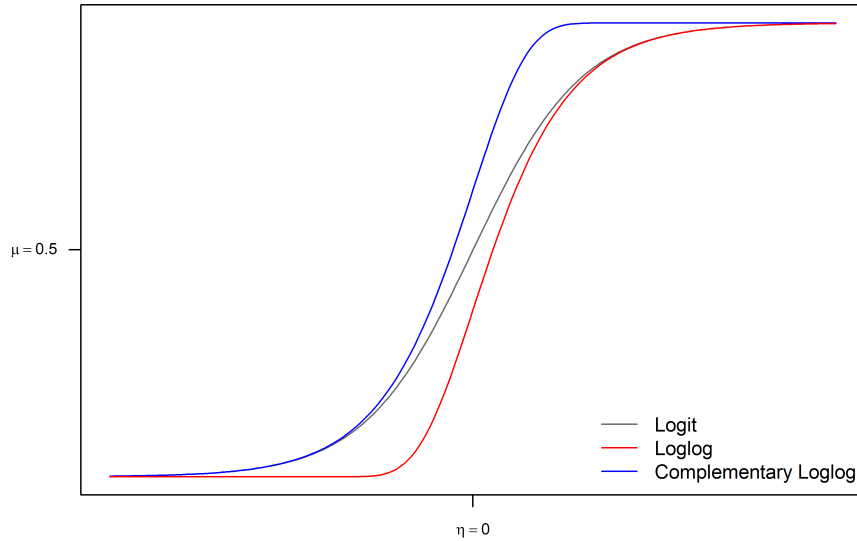


Figure 12.4: A graph of two asymmetric links (complementary-log-log and log-log functions) with the symmetric logit link for comparison.



Warning: While the actual predictions will differ, they should only do so slightly. The rule is that all models that are “appropriate” should provide similar conclusions and predictions. If they do not, then your model is too fragile... a bad thing. Build for model robustness.

12.4: Modeling with the Logit

For a binary response variable, the canonical link function is the logit link (Section 12.3.1). This link is characterized by being symmetric and having relatively thin tails (see Figure 12.3). This symmetry may be important when you are dealing with events that are balanced — neither rare nor frequent. The tail thickness may be important when you think there is a sharp transition between success and failure in your data. In reality, current social science theory is rarely so clear as to give you guidance in which link function you should use. As such, try several and see which one gives the best fit.⁴

Of course, if there is a traditional link function used in your field, you should use it as the default. Thus, social scientists should start with the logit, while the health science researchers should start with the probit.

Example 2

Since binary dependent variable regression is so very important to understand, let us look at it from a different direction:

Let us imagine an experiment where we have a series of 100 coins. Were these coins all fair, then the probability of getting a Head on any throw would be $\pi = \frac{1}{2}$. However, let us assume these coins are not necessarily fair, and that they are weighted in a very specific manner: Coin i has a probability of flipping a Head of π_i , which increases as i increases. That is,

$$\pi_{i+1} > \pi_i, \forall i \quad (12.32)$$

Now, if we were allowed to flip each coin *only once*, how can we estimate π_1 from the data?

⁴There is another reason to try several link functions. Since the “population” link function is not known, the predictions of the model should be robust to the choice of link function: Test several and see if the predictions are stable. If not, then the quality of your model depends heavily on something you cannot measure.

Solution: As we have no evidence to the contrary, let us use the canonical link function, the logit. Our steps are quite similar to the steps we performed when we had to transform the dependent variable:

1. Read in the data
2. Model the dependent variable using the GLM paradigm (specify the distribution, the linear estimator, and the link function)
3. Predict outcomes using your model
4. Back-transform the predictions using the inverse of your chosen link function

Note: There is a step missing from when we previously transformed our dependent variable: We do not have to transform the dependent variable. Generalized linear modeling does that for us in R. We do, however, have to back-transform the predictions. Be aware of this!

In R, the general form of the command is, showing the most important parameters,

```
glm(formula, family(link), data)
```

Only `formula` is required. If `family` is missing, the Gaussian (or Normal) distribution will be assumed. If `link` is missing, the canonical link for that family will be assumed. If `data` is missing, the current dataframe will be assumed.

For binary response variables, the family will need to be the Binomial distribution.⁵ Thus, for the example using the `coin` data file, the command will be

```
|| m1 = glm(head ~ trial,  
||       family=binomial(link="logit"),  
||       data=coin)
```

⁵Remember that the Bernoulli distribution is a special case of the Binomial. $Bern(\pi) = Bin(n = 1; \pi)$.

	Estimate	Std. Error	z-value	p-value
Constant Term	-2.2929	0.5384	-4.26	$\ll 0.0001$
Trial Number	0.0345	0.0087	3.97	0.0001

Table 12.3: Results of performing logistic regression on the coin flip data, `coinflips`. Note that these coefficient estimates are in logit units. As such, any predictions done using them will have to be transformed into level units using the inverse of the link function (the logistic).

I used the `data` parameter, as I did not attach the data earlier. If you attached, then you do not need to include this parameter. I also included `link="logit"` even though this is the default setting for the Binomial family in order to remind myself of the link function I used in this analysis.

The results from this command are summarized in Table 12.3. Again, note that the parameter estimates (and predictions) will be in “logit units.” You will have to use the logistic function (the inverse of the logit) to get the predictions in units of probability.

Recall that the original question asked us to determine π_1 , the probability of getting a Head on the first coin. There are a couple ways of doing that. The best will depend on the numbers involved. Since we want π_1 , we know it is equal to the logistic of the intercept plus *one* times the coefficient:

$$\pi_x = \text{logistic}(-2.2929 + 0.0345 x) \quad (12.33)$$

$$\pi_1 = \text{logistic}(-2.2929 + 0.0345 (1)) \quad (12.34)$$

$$= 0.0946 \quad (12.35)$$

The other way is to use the `predict` function and take the logistic of that value. You will get the same answer (within rounding error). The function call used is

```
|| predict(m1, newdata=data.frame(trial=1))
```

This gives an answer of `-2.2584`. The logistic of `-2.2584` is our estimate of π_1 , which is $\pi_1 = 0.0946$.

If we so desire, we can also plot the probability curve on a graph of the outcomes (see Figure 12.5). With such a graph, we could estimate which coin is most fair. With the graph, we could also get a feel for how well the model represents the data. ◆

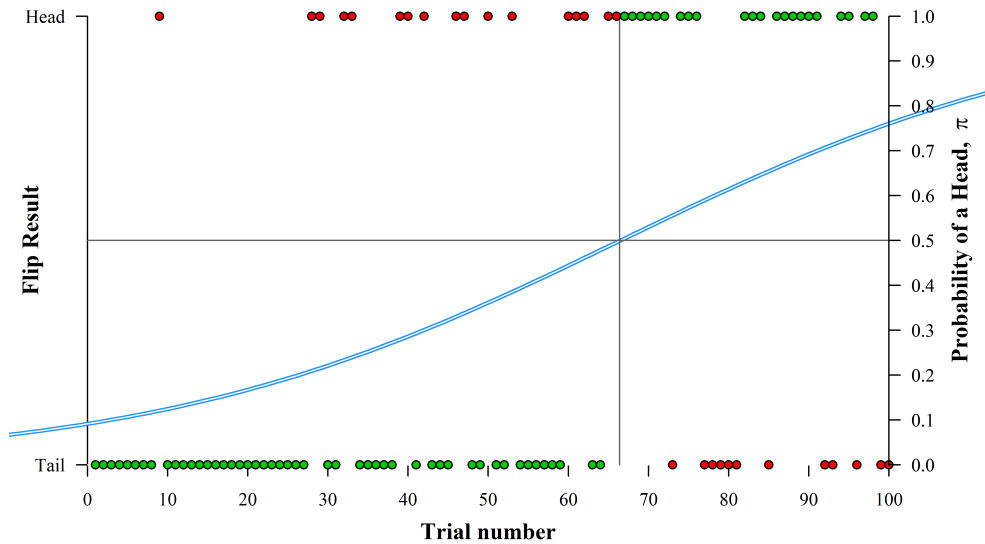


Figure 12.5: Overlaid plot of the outcome of the experiment with the estimated probabilities superimposed. The horizontal line is the $\tau = 0.500$ threshold. The vertical line corresponds to a trial number corresponding to that threshold ($\tau = 66.4$). Thus, this model predicts that all coins above number 66.4 have a probability of greater than half of coming up Heads. Red dots are misclassified, green dots are properly classified.

Note: The linear predictor is represented in the curve graphed in Figure 12.5. Note, however, that the curve is *not* linear. This is because the curve in Figure 12.5 is actually the logistic of the linear predictor.

With that said, the curve *is* linear in the transform space. If you graph the coin number against the logit of the head probability, the line of best fit is, indeed, a line (see Figure 12.2.)

12.5: Prediction Accuracy

Naturally, the next questions concern issues of goodness-of-fit: How good is the model? This question can be answered in many ways using many related accuracy measures.

Recall that in linear regression, we used R^2 to help us determine how well the model fit the data — an R^2 value close to 1.00 indicated good fit, while an R^2 value close to 0.00 indicated a poor fit. If we recall, the R^2 value — a PRE measure — was calculated using a ratio of the original variability in the data and the variability explained by the model (Section 2.4). The R^2 value was not the only PRE we have covered. Many others exist. Similar processes can be used in *this* context to create a pseudo- R^2 measure.

pseudo- R^2

Note: This measure is a pseudo- R^2 measure primarily because it shares some of the characteristics of the true R^2 measure, namely that it measures the decrease in prediction variability due to the model. It is a PRE measure.

The reason it is not called the R^2 measure is only because that name is taken elsewhere.

12.5.1 ACCURACY RATE Let us define the accuracy rate to be the number of correct predictions divided by the total number of predictions. This makes inherent sense as a measure of goodness of fit since it reads as the proportion of correct predictions.

There is no native accuracy function in \mathbb{R} (this should raise a red flag). However, the `RFS` package provides one. The `accuracy` function takes four parameters: `data` (the data variable), `y` (the binary dependent variable), `model` (the model you fit with the data), and `t` (the threshold). The optional parameter, `rate`, tells the function to return the accuracy rate (default) or the number of accurate predictions (`rate=FALSE`).

Thus, to determine the accuracy of this model for this data using the usual threshold value of $\tau = 0.500$, we would use

```
|| accuracy(data=coin, y=coin$head, model=m1, t=0.500)
```

The result of this command is 0.710, which agrees with our by-hand calculations. Thus, we conclude that this model correctly predicts 71% of the time for *this* data.

12.5.2 RELATIVE ACCURACY Of course, having an accuracy rate of 0.710 does not tell us the entire story. Just as the R^2 from Section 2.4 was based on a ratio of the model variance to the data (null) variance, a better accuracy number would be the accuracy of the model relative to the accuracy of the null model. The accuracy of the null model refers to merely selecting the modal category as our prediction. In this example, the modal category is Tails, as there were 61 Tails in the data. Thus, the accuracy of merely selecting the modal category is $61 \div 100 = 0.610$. So, the *relative* accuracy is

$$A_R = \frac{0.710}{0.610} = 1.164 \quad (12.36)$$

Thus, the model does a 16.4% better job of prediction than does just predicting ‘Tail’ all of the time.

There actually is a proportional reduction in error (PRE) measurement associated with the relative accuracy. Recall that the R^2 value was valuable because it measured the proportion of error explained by the model. For binary dependent variable regression, we can calculate something similar.

$$PRE = 1 - \frac{\text{error with model}}{\text{error without model}} \quad (12.37)$$

Here, we can see that a pseudo- R^2 measure for this data and this model (and this threshold) is

$$1 - \frac{1 - 0.710}{1 - 0.610} \approx 0.2564 \quad (12.38)$$

Thus, we can state that this model (and this threshold) reduced the error by 25.64%. Note that there is a bad quality of this measure: while it can never be greater than 1.0, it *can* be less than zero. However, it will only be less than zero when *your* model is worse than no model at all.

Note: There are *many* different ways of calculating pseudo- R^2 measures. Each of the measures are based on different definitions of ‘error’ or of ‘variability,’ just as the R^2 and the adjusted R^2 are both based on different definitions of variability. Researchers do not agree on much about pseudo- R^2 measures except that they are not useful *in vacuo*, and rarely useful in concert with other measures.

This is why I am offering it here, alongside many other measures of fit. Getting to know your model results is just as important as getting to know your data.

12.5.3 MAXIMUM ACCURACY In each of the above measures, we assumed our threshold was $\tau = 0.500$. In some cases, this is a logical threshold. In some cases, it is chosen arbitrarily. If we treat τ as a parameter, we may be able to get a better prediction model.

The plan is straight forward: Calculate the accuracy for various values of the threshold. The threshold that gives us the best accuracy will be our optimal threshold. Doing this by hand is prohibitive. Using a script to loop through all threshold values is much easier:

```
|| a = numeric()
|| for(i in 1:100) {
||   t = i/100
||   a[i] = accuracy(coin, coin$head, m1, t=t)
|| }
```

Figure 12.6 is a plot of the calculated accuracy for various thresholds. Note that the ‘optimal’ threshold is not $\tau = 0.50$, but $\tau = 0.48$, and the maximal accuracy is 0.73 for that threshold. Note, however, that there is little difference in accuracies between this optimal threshold ($\tau = 0.48, A = 0.73$) and the traditional threshold ($\tau = 0.50, A = 0.72$).

Note: Recall that the standard deviation for (variability of) a binomial random variable is $\sigma_x = \sqrt{n\pi(1-\pi)}$. This takes on a maximum value at $\pi = 0.500$... the success probability for a fair coin. This means that we are *least sure* of our answer nearest $\pi = 0.500$.

The blue envelope of Figure 12.6 contains 95% of the calculated accuracies based on the true population; that is, 95% of the accuracy curves are contained in that envelope. It is very wide. It supports the contention that accuracy (relative or otherwise) matters little in the estimation of an optimal threshold τ .

By the way, since this is all based on generated data, we know the true threshold for a fair coin: $\tau = 0.500$. Binomial random variables contain small amounts of information.

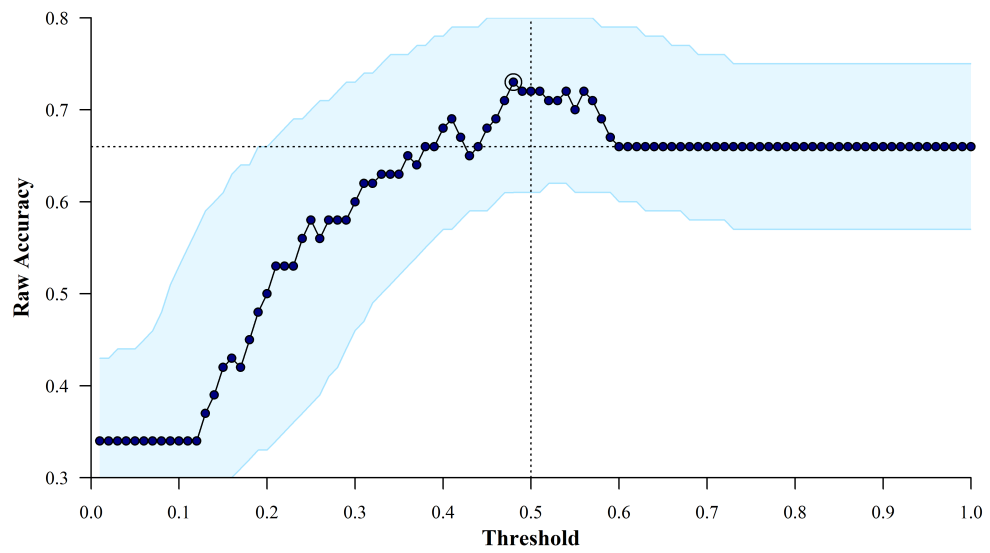


Figure 12.6: A plot of the accuracy of the model against various thresholds. The horizontal line corresponds to the accuracy of selecting the modal category (the base accuracy). The vertical line corresponds to the threshold $\tau = 0.50$. The circled point represents the maximal threshold, $\tau = 0.48$ and accuracy = 0.73. The light blue envelope consists of a 95% confidence interval for coin accuracy, based on Monte Carlo simulation.

Finally, here is the script to estimate the confidence bounds above using Monte Carlo simulation:

```

B = 1e4
acc = matrix(NA, ncol=100, nrow=B)

for( j in 1:B ) {
  thisSample = sample(100, replace=TRUE)
  mod = glm(H[thisSample] ~ cnum[thisSample], family=binomial)
  a = numeric()

  for(tau in 1:100) {
    a[tau] = accuracy( cnum, H, mod, t=tau/100 )
  }
  acc[j, ] = a
}

```

From reading through this script, you should be able to tell what the variables `H` and `cnum` represent. You should also be able to explain the purpose of each line. In fact, you should be able to use this code as the basis of future accuracy investigations.

12.5.4 THE ROC CURVE There are other types of errors, more-specific types, that are useful in other fields. If we look back to Figure 12.5, we see that the threshold line (horizontal) and the corresponding trial line (vertical) divide the dataset into four parts. The lower-left quadrant are those Tails that are correctly predicted by the model and the threshold value to be Tails. The upper-right quadrant are those Heads that are correctly predicted to be Heads. The lower-right quadrant are Tails incorrectly predicted to be Heads. The upper-left quadrant are Heads incorrectly predicted to be Tails. These four types of errors are also referred to as True Negatives, True Positives, False Positives, and False Negatives, respectively.

For our coin flipping example (and with $\tau = 0.500$), we can write out a confusion matrix to show all four of these, both in magnitude and in rates:

confusion matrix

$$\begin{bmatrix} TP = 22 & FP = 12 \\ FN = 17 & TN = 49 \end{bmatrix} \iff \begin{bmatrix} TPR = \frac{22}{17+22} = 0.5641 & FPR = \frac{12}{49+12} = 0.1967 \\ FNR = \frac{17}{17+22} = 0.4359 & TNR = \frac{49}{49+12} = 0.8033 \end{bmatrix} \quad (12.39)$$

The true negative rate (TNR) is also called *specificity*, and the true positive rate (TPR) is called the *sensitivity*. You will come across these two terms in the field of biostatistics and clinical trials because they mirror what physicians and biomedical researchers want out of their diagnostic tests.

The **receiver operating characteristic** (ROC) curve is a graphical representation of the true positive rate against the false positive rate (sensitivity against

ROC Curve

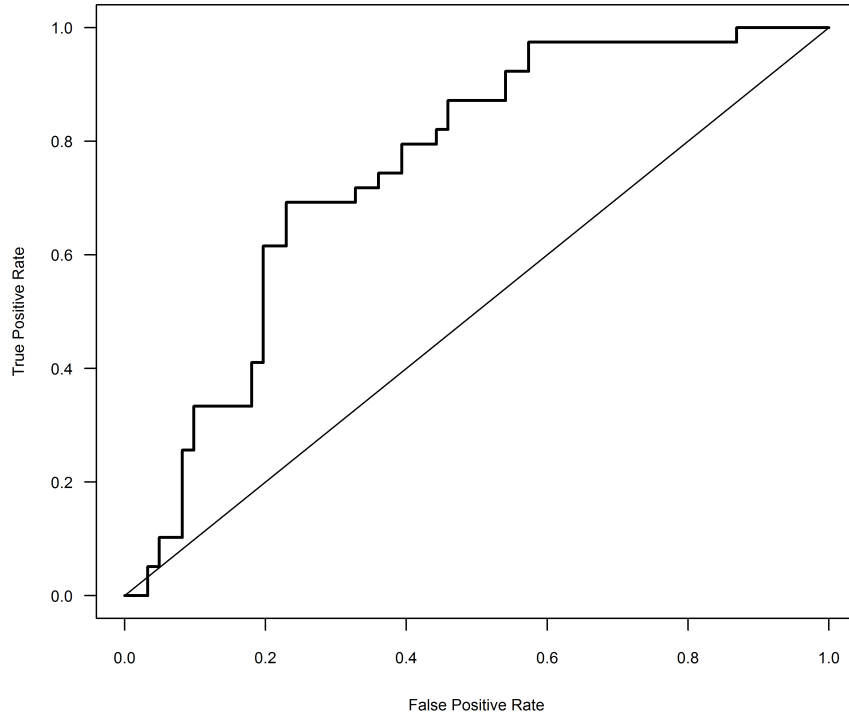


Figure 12.7: A receiver operating characteristic curve for the coin flipping model. The diagonal line represents a random model. The thicker line represents our model. The farther the ROC curve is above the random line, the better the model is at distinguishing between the two cases (Head and Tail, here). The area under the ROC curve is a measure of the goodness of the model. Here, $A' = 0.7516$.

1-specificity) as the threshold is changed. Thus, to plot a ROC curve, one would calculate the true positive and the false positive rates for various values of the threshold, then plot the first against the second. Figure 12.7 shows the ROC curve for our coin model.

In general, a model whose ROC curve is closer to the left and upper axes is the better model. As such, we can define a single number that tells us how good our model is — the area under the ROC curve (AUC, A').

The area under the ROC curve is a useful number in that it equals the probability that a model will classify a positive instance higher than a negative one. In other words, A' is the probability that the model scores a true Head (success) higher than a true Tail (failure). Calculating the area is very straight forward, in a geometry/Riemann Sum manner.

probability

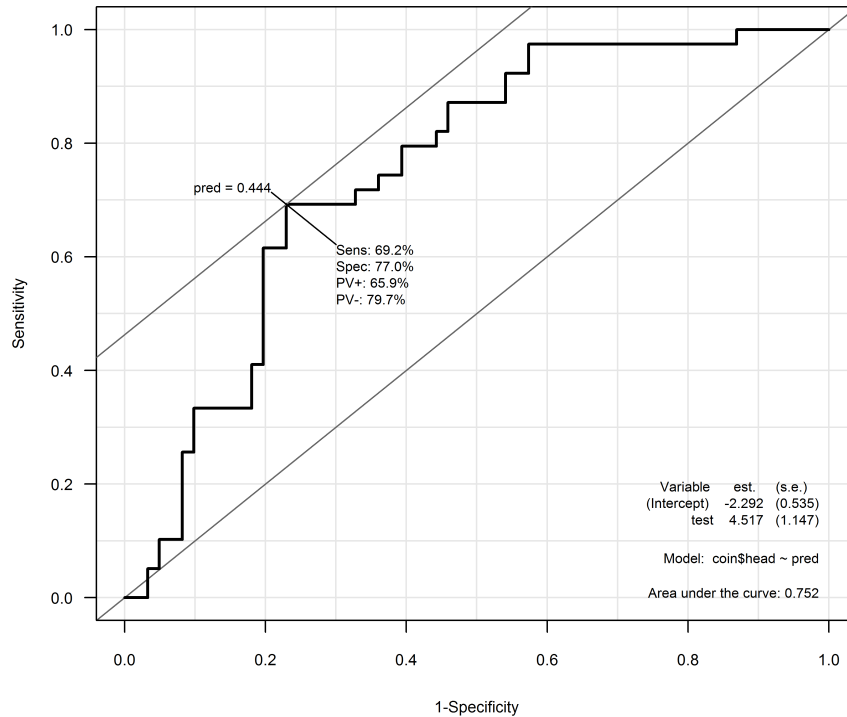


Figure 12.8: The receiver operating characteristic curve for the coin flipping model using the `ROC` command from the `Epi` package.

Note: There is an entire R package dedicated to ROC curves, `Epi`. To create ROC graphs and to calculate the area under the curve in that package, first load it using `library(Epi)`, then use the command

```
ROC(test, stat, plot="ROC")
```

Here, `test` is the predicted probability of success for each datum from model (a continuous variable bounded by 0 and 1), `stat` is the binary dependent variable, and `plot="ROC"` produces a ROC plot. This graph (Figure 12.8) is a bit more useful than the simple graph in Figure 12.7, as it contains some useful statistics, including the AUC and the optimal threshold, τ , which is the threshold value closest to the upper-left corner.

ERROR COSTS: Note that this optimal value is only optimal if the costs of making each type of error is **the same**. If the cost of a Type I Error is greater than that of a Type II Error (or vice-versa), then one should take those costs into consideration when determining the 'optimal' threshold τ .

Not all errors hurt the same.

For instance, if one is modeling fraudulent credit card transactions, then a false positive would happen if the model flagged the transaction as being fraudulent (but it isn't). A false negative would occur if the model did not flag a transaction as fraudulent (but it is).

A high false positive rate would inconvenience the credit card holder. It would reduce their ability to use it. A high false negative rate would inconvenience the credit card bank by forcing them to pay for fraudulent uses of the card.

Not all errors hurt the same (or even the same people).

For instance, given the following table, which of the two models should be used?

Error Type	Error Cost	Model 1	Model 2
FPR	\$50	0.10	0.15
FNR	\$10	0.15	0.10

The first model costs \$6.50; the second, \$8.50. Thus, we should select the first model (or threshold).

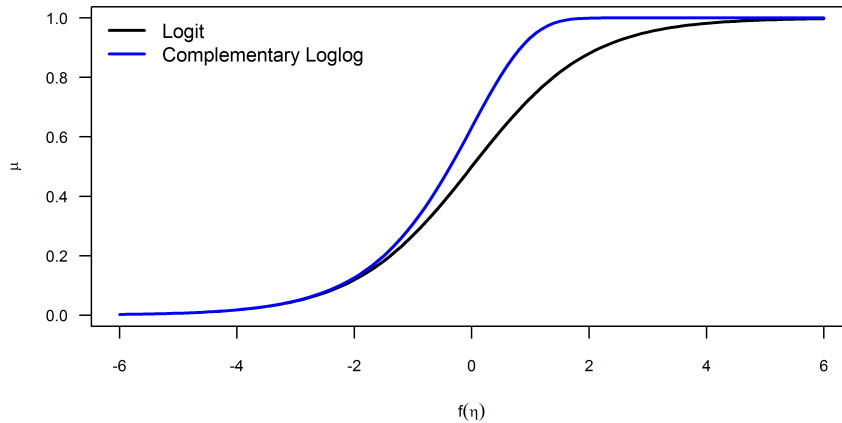


Figure 12.9: Plot of the complementary log-log function (upper curve) on top of the logit. Note the difference in shapes between the two curves. The asymmetric complementary log-log function approaches its maximum value much faster than does the symmetric logit.

12.6: Modeling with Other Links

The logistic regression (or “logit” regression) we did above is quite sufficient if all you want to do is fit the data using logistic regression. If, on the other hand, you want to better understand the process that gave you the data, you will want to try different link functions to determine if any of the alternative links do an appreciably better job of fitting your data. The logit link is symmetric. You should also use the probit link as a check on your model: If the results are comparable, then the conclusions are strengthened; if not, there is something wrong with your model.

In addition to using a second symmetric link function, you should use the two main asymmetric link functions: the complementary log-log and the log-log link function.

12.6.1 THE COMPLEMENTARY LOG-LOG LINK As mentioned earlier, there are several other available link functions beyond the logit link (see Table 12.2). Actually, for binary response variables, all that is required of the link function is for it to be increasing, to smoothly map $g: (0,1) \mapsto \mathbb{R}$, and to have an inverse that smoothly maps $g^{-1}: \mathbb{R} \mapsto (0,1)$. As mentioned earlier, the logit link is symmetric. If you are dealing with rare-events data, you may not want to use a symmetric link function. The complementary log-log link is asymmetric and is often useful (Figure 12.9).⁶

The formula for the complementary log-log is

⁶You may see the complementary log-log link function referred to by its abbreviation — cloglog.

$$g(\pi) := \log(-\log(1 - \pi)) \quad (12.40)$$

Its inverse is

$$g^{-1}(\eta) = 1 - \exp(-\exp(\eta)) \quad (12.41)$$

The plot of the complementary log-log function is seen in Figure 12.9, overlaid with the same plot for the logit link. Note the difference in shapes. Recall that the logit link is symmetric. The complementary log-log is not; it approaches its maximum value more steeply than the logit.

Because of this asymmetry, it will fit models differently. Let us fit the coin data with a complementary log-log link. The command is

```
|| glm(head ~ trial, family=binomial(link="cloglog"), data=coin)
```

Note that the only change is in the link clause. The results of this new model are provided in Table 12.4. Note that the direction of effect is the same in both models. Unfortunately, as the first model is in logit units and the second model is in complementary log-log units, comparing the magnitude of the coefficients tells us nothing. Comparing predictions tells us much more.

Using the logit model, the prediction for π_1 was 0.095. Using the complementary log-log model, the prediction is $\pi_1 = 0.122$, which is closer to the true value of $\pi_1 = 0.150$.

12.6.2 THE LOG-LOG LINK A second useful asymmetrical link function is the log-log link (Figure 12.10). Note that the asymmetric log-log link rises to its maximum much slower than either the symmetric logit link or the asymmetric complementary log-log link. Because of this functional shape, it will be better at fitting certain data sets better than the other link functions discussed.

In reality, there is a functional relationship between the complementary log-log and the log-log link functions. They are 180° rotations of each other. Thus, statistical programs either have no support for either or have support only one. Like most statistics packages, R has native support for only one of the two. For R, it is the complementary log-log link.⁷ However, with the `RFS` package, it is straight forward to perform binary regression using the log-log link.

The command to perform the log-log regression on this data is the same as before, except for the link parameter, which is now

```
link=make.link("loglog")
```

⁷This is actually a decision of history. From how I (and most) have presented the binary dependent variable models, it seems as though we statisticians started with the logit. The first use of this type of regression, however, used the complementary log-log function (Fisher). It was not pretty, but it was a fantastic step in the right direction!

	Estimate	Std. Error	z value	Pr(> z)
Constant term	-2.0651	0.4353	-4.74	$\ll 0.0001$
Trial number	0.0244	0.0063	3.86	0.0001

Table 12.4: *The results of fitting the coin flip data with a complementary log-log link (cf. Table 12.3). As before, the magnitudes of the estimates cannot be compared across different link functions; however, the direction of effect can.*

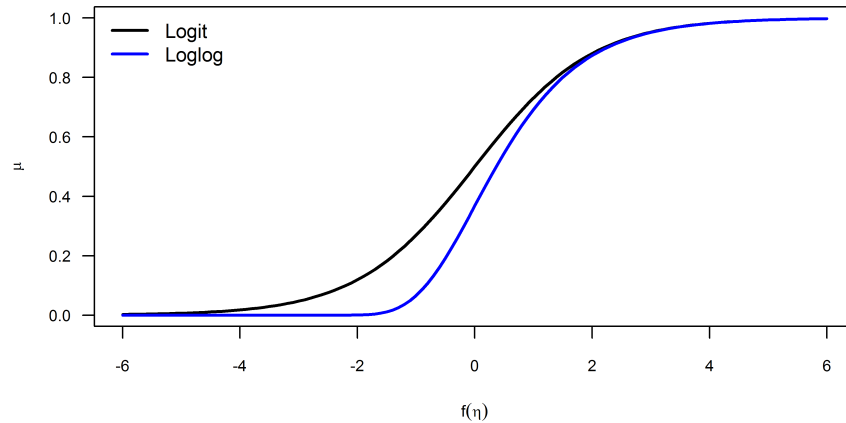


Figure 12.10: Plot of the log-log function (upper curve) on top of the logit. Note the difference in shapes between the two curves. The asymmetric log-log function approaches its maximum value much slower than does the symmetric logit.

With this, I leave it as an exercise for you to show that the effect of trials in the log-log model is 0.0233 and that the predicted probability of a head for Coin 1 using this model is 0.0498.

12.7: Model Selection

Which of the models is best? That is a model selection question. Model selection *procedures* (not tests) attempt to balance competing desires — accuracy and parsimony — to create the ‘best’ model, by some standard. For linear models, we discussed the R^2 value as a measure of accuracy. However, we noted that adding variables to the model can never decrease the R^2 value, and will usually increase it. Thus, there is a pressure to increase the number of variables. However, science is guided by the philosophy of William of Occam and his Razor:

Numquam ponenda est pluralitas sine necessitate.

Plurality must never be posited without necessity. That is, models should be as simple as possible, but no simpler. In other words, as scientists, we should only include variables if the theory warrants it.

Note: Make no mistake, all models are wrong. As scientists, we are merely searching for the useful ones.

shaving

In linear regression, we corrected for the pressure to keep adding variables by using the adjusted R^2 as a guide. This value penalizes the model for the number of variables it has. Thus, unless the variable is statistically significant, there is no benefit to adding it to the model. This is why many scientists use the adjusted R^2 measure to help them model select.

There is neither a true R^2 nor a true \bar{R}^2 value for discrete dependent variable models. Thus, there has been much work in creating an appropriate measure to use for model selection. Three different measures are frequently used in the literature: Akaike's Information Criterion (Akaike 1974), Bayesian Information Criterion (Schwarz 1978), and Likelihood Ratio Test (Wilks 1938). Each of these three penalizes additional variables in a different manner and to a different degree. The one you select depends on the one available to you and the relationship between the two models.

12.7.1 AKAIKE INFORMATION CRITERION One of the first attempts to explicitly penalize for additional parameters (variables) was done by Hirotugu Akaike (1974). In his paper, he developed (albeit without much mathematical rigor) a comparative measure of ‘model goodness’ that can be used to select the better of two models based on the log-likelihood measure. The Akaike Information Criterion (AIC) score can be calculated whenever Maximum Likelihood Estimation is used to estimate the model parameters. The formula for the AIC is

$$AIC := -2\ln(\mathcal{L}) + 2k$$

Here, k is the number of parameters being estimated in the model and \mathcal{L} is the likelihood of the data with the model.⁸

The procedure to determine if one model is better than the other is straightforward:

1. Calculate the AIC for Model A.
2. Calculate the AIC for Model B.
3. The model with the *lower* AIC score is the preferred model.

Its simplicity is its strength. Its weakness is that this measure, called the minimum information theoretical criterion (MAICE) in the paper, has no known probability distribution. As such, there is no way to determine whether the model with the lower AIC is *enough* better to justify eliminating the other from the discussion: If the AIC of Model 1 is 3 less than the AIC of Model 2, do we completely ignore Model 2?

Rule of Thumb

This question actually leads to several “rules of thumb” that determine when that difference is “large enough.” The usual rules of thumb are to drop the model with the higher AIC if the difference is at least 5 (or 8 or 10).

That there is no *a priori* statistical distribution to the AIC score only means the test is not optimal. In his paper, Akaike concurs (1974: 722):

Although the present author has no proof of optimality of MAICE it is at present the only procedure applicable to every situation where the likelihood can be properly defined and it is actually producing very reasonable results without very much amount of help of subjective judgment.

The R function that calculates the Akaike Information Criterion is `AIC`. Using this function, the AIC for each of the three coin models are $AIC_{logit} = 118.48$, $AIC_{cloglog} = 119.74$, and $AIC_{loglog} = 117.05$. Thus, while the log-log is the ‘best’ model from the AIC standpoint, it is not sufficiently better to completely ignore the other two

⁸The quantity $-2\ln(\mathcal{L})$ is often called the **deviance** of the model, which will be used in Section 12.7.3.

models (the AIC improvement is not greater than 5). As such, this procedure is inconclusive with respect to the single model we should choose.

Note: Please keep in mind that for the AIC to be valid in comparing models, the dependent variable values must be the *exactly* same across the models. If not, then this process cannot be used (nor any of these methods). Thus, transformations of the dependent variable mean the AIC cannot be used. Similarly, if data points are removed between two models, the AIC **cannot be used**.

Warning: *Keeping multiple appropriate models is a good idea. Since sufficient science theory does not exist to determine the “right” link, we should keep as many as possible. This will allow us to better understand how much our conclusions depend on our choice of the link function.*



robust analysis

12.7.2 BAYESIAN INFORMATION CRITERION Akaike’s paper did not give a mathematically solid reason why there should be a 2 point penalty for each additional estimated parameter (the $2k$ factor). This created an opening for other researchers to improve upon Akaike’s proof and to create different penalty factors. Schwarz (1978) took Akaike’s idea and put it on a more solid foundation. He humbly called his measure the Bayesian Information Criterion (BIC), others may refer to it as the Schwartz Information Criterion (SIC) or the Schwarz Bayesian Criterion (SBC).

Its formula is quite similar to the AIC:

$$BIC := -2\ln(\mathcal{L}) + k\log(n) \quad (12.42)$$

Here, k is the number of parameters being estimated, n is the number of data points, and \mathcal{L} is the likelihood of the model. Thus, the difference between the AIC and the BIC is the effect of the additional parameter. In the AIC, each additional parameter penalizes the score by 2 points; in the BIC, $\log(n)$ points — usually a much greater penalty.

The process to select the better of two models is the same as for the AIC: Select the model with the lower BIC score (including the rules of thumb). Furthermore, the requirement that the dependent variable values are the same between the models remains.

12.7.3 LIKELIHOOD RATIO TEST Frequently, we wish to determine if a group of variables are **jointly significant** in the model. To do this, we compare the two nested models. We say that Model B is *nested* in Model A if Model A contains all the same variables as does Model B, plus at least one other. For instance, let Model A contain

the variables x_1, x_2, x_3, x_3^2 , and x_4 . Let Model B contain variables x_1, x_2 , and x_3 . Here, Model B is nested within Model A. Now, if we want to determine if variables x_3^2 and x_4 are jointly significant, then we merely compare Models A and B. To do this, we can use the AIC or the BIC, but the Likelihood Ratio test is more statistically clean.

The Likelihood Ratio test is superior to the AIC and BIC — when it can be used — because there is a known asymptotic probability distribution for the test statistic. As such, we can determine whether Model A is *significantly* better than Model B — i.e., whether variables x_4 and x_5 are jointly significant.

jointly

The procedure is also straight forward:

1. Calculate the deviance for Model A.
2. Calculate the deviance for Model B.
3. The difference between the two deviances is distributed as a chi-squared random variable with degrees of freedom equal to the parameter (variable) difference in the two models.⁹

The deviance of a model is defined as

$$D := -2\ln(\mathcal{L}) \tag{12.43}$$

⁹Technically, the distribution of the test statistic is only *asymptotically* chi-square. For small sample sizes, you may want to use simulation to obtain a more accurate test.

Thus, if Model B is nested in Model A, the test statistic is equal to

$$TS := D_B - D_A \sim \chi_{v_A - v_B}^2 \quad (12.44)$$

Here, v_A is the number of parameters in Model A; v_B , in Model B.

Example 3

Let us assume that Model A uses three variables, X1, X2, and X3, and has a log-likelihood of -20, and Model B uses one variable, X1, and has a log-likelihood of -22. Are variables X2 and X3 jointly significant?

Solution: This is an application of the Likelihood Ratio test. The test statistic is

$$TS := D_B - D_A \quad (12.45)$$

$$= (-2 \ln(\mathcal{L}_B)) - (-2 \ln(\mathcal{L}_A)) \quad (12.46)$$

$$= (-2(-22)) - (-2(-20)) \quad (12.47)$$

$$= 44 - 40 = 4 \quad (12.48)$$

This test statistic is approximately distributed as a chi-squared random variable with $3 - 1 = 2$ degrees of freedom; that is, $TS \sim \chi_2^2$.

A chi-squared table gives us a p-value of approximately $p = 0.15$. This is close to what R gives us:

$$\text{pchisq}(4, \text{df}=2, \text{lower.tail}=\text{FALSE}) = 0.135$$

Thus, we conclude at the $\alpha = 0.05$ level that we cannot reject the null hypothesis and we conclude that the restricted model is not significantly different from the full model.

That is, we conclude that the two variables are not jointly significant and we can use Model B in lieu of Model A with little loss of precision. ♦

12.8: Conclusion

This chapter covered a lot of material. First, we examined how to fit binary dependent variable models. The GLM paradigm allows us to easily fit such models. As in all uses of the GLM paradigm, we need to know three things: the conditional distribution of the dependent variable, the linear predictor, and the link function that connects the two.

For binary dependent variables, the dependent variable is distributed Bernoulli. The linear predictor is the usual combination of our independent variables. The canonical link is the logit link. Additional link functions include the probit, log-log, and complementary log-log functions.

The chapter proceeded to examine issues of determining how well a model fits the data. Accuracy, relative accuracy, and maximum accuracy measures were examined. Additionally, we examined the ROC curve and how it gives us additional information about our model.

Finally, we examined general techniques to select between two models. Three methods were examined. The first two did not require that the two models be nested. Both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) only required that the dependent variables be the same. Also in both cases, the model with the lower score was the preferred model, although when that difference was less than 8 there was no reason to jettison the higher-scoring model.

The Likelihood Ratio Test was superior to the two Information Criterion tests as the test statistic has a known asymptotic distribution (χ^2). Thus, we could test the statistical significance of multiple variables at once. The drawback to using the Likelihood Ratio test is that the compared models needed to be nested.

The next chapter continues our examination of discrete dependent variables. Frequently, our outcome variable is a *count* of events. In such a case, we cannot use the techniques discussed in this chapter as the dependent variable takes on more than just two values. We also cannot apply the techniques of Chapter 11, as the dependent variable is not continuous.

Staying in the realm of GLMs allows us to fit such variables easily. All we need to do is determine the appropriate distribution, the linear predictor, and the link function.

12.9: End-of-Chapter Materials

12.9.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

Epi This package adds several functions and procedures related to epidemiology. As it is not a part of the base installation for R, you will need to install it before you can load it with `library(Epi)`.

STATISTICS:

lm(formula) This function performs linear regression on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable.

glm(formula) This function performs generalized linear model estimation on the given formula. There are three additional parameters that can (and often should) be specified.

The `family` parameter specifies the distributional family of the dependent variable, options include `gaussian`, `binomial`, `poisson`, `gamma`, `quasibinomial`, and `quasipoisson`. If this parameter is not specified, R assumes `gaussian`.

The `link` parameter specifies the link function for the distribution. If none is specified, the canonical link is assumed.

Finally, the `data` parameter specifies the data from which the formula variables come. This is the same parameter as in the `lm()` function.

predict(model, newdata) As with almost all statistical packages, R has a `predict` function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

accuracy(model) This function in the `RFS` package determines the predictive accuracy of a provided model. It takes three necessary parameters: `data`, `truth`, `model`, and `threshold`. It has the optional parameter of returning the number of correct classifications (`rate=FALSE`).

AIC(model) This function calculates the Akaike Informations Criterion score for the provided model. The model needs to have been fit using Maximum Likelihood Estimation.

BIC(model) This function in the `RFS` package calculates Schwarz's Bayesian Information Criterion (BIC) for the provided model.

deviance(model) This function returns the deviance in the model. This value is useful in the Likelihood Ratio Test.

pchisq(x) This gives the value of the cumulative distribution function (CDF) under the Chi-squared distribution. The necessary parameter is the number of degrees of freedom, `df=`. By default, it returns the lower-tail probability. Usually, we will want to have the upper-tail probability, thus we will use the `lower.tail=FALSE` parameter.

var.test(x,y) This function performs an F test, which compares the variances of two samples (`x` and `y`) from Normal populations. It can only compare two samples. If you need to compare more than two samples for equality of variance, you will need to perform either a Bartlett test or a Fligner-Killeen test.

GRAPHICS:

ROC(formula) This function in the `Epi` package performs ROC analysis on the data. It provides a ROC graph as well as some statistical values.

PROGRAMMING:

for This command is one of the basic control-constructs in the R language (as in most programming languages). The usual use is `for(var in seq) expr`, where `var` is the looping variable (the variable that equals the current loop number). The parameter `seq` is a vector of values. Usually, `seq` = something like `1:100`, which is a vector of values from 1 to 100. Finally, `expr` is the expression (or series of expressions) that are performed for each value in the `seq` vector.

12.9.2 EXERCISES This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 12 folder. For each of the following problems, please save the associated R script in the chapter folder as `ext0x.R`, where `x` is the problem number.

1. In Example 12.1, we suggested that you fit the provided pseudo data with linear regression and the OLS method. Please do so now.
2. From Section 12.6.2, please fit the `coin` data with the formula `head~trial` and the log-log link. What is the predicted probability of getting a Head on Coin 15?
3. Use the coinflip data (`coinflips.csv`) to estimate the coin that is closest to being fair (a probability of producing a head is closest to 0.500). Use multiple link functions and select which you think is the best.
4. Let us revisit the `cows` data. One of the variables is `passed`, which is a binary variable indicating whether the ballot measure passed. Your job is to predict the proportion of voters in Děčín who will vote in favor of the bill to limit cows. Do not use the `pctFavor` variable. Decide which model you are supposed to use. Prove that your model is the best model available. Make your prediction of the vote share. Include graphs if you would like, but only if the graph helps to illustrate your point.

12.9.3 APPLIED READINGS

- Judi Bartfeld and Myoung Kim. (2010) "Participation in the School Breakfast Program: New Evidence from the ECLS-K." *Social Service Review* 84(4): 541–62.
- Regina P. Branton. (2009) "The Importance of Race and Ethnicity in Congressional Primary Elections." *Political Research Quarterly* 62(3): 459–73.
- Denise Gammonley, Ning Jackie Zhang, Kathryn Frahm, and Seung Chun Paek. (2009) "Social Service Staffing in U.S. Nursing Homes." *Social Service Review* 83(4): 633–50.
- Michael A. Neblo. (2009) "Meaning and Measurement: Reorienting the Race Politics Debate." *Political Research Quarterly* 62(3): 474–84.
- Lenna Nepomnyaschy and Irwin Garfinkel. (2011) "Fathers' Involvement with Their Nonresident Children and Material Hardship." *Social Service Review* 85(1): 3–38.
- Joseph G. Pickard, Megumi Inoue, Letha A. Chadiha, and Sharon Johnson. (2011) "The Relationship of Social Support to African American Caregivers' Help-Seeking for Emotional Problems." *Social Service Review* 85(2): 247–66.
- Brian Kelleher Richter, Krislert Samphantharak, and Jeffrey F. Timmons. (2009) "Lobbying and Taxes." *American Journal of Political Science* 53(4): 893–909.
- Lori E. Ross, Rachel Epstein, Corrie Goldfinger, and Christina Yager. (2009) "Policy and Practice regarding Adoption by Sexual and Gender Minority People in Ontario." *Canadian Public Policy / Analyse de Politiques* 35(4): 451–67.

12.9.4 THEORY READINGS

- Hirotugu Akaike. (1974) “A New Look at Statistical Identification Model.” *IEEE Transactions on Automatic Control* 19(6): 716–23.
- Hirotugu Akaike. (1977) “On Entropy Maximization Principle.” In: P. R. Krishnaiah (Editor). *Applications of Statistics: Proceedings of the Symposium Held at Wright State University, Dayton, Ohio, 14-18 June 1976*. New York: North Holland Publishing, 27–41.
- George Casella and Roger L. Berger. (2002) *Statistical Inference*, Second edition. New York: Duxbury.
- Peter McCullagh and John A. Nelder. (1989) *Generalized Linear Models*. London: Chapman and Hall.
- Gideon E. Schwarz. (1978) “Estimating the dimension of a model.” *Annals of Statistics* 6(2): 461–64.
- Samuel S. Wilks (1938) “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.” *The Annals of Mathematical Statistics* 9(1): 60–62.

CHAPTER 13:

BINOMIAL DEPENDENT VARIABLES

OVERVIEW:

In the previous chapter, we examined what we can do if the dependent variable is dichotomous, has only two possible outcomes. The response variable followed a Bernoulli distribution. In this chapter, we extend this idea to where the outcome variable follows a Binomial distribution — conditional on the values of the independent variable(s).

In this chapter, we first define a Binomial random variable, we then show that the Binomial distribution is a member of the Exponential Class of distributions (EC). With that, we can look at the assumptions and a couple extended examples.

Chapter Contents

13.1	Binomial Distribution	377
13.2	The Mathematics.	379
13.3	Full Example: O Canada!.	381
13.4	Full Example: Sri Lanka in 2010	392
13.5	Beta-Binomial Regression*	397
13.6	Conclusion	400
13.7	End-of-Chapter Materials	401



In the previous chapter, we explored a situation where the classical linear model utterly failed. That failure has been known since the early parts of the 20th century. As a result, statisticians created several specialized fitting techniques for binary dependent variables.

The importance of the generalized linear model is not that it can fit models with a binary dependent variable, but that it is a technique that can fit those models *and* many others. This allows the researcher to pay closer attention to the dependent variable and use more information.

In *this* chapter, we cover the case where the dependent variable represents a count of successes out of a known number of attempts, n .

While least squares can be used in such cases, even without clearly violating its assumptions, it is inefficient. You will be throwing out important information, namely the distribution of the dependent variable. Statisticians hate to throw out information.

This chapter starts with exploring the most common type of variable representing a count of successes out of a known number of attempts. After working with the Binomial, it looks into what happens when the data show the wrong level of variability... and what to do about it.

13.1: Binomial Distribution

The Binomial distribution is defined as the sum of independent and identically distributed Bernoulli random variables. Subsequently, this results in five requirements for a random variable to follow a Binomial distribution:

1. the number of trials, n , is known;
2. each trial has two possible outcomes: success and failure;
3. the success probability for each trial, π , is constant;
4. each trial is independent from the others; and
5. the random variable is the number of successes in those n trials.

One can also think of the Binomial distribution as a generalization of the Bernoulli distribution to $n > 1$. Similarly, one can think of the Bernoulli distribution as a special case of the Binomial, where $n = 1$. Whichever way you look at it, there are a number of similarities between the two distributions.

The sample space for the Binomial distribution is key to understanding when it can be used, $\mathcal{S} = \{0, 1, 2, \dots, n\}$. Thus, the Binomial distribution can be used to model counts of successes when the number of attempts (trials) is known. The following are variables that could follow a Binomial distribution:

- number of students passing a class
- number of Euchre games a person wins in a tournament finals
- number of college students in a class who can locate Ruritania on a map
- number of football games the SUR Hawks win in a year
- number of pages in a book that have a typographical error
- number of fireworks in a shipment of 144 that are duds
- number of cast ballots declared invalid in an electoral division

Note that each of these examples start with “number of.” This is because the Binomial distribution models the “number of” successes. Second, note that the sample space has both a lower (no successes) and an upper bound (no failures). The upper bound for the first example is the number of students taking that statistics class. The upper bound for the second is the number of hands played by a person in a poker game; for the third, the number of students in that class; for the fourth, the number of games the SUR Hawks play in a year; etc.

The following variables *cannot* follow a Binomial distribution:

- number of crimes in D  c  n this year

- number of injuries experienced by the SUR Hawks
- number of errors in a book
- number of dents on a car

While each of these also starts with “the number of,” none of these can follow a Binomial distribution. In each case, there is *no upper bound*. The first example measures the number of successes over a time period. There can be multiple crimes on a given day, so there is no upper bound. To make this a Binomial random variable, one could measure instead the number of Dêcín residents who are the target of a crime in a given year. In that case, there is an upper bound — the population of Dêcín.

The second example also has no upper bound. Each player can have multiple injuries. To make this a possible Binomial random variable, one could measure the number of players injured. Note that the upper bound would then be set at the number of players on the Hawks.

It is interesting that these four random variables could be examples of Poisson random variables. We will be covering how to analyze such count data later (see Chapter 14).

Note: One thing that may help you determine whether a variable follows a Binomial or a Poisson distribution could be this: If you can represent a similar variable as a proportion, then it is Binomial; if as a rate, then Poisson.

13.2: The Mathematics

Remember from Chapter 11 that performing generalized linear modeling requires that we specify three things about our model:

- the linear predictor;
- the conditional distribution of the dependent variable; and
- the function that links the two.

13.2.1 LINEAR PREDICTOR As usual, the linear predictor is the function that relates the independent variable(s) with the dependent variable. For k predictor (independent) variables, the linear predictor is

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (13.1)$$

Frequently, this is what the researcher cares most about. This is the start of where we can test if certain variables can help in better understanding the data-generating process.

13.2.2 CONDITIONAL DISTRIBUTION The second need is the conditional distribution of the dependent variable, the distribution of Y given the values of the x -variables. For the Binomial distribution, the probability mass function (pmf) is

$$f(y, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad y \in \{0, 1, 2, \dots, n\} \quad (13.2)$$

I leave it as an exercise to show that the Binomial distribution is a member of the Exponential Class of distributions. In other words, you will need to show that the above probability mass function can be written as

exercise

$$f(y, \pi) = \exp \left[\frac{y \log(\pi) + n \log(1 - \pi)}{1} + \log \binom{n}{y} \right] \quad (13.3)$$

With this, we can calculate $\mathbb{E}[Y]$ and $\mathbb{V}[Y]$.

exercise

Note that Equation 13.3 shows us that the canonical link is the logit function, $g(\pi) := \text{logit}(\pi)$. As always, the canonical link offers some mathematical cleanness but little else. If the situation calls for a different link function, you should use it.

13.2.3 LINK FUNCTION From the previous section, we know that the canonical link function is the logit function. However, as discussed in Section 12.3.1, many alternative link functions are available. If the model is sound, then predictions based on those alternatives will tend to be similar. Let me emphasize that here:



Warning: *It is extremely rare that the link function can be determined from the scientific theory — extremely rare. Thus, if the model significantly depends on the choice of link, then the model is weak. You should improve the model.*

This also suggests another model test. Fit the model using several link functions. That the results are substantively the same across the link functions supports the goodness of your model.

13.2.4 ASSUMPTIONS/REQUIREMENTS As you have read through this chapter, what assumptions were made? Those are the requirements you need to check. Allow me to repeat this *extremely* important point:

Note: As you have read through this chapter, what assumptions were made? Those are the requirements you need to check.

13.3: Full Example: O Canada!

Let our first full example concern finding Canada on a world map. The research question is

Question

What is the relationship between the type of course being taken, the average age of the students in the class, and the proportion of students who can locate Canada on a map?

According to the wording, the measurement unit is the classroom; that is, each record represents a classroom. For each classroom, four measurements are taken: number of students in the class (trials), number of students who could locate Canada on the map (successes), type of class (independent variable), and average age of students in the class (independent variable).

The pseudodata are located at

<http://rur.kvasaheim.com/data/ocanada.csv>

The first step is to determine the **linear predictor**. The two independent variables are the course type taught to the class, `course`, and the average age of the students in the class, `averageAge`.



Figure 13.1: A map of the world, as of 2007. See if you can locate Canada on it.

Thus, the linear predictor is

$$\eta = \beta_0 + \beta_1 \text{course} + \beta_2 \text{averageAge} \quad (13.4)$$

The second thing is to determine the **conditional distribution** of the dependent variable. Here, note that the dependent variable is the number of students in the class who can locate Canada on the map. This random variable is the number of successes out of a total number of trials (number of students in the course). Because of this, the dependent variable may follow a Binomial distribution.¹

$$Y | x \sim \text{Bin}(n, \pi) \quad (13.5)$$

The third, and final, aspect of a generalized linear model that needs to be specified is the **link function**. This is the function that links the ranges of the distribution's expected value and the unbounded linear predictor, η .

Before we can move ahead, we need to think about this expected value. From elementary statistics, we know $\mathbb{E}[Y | x] = n\pi$ for a Binomial distribution. However, because n varies from classroom to classroom, and because we really care about the proportion of students who can find Canada on the map, it is *always* preferable to “divide out by n ” and focus on modeling π , our parameter of interest. This is what actually happens in the estimation procedure.

Back in Section 13.2.2, we decided that the canonical link was the logit function. In fact, any function that maps $(0,1) \mapsto \mathbb{R}$ would be appropriate. Using the logit function, we now have

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{course} + \beta_2 \text{averageAge} + \varepsilon \quad (13.6)$$

Furthermore, to increase the evidence that our basic model is appropriate, we should fit with multiple link functions (see Sections 12.3.2 and 13.2.3). Perhaps we should also fit with the probit and cauchit link functions.

¹Note that it is likely that the observations are not independent: those in a class will be more similar in terms of geographic knowledge. Thus, while the Binomial distribution is likely, we may find need to use a different estimation method than maximum likelihood to take into consideration the dependence (a.k.a. clumping).

13.3.1 PUTTING IT TOGETHER Note that the three parts combine to make the following distribution statement:

$$Y_i \sim \text{Bin}\left(n_i, \underbrace{\text{logistic}(\beta_0 + \beta_1 \text{course}_i + \beta_2 \text{averageAge}_i + \varepsilon_i)}_{\pi_i}\right) \quad (13.7)$$

The importance of Equation 13.7 is that it shows how the three parts fit together into a coherent whole.

13.3.2 THE MODELS Here is how to fit the model in R using these three link functions

```
depVar = cbind(correct, classSize-correct)
modL = glm( depVar ~ course + averageAge,
            family=binomial(link=make.link("logit")) )
modP = glm( depVar ~ course + averageAge,
            family=binomial(link=make.link("probit")) )
modC = glm( depVar ~ course + averageAge,
            family=binomial(link=make.link("cauchit")) )
```

Note that the first thing we need to do is specify both the number of successes and the number of failures. This is the variable `depVar`, the successes (`correct`) stacked against the failures (`classSize-correct`) as a single dependent variable.

Closely study the three modeling statements. In them you will find the linear predictor, the conditional distribution, and the link function.

13.3.3 THE ASSUMPTIONS Before reading this section, read back through Section 13.2 and determine the assumptions. Once you have them, compare your list to the following:

The first assumption is that the conditional distribution of the dependent variable is the Binomial distribution. Section 13.1 provides the five requirements for a Binomial distribution. Usually, it will be quite easy to meet requirements 1, 2, and 5. The other two requirements may, or may not, be met by the data.

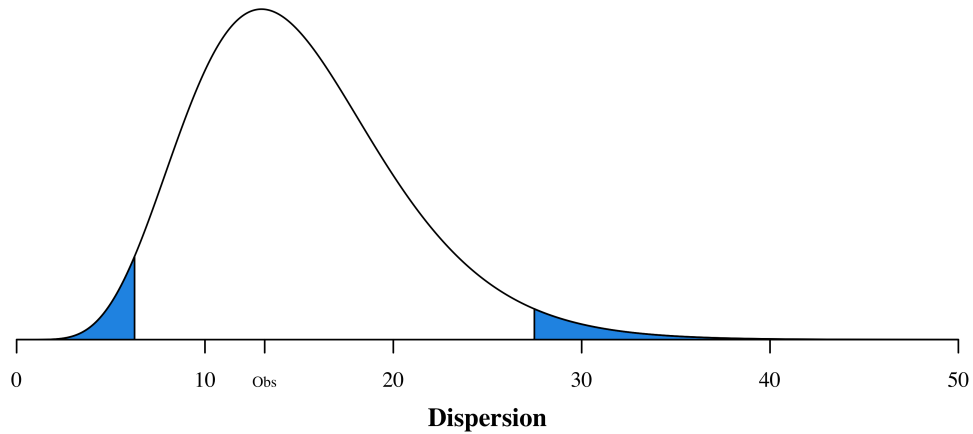


Figure 13.2: The distribution of the dispersions for the O Canada example. Note that the shaded regions are the rejection region. That the observed dispersion (“Obs”) is not in the rejection region tells us that there is no sufficient evidence that the population dispersion is anything other than 1.

exactly

VARIANCE (DISPERSION): At this point, the usual result of violating Assumptions 3 and/or 4 is to have the variance of the variable *not* be what is expected. Recall the value of $a(\phi)$ in Equation 13.3 is 1. That means that *if* the random variable exactly follows a Binomial distribution, then the data’s dispersion is approximately 1. To calculate the observed dispersion parameter, divide the residual deviance by the residual degrees of freedom.

It can be shown² that the deviance asymptotically follows a Chi-square distribution with $n - p$ degrees of freedom:

$$-2\ln \mathcal{L} \sim \chi_{n-p}^2 \tag{13.8}$$

As such, we can create a test for overdispersion... or for “non-unit dispersion.”

Running `summary(modL)` gives the residual dispersion as 197.56 and the degrees of freedom as $\nu = 15$. The ratio for this sample is 13.17067. Do we have sufficient evidence that the ratio differs from 1 for the population?

²... meaning that this proof is beyond the scope of this course...

Well, the usual 95% confidence interval for the sample dispersion, given that the population dispersion is 1, is from 6.26 to 27.49:

```
|| qchisq(c(0.025,0.975), df=15)
```

Since the observed value of 197.56 is above this interval, we can conclude that there is significant evidence of overdispersion (see Figure 13.2).³

What do we do when **there is evidence of overdispersion**? That depends on the effects of overdispersion. At this level, the usual effect of overdispersion is that the variance estimate is not correct. In maximum likelihood estimation, the usual method for estimating all parameters in generalized linear models, the dispersion is required to be 1 in the Binomial case because the value of $a(\phi) = 1$.

The maximum *quasi*-likelihood method includes an additional variable in the denominator, thus changing it from $a(\phi)$ to $za(\phi)$. In the case where $a(\phi)$ is not a constant, such as the Gaussian, this addition offers absolutely nothing. Estimating $a(\phi)$ from the data gives no information above estimating $za(\phi)$ from the data.

However, when $a(\phi)$ is a constant, such as with the Binomial (or the Poisson), including z allows more flexibility in the model. Make no mistake, if $z \neq 1$ then the actual distribution is *not* Binomial. The inclusion of z allows for distributions that are “the same” as the Binomial with the exception of the variance.⁴

Fitting the model using maximum quasi-likelihood is almost as easy as fitting it using maximum likelihood estimation. The only difference is adding `quasi` to the family name:

```
|| QmodL = glm( depVar ~ course + averageAge,
               family=quasibinomial(link=make.link("logit")) )
```

Compare the summary of *this* model, `summary(QmodL)`, with the summary from the MLE model, `summary(modL)`. Notice that the effect estimates are identical. The difference is in the estimate of the standard errors, ... which affects the estimates of the test statistics, ... which affects the estimates of the p-values.

Note: When the dispersion is greater than one, the p-values estimated using maximum likelihood are too low because they assume $a(\phi) = 1$. When the dispersion is less than one (a rare event), the p-values estimated using maximum likelihood are bigger than reality.

³Note that we are working with both ratios *and* totals in this discussion. Knowing the difference in terminology is important in understanding. The statement “the population dispersion is 1” means that the *ratio* is 1; that is, the dispersion is 1 for *each* degree of freedom. That is why the number of degrees of freedom, $n - p$, is within the interval while 1 is not.

⁴Such distributions are termed “overdispersed Binomial” distributions.

The abbreviated output from the `summary(QmodL)` function is

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -9.9030     6.2123  -1.594  0.1318
courseHumanities -0.1049     1.0654  -0.098  0.9229
coursePhysical Science  1.2328     1.2901   0.956  0.3544
courseSocial Science -0.9069     1.0270  -0.883  0.3911
averageAge      0.5342     0.3002   1.780  0.0954
---
(Dispersion parameter for quasibinomial family taken to be
 11.95446)

Null deviance: 276.86  on 19  degrees of freedom
Residual deviance: 197.56  on 15  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

```

Most of the output table should be familiar by now. Note that the “Number of Fisher Scoring iterations” is not important at this point. This is the number of “loops” through the estimating procedure needed until the computer determined the estimates converged.

Since we are fitting this generalized linear model using maximum quasi-likelihood estimation, there is little useful information contained in the deviances. You *could* use them to calculate a measure similar to the infamous R^2 value. Recall that the R^2 value is a PRE (proportional reduction in error) measure (see Section 12.5.2). Understanding what a PRE measure is supposed to measure allows us to create a formula for one:

$$\text{pseudo-}R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}} \quad (13.9)$$

pay attention

Does Equation 13.9 makes sense to you? The better you understand the meanings of the terms, the better your understanding of Equation 13.9. Make sure you realize that these equations are not unique. There are other ways of measuring both “before uncertainty” and “after uncertainty.” Different measures of each will lead to different equations for the PRE.⁵

From the regression table, we can tell that there is no relationship detected between the average age in the class and the proportion of students correctly finding Canada on a map ($p = 0.3911$). Thus, should we be in the model selection phase, we should drop this variable and fit the new model.

⁵This is why there are so many PRE measures. Typically, specific measures are traditional/expected for specific models. Usually, the only thing special about those PRE measures is that they came first. Regardless, including them in a report is expected.

Is the course-type variable statistically significant? From this table, we cannot tell. Each of the three lines in the regression table devoted to that variable are *actually* specifying the level's effect *with respect to* some base category ("Arts" here). That means the table can only tell us if one level is statistically different than the effect of the Arts level. We cannot tell whether the *variable* is significant as a whole.

Aretha Franklin

Think through that last paragraph. It gives you a hint on how to determine if a categorical *variable* significantly affects the dependent variable.

contemplate

If the *variable* is significant in modeling the dependent variable, then removing it from the model will produce a significantly worse model. So...how do we measure "significantly worse"? There are many ways. One is to compare the AIC or BIC values between the two models (Section 12.7.1). Note that both AIC and BIC depend on the value of the likelihood. When using maximum quasi-likelihood estimation, there is no likelihood (there is only quasi-likelihood).

worse

only Xul

We could also use the likelihood ratio test. This test relies on the asymptotic distribution of the difference in deviances. Thus, it is great for large-sample cases. How large? It depends on how closely the deviances follow the Chi-square distribution... which means (at its core) this test relies on the Central Limit Theorem.⁶

LRT

⁶From experience, as long as no one will die from you being wrong, a sample size in excess of 50 will suffice in most cases. For a Binomial random variable, if $n\pi$ and $n(1 - \pi)$ are both at least 15, things will tend to be acceptable. When in doubt, make sure you check this assumption for your specific case.

Performing a likelihood ratio test is as easy as fitting the full model, fitting the reduced (constrained) model, and testing if the dispersions are significantly different:

```
QmodLreduced = glm( depVar ~ averageAge,
                    family=quasibinomial(link=make.link("logit")))
anova(QmodLreduced, QmodL, test="LRT")
```

After running these lines, we have the following output

```
Analysis of Deviance Table

Model 1: depVar ~ averageAge
Model 2: depVar ~ course + averageAge
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         18      246.20
2         15      197.56  3   48.635  0.2542
```

Since the p-value is greater than our $\alpha = 0.05$, there is no significant difference between the two models. This means that including `course` does not significantly improve our model ($p = 0.2542$). If we are trying to create our final model, we would drop `course` from consideration and deal with this reduced model (exploratory analysis). If we are testing whether the `course` matters in being able to locate Canada on the map, we would have to conclude that there is no significant evidence that it does (confirmatory analysis).⁷

MODEL FIT: The second assumption is that the “curve of best fit” consistently fits the data. This is the same assumption we met back in Section 5.2. Its tests are the same as always...no change, whatsoever. Thus, the code I would run to check that the link function is appropriate is

```
eL = residuals(modL)
runs.test(eL, order=averageAge)
summary(aov(eL ~ course))
```

The first line calculates the residuals. The second tests the residuals against the numeric variable; the third, against the categorical variable. In neither case is the p-value less than our $\alpha = 0.05$. In fact, neither are even close to 0.05. Thus, there is no evidence that the link we used, the logit link, is incorrect.

⁷There are other testing options in the `anova` function. There are also other ways of performing the likelihood ratio test. In very large sample sizes, the tests will return the same substantive conclusions. In smaller sample sizes, the conclusions may differ. The `anova` function uses an unbiased estimator of the variance, while others may use biased estimators. Thus, it may be better to use the `anova` function...in general.

Since the p-value for the tests of model fit are greater than $\alpha = 0.05$ and since we adjusted for overdispersion, this is an acceptable model. The results of fitting this model using the logit link are given by [summary \(QmodL\)](#). From that model and the above analysis, we can make the following conclusions:

1. There is no significant evidence that the average age in the class has an impact on the class's ability to identify Canada on the map.
2. There is also no significant evidence that the class type affects the ability to identify Canada.

In other words, we have no new information on what influences a class's ability to find Canada on the map.

13.3.4 ETHICALITY Were we unethical, we could create a graphic to illustrate our non-results (Figure 13.3). This graphic strongly indicates that those in physical science classes are much more adept at geography than those in social science classes. Note how much higher the blue curve is than the red curve.

However, Figure 13.3 is misleading — it is *very* misleading. The statistics tell us that there is no evidence that the abilities differ between those in those two types of classes. The graphic lies by suggesting such a difference exists.

As an ethical statistician, be aware that your graphics must illustrate what the statistics actually tell us. They should not suggest that which does not exist. Similarly, they should not minimize an effect that *does* exist. A statistician needs to use graphics to tell the story of the data — and nothing else.

This is not as easy as it seems. Violations of ethics happen when you violate these precepts *by design*. Violations of these tenets may still happen by accident. To avoid claims that you are being unethical, make sure you are clear on your conclusions and why you are making those specific conclusions.

It is possible that two ethical researchers come to different conclusions. It is unlikely that those conclusions are substantively different. If this happens, then ethical researchers will find the results interesting and seek to understand why the models produced such different results.

I would argue that this is the hallmark of scientists. When proven wrong, an ethical scientist will try to better understand the phenomenon to explain the differences in the conclusions.

Remember this: The choice is always between humility and humiliation.

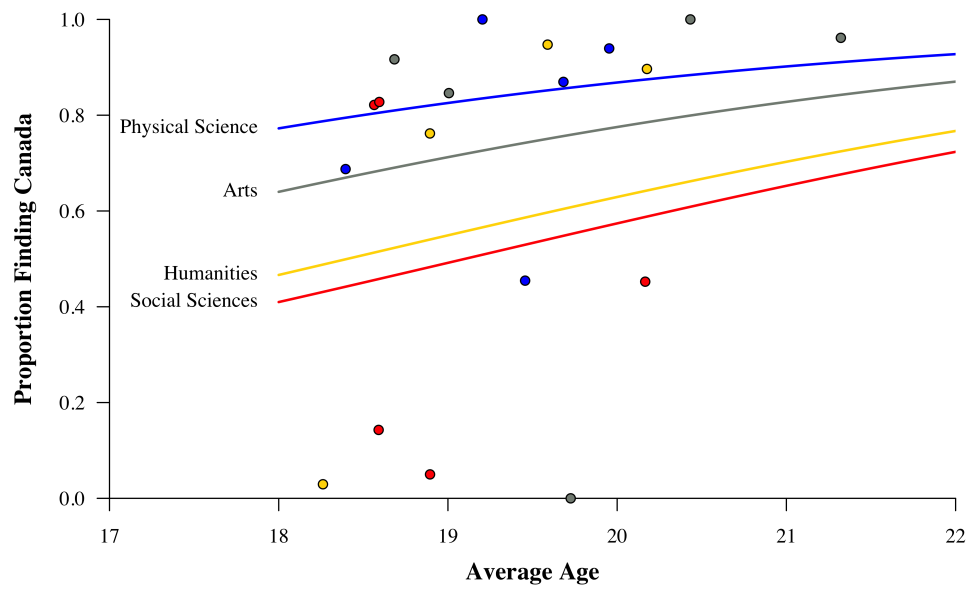


Figure 13.3: A plot of the 'O Canada' data with prediction curves for each class type. The blue is for the Physical Sciences; red, Social Sciences; gold, Humanities; and grey, Arts. The curves are regression curves; the dots, observations. Note that the actual model was unable to detect a difference in effects among the four course types. Thus, this graphic is unethical.

Note: We could actually obtain stronger results if we had data at the student level. This would allow us to forego aggregation and measure the relationship of interest (what allows students to find Canada on the map) directly.

The vast majority of the time, this is a true statement: We are interested in relationships at the individual level, but our data only exists at some level of aggregation.

If we are careful with our wording, this aggregation results in no more than a loss of power. If we are not careful, then we are drawing conclusions not warranted by the model. Strictly speaking, we can only draw clear conclusions on our measurement unit (classes in this example). However, we *can* say that our conclusions are **consistent** with hypotheses about the unit of interest (students in this example).

consistent

[For the following interpretation, let us *pretend* our variables were significant. I am doing this to illustrate how to write out a conclusion.]

pretend

We *can* conclude:

Physical Science classes with older students tend to do better at locating Canada on the map.

How boring and unconnected with a really interesting question about the students themselves.

We could also conclude:

The results are consistent with the hypothesis that Physical Science majors who are older tend to be better able to locate Canada on the map.

Much better, but less decisive. To be “consistent with” is logically quite weak. There are many results “consistent with” any hypothesis. If relying on consistency, your theory needs to be **very strong**. . . or your results *very* interesting.

The best of both worlds is having data at the individual level. Unfortunately, that is not always available. As such, we do need to understand the limitations on our conclusions due to the limitations of our data.

Again, know what you *can* conclude and conclude **no farther**.

13.4: Full Example: Sri Lanka in 2010

For our second full example, let us examine the official returns from the 2010 presidential election in Sri Lanka. In fact, let us check to see if there is evidence of differential invalidation. The term “differential invalidation” refers to certain *types of* ballots having a higher probability of being invalidated. If this invalidation is a function of the ballot recipient, then the election is unfair.

If ballots cast for Candidate X have a higher probability of being invalidated, then the election is unfair against Candidate X. Since we only have ballot counts at the electoral division level, such an unfair election would show itself by having a significant relationship between the invalidation rate at the division level and the support level for the candidate.

Thus, the dependent variable is the proportion of ballots declared invalid by the official counters. The independent variable is the level of support for Candidate X. Both are measured at the electoral division level. If the relationship between these two variables is statistically significant, then we have evidence of differential invalidation. If the slope is also negative, then this helps Candidate X.

differential
invalidation

a very deep
paragraph

measurement unit

why?

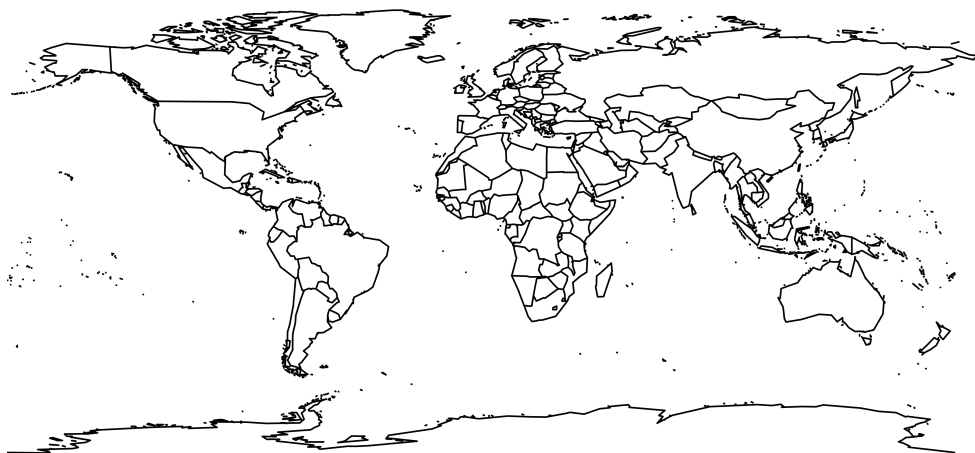


Figure 13.4: A map of the world, as of 2007. Now, see if you can locate Sri Lanka.

Note: These previous paragraphs illustrate limitations of using aggregate data (measurement unit) when we want to draw conclusions about the individual (experimental unit). We must say that the results are “consistent with” our hypothesis about the individual. We can conclude there is “evidence of” our hypothesis. That is about all.

ecological fallacy

The **ecological fallacy** is a logical error that arises when characteristics observed at the *group* level are attributed to *individuals* within that group. Essentially, it involves assuming that what’s true for a population is automatically true for every individual within it. In statistics, this means that findings on aggregated data are not necessary true for the individuals being aggregated.

For example, we know that states with higher percentages of college graduates also tend to have lower rates of obesity. It would be fallacious to conclude that college graduates are less likely to be obese.

An example in US elections: We know that states with higher proportion of their residents being African-American tend to be more Republican. It would be wrong to conclude that African-Americans tend to be Republican.

13.4.1 THEORY APPLICATION An applied statistician needs to be able to translate between the theory of the scientist and the theory of the statistician. This frequently takes a lot of time and practice. Remember to be able to ask questions and to present results from different standpoints.

To illustrate this symbiosis between science and statistics, let’s apply this theory to the 2010 Sri Lankan election. In 2010, Sri Lanka held a presidential election between incumbent Mahinda Rajapaksa and challenger General Sarath Fonseka. Both were instrumental in ending the Sri Lankan civil war between the Sinhalese and the Tamils.⁸

As was expected for Sri Lanka at this time, there was violence on election day, intimidation throughout the campaign, and claims of fraud by both parties. To secure his victory, Rajapaksa had Fonseka arrested and imprisoned. The entire scene is explored by Ratnayake in *That Blue Thing: An Engineer’s Travel*.

With that brief background, the data are located at

<http://rur.kvasaheim.com/data/sri2010pres.csv>

Load the data as usual, but do not attach it... yet. Examine the data. Look at the relationships in the data. Think of the meanings of the data. Become one with the data.

zen

⁸Arguably, the two were not as effective as nature. The 2004 Boxing Day Indian Ocean earthquake caused a tsunami that swamped the Tamil navy — a blow from which the rebel Tamils did not recover.

always

Note: *Always* do this if you are exploring the data; you need to be as aware of the data as possible.

Arguably, this is the most under-appreciated part of statistical analysis. It is also the most tricky. As always, know the difference between exploratory and confirmatory analysis. If you are exploring the data and relationships in the data, then make sure you explore all of their nooks and crannies.

If, on the other hand, you are testing hypotheses about the data (confirmatory analysis), then you will *not* be doing any explorations. You will simply be testing for statistically significant relationships in your model. Always be aware of your purpose.

Doing this here, you will see several records identified as “postal” and several as “displaced.” The former refer to votes sent in the mail. The latter refer to votes by people outside their division. At this point in Sri Lankan history, the country was still suffering the after-effects of a decades-long civil war. A large proportion of the population was displaced from their homelands... especially from the north and east of the country. The “displaced” ballots are the votes of these people.

Note that the number of displaced and postal votes is rather low. We probably should consider dropping these records for a couple of reasons. The most important reason is that the number of invalidated ballots in many of these records is 0, and zeroes tend to cause problems in analyses.⁹

A second reason is that it is unclear where these votes were counted and who counted them. If we are interested in checking how the government in each division affected the vote counting, then the Postal and Displaced votes muddy the argument.

best

Of course, it is best to create two models (as I do here), one with the displaced and postal votes and one without, and fit both models. Remember the goal of science is to better understand relationships.¹⁰ Creating and interpretation *multiple* models allows us to have a greater insight into the relationships.

insight

For the record, there is not much of a difference between the two models. Figure 13.5 is a graphic for the models fit on the data with the Postal and Displaced ballots removed; Figure 13.6, for the models fit on the entire data set. Since the two graphics are very similar, we have more evidence that the underlying model *does* describe reality well. That is, we are much more confident in our conclusion that there is differential invalidation *and* that it benefited Mahinda Rajapaksa.

⁹As an illustration, calculate a 95% confidence interval for a proportion when $x=0$ and $n=10$. The standard Wald confidence interval gives a confidence interval from 0 to 0. This is one reason Agresti and Coull (1998) created a new confidence interval for proportion data.

¹⁰It just so happens, that the substantive conclusions remain the same, regardless of whether or not the displaced and postal votes are dropped.

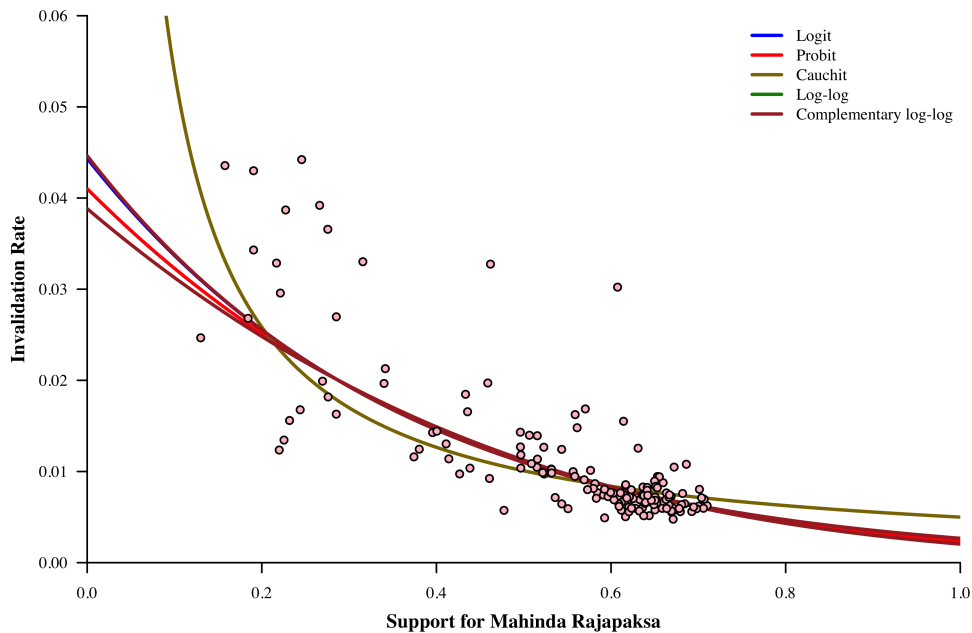


Figure 13.5: An invalidation plot of the Sri Lankan presidential election in 2010 with the postal and displaced ballots removed. The curves are prediction curves for the five main link functions. Only the cauchit link (gold) produces estimates that practically differ from the others. Regardless, all models tell the same substantive story: there is evidence of differential invalidation in favor of Mahinda Rajapaksa.

13.4.2 INTERPRETATION From the graphics (Figures 13.5 and/or 13.6) and the results of the statistical analysis, we can conclude that there is a statistically significant relationship between the invalidation rate and the support level for Mahinda Rajapaksa ($p \ll 0.0001$). Furthermore, that relationship is negative. This means that those divisions that supported Rajapaksa more had a higher proportion of their votes counted.¹¹

Since the relationship is significant *negative*, electoral theory lets us conclude that there is strong evidence of differential invalidation in the 2010 election, *and* that it helped Rajapaksa retain the presidency.

¹¹Did you remember to check if these models violate the assumptions? Checking for overdispersion leads us to use maximum quasi-likelihood as our estimation method. Checking model fit (runs test) tells us that these models are appropriate. **Always check your models.**

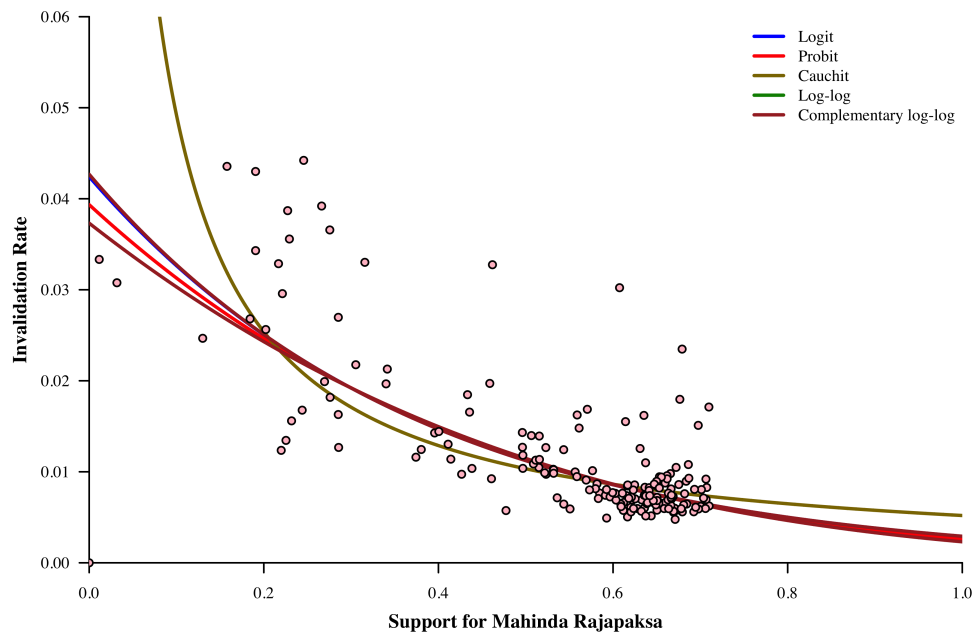


Figure 13.6: An invalidation plot of the Sri Lankan presidential election in 2010 with the entire data set. The curves are prediction curves for the five main link functions. Only the cauchit link (gold) produces estimates that practically differ from the others. Regardless, all models tell the same substantive story: there is evidence of differential invalidation in favor of Mahinda Rajapaksa.

13.5: Beta-Binomial Regression*

There is yet another option for modeling dependent variables that are a count with an upper bound. The Binomial distribution is rather restrictive in that the dispersion parameter must be 1. We were able to relax this requirement by including a new variable to be estimated from the data. All this did, however, was admit that the Binomial distribution is not the correct distribution. Using MQLE allowed us to estimate the standard errors better, but we still clung to the Binomial distribution as the “true” distribution of the data.

It is better to select an entirely new distribution than continue to twist the Binomial into something that it is not. This new distribution must be known, *and* it must allow for non-unit dispersion.

Thankfully, we have such a distribution — the **Beta-Binomial distribution**. Here is its probability mass function:

$$f(y; n, \alpha, \beta) = \binom{n}{y} \frac{B(y + \alpha, n - y + \beta)}{B(\alpha, \beta)} \quad (13.10)$$

Here $B(x, y)$ is the “beta function” defined as

$$B(x, y) = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x + y)} \quad (13.11)$$

Unsurprisingly, the expected value is

$$\mathbb{E}[Y] = n \frac{\alpha}{\alpha + \beta} \quad (13.12)$$

Thankfully, the variance is

$$\mathbb{V}[Y] = n \frac{\alpha \beta (\alpha + \beta + n)}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \quad (13.13)$$

$$= \mathbb{E}[Y] \frac{\beta (\alpha + \beta + n)}{(\alpha + \beta) (\alpha + \beta + 1)} \quad (13.14)$$

Why “thankfully”? Quite simply because dispersion is not an issue. Like the Gaussian, there are enough variables in the probability mass function to estimate it from the data without forcing it to have a specific value (as in the Binomial and Poisson cases).

With all of that good news being said, it is unfortunate that the Beta-Binomial is not EC. Thus, one cannot use the techniques defined by Nelder and Wedderburn (1972). However, one can *still* estimate the parameters of interest using maximum likelihood estimation.

While it is clear that the mathematics of the problem will be more complicated, the coding will be about the same as when we used OLS and GLMs. Compare and contrast the following three lines of code. Determine what each does.

```
|| lm(y ~ x)
|| glm(y ~ x, family=binomial)
|| vglm(y ~ x, family=betabinomial)
```

So, what are the differences? How nice is it that the researcher who created Vector Generalized Linear Models (VGLMs) believed in following good programming practice (Yee 2010 and 2015). Note that the coding is very similar. However, the function requires the [VGAM](#) package.

After running the `vglm` function line, perform the usual summary of the estimated model. Note that this will take awhile. The function is performing many estimation steps in that `summary` call. As a result, there are a few things that are included in the regression output.

To see this, here is the output from the Binomial regression using MQLE.

```
|| Coefficients:
||           Estimate Std. Error t value Pr(>|t|)
|| (Intercept) -6.6636     6.7370  -0.989   0.336
|| averageAge   0.3766     0.3469   1.086   0.292
```

Here is the output from the Bet-Binomial regression (using MLE).

```
|| Coefficients:
||           Estimate Std. Error z value Pr(>|z|)
|| (Intercept):1 -5.8076     5.7313  -1.013   0.311
|| (Intercept):2 -0.3309     0.3335  -0.992   0.321
|| averageAge     0.3233     0.2944   1.098   0.272
```

The effect of the average age is interpreted in the same manner (note that the logit link was used in both cases). The difference in the estimated effects in the two models is slight.

The big difference is that the intercept is estimated using two values in the VGLM. This is because the Beta-Binomial distribution has two parameters to estimate, α and β .

Compare Figure 13.7, which was fit using the Beta-Binomial model, to Figure 13.3, which was fit using the Binomial distribution. Note that the story told by the Beta-Binomial is approximately the same. The only difference I can see is that the Humanities and Social Sciences are not as close together in this model as they were in the Binomial model.

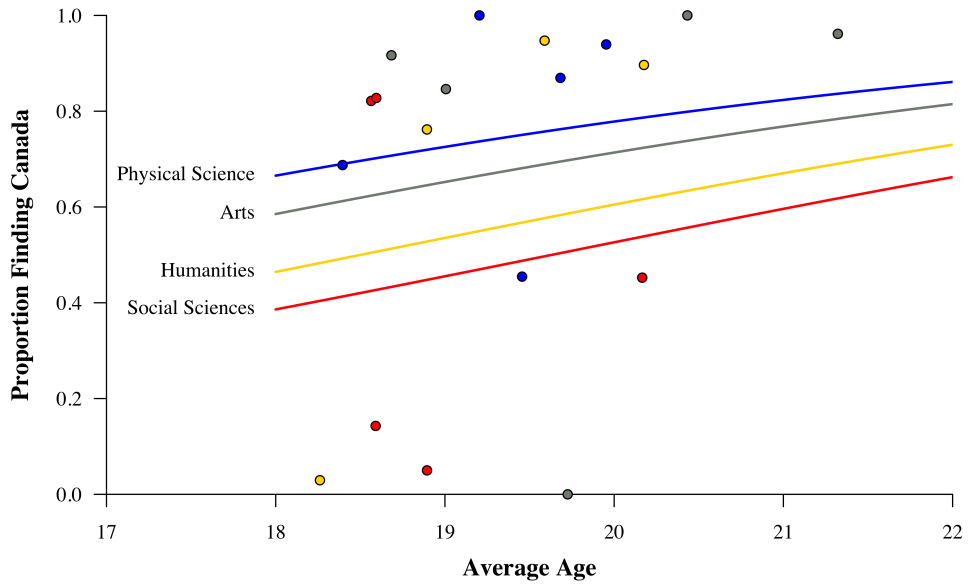


Figure 13.7: The estimated results of modeling the “O Canada” data using the Beta-Binomial distribution. Compare to Figure 13.3, which used the Binomial distribution.

Note: The key reason to choose the Beta-Binomial over the Binomial is that the overdispersion guarantees that the Binomial is *not* the right distribution. Thus, logically, the Beta-Binomial is the better distribution.

13.6: Conclusion

In the previous chapter we examined how to model dichotomous dependent variables. Such variables follow a Bernoulli distribution and should be modeled as such. In this chapter, we looked at modeling random variables that are sums of Bernoulli random variables. As a first approximation, these follow a Binomial distribution.

As a second approximation, we need to recognize that there may be evidence of dependence in the data. Evidence of such is frequently seen in overdispersion. The adjustment is to fit the model using a different estimation procedure: maximum quasi-likelihood estimation.

Once these topics were covered, this chapter provided a (far too-) brief discussion of ethics, a refocus on the measurement unit and what we are allowed to say about the experimental unit, and two extended examples.

13.7: End-of-Chapter Materials

13.7.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

RFS This package does not (yet) exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:

```
source("http://rfs.kvasaheim.com/rfs.R")
```

Epi This package adds several functions and procedures related to epidemiology. As it is not a part of the base installation for R, you will need to install it before you can load it with `library(Epi)`.

VGAM This package allows one to model dependent variables that are conditionally Beta-Binomial distributed. As it is not a part of the base installation for R, you will need to install it before you can load it with `library(VGAM)`.

STATISTICS:

lm(formula) This function performs linear regression on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable.

glm(formula) This function performs generalized linear model estimation on the given formula. There are three additional parameters that can (and often should) be specified.

The `family` parameter specifies the distributional family of the dependent variable, options include `gaussian`, `binomial`, `poisson`, `gamma`, `quasibinomial`, and `quasipoisson`. If this parameter is not specified, R assumes `gaussian`.

The `link` parameter specifies the link function for the distribution. If none is specified, the canonical link is assumed.

Finally, the `data` parameter specifies the data from which the formula variables come. This is the same parameter as in the `lm()` function.

vglm(formula) This function performs vector generalized linear model estimation on the given formula. As with the `glm` function, there are additional parameters that can (and often should) be specified. Note that the `link` option is not allowed in `vglm` models at this point.

predict(mod, newdata) As with almost all statistical packages, R has a `predict` function. It takes two parameters, the model, and a dataframe of the independent

values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

accuracy(mod) This function in the `RFS` package determines the predictive accuracy of a provided model. It takes three necessary parameters: `data`, `truth`, `model`, and `threshold`. It has the optional parameter of returning the *number* of correct classifications (`rate=FALSE`).

AIC(mod) This function calculates the Akaike Informations Criterion score for the provided model. The model needs to have been fit using Maximum Likelihood Estimation.

BIC(mod) This function in the `RFS` package calculates Schwarz's Bayesian Information Criterion (BIC) for the provided model.

deviance(mod) This function returns the deviance in the model. This value is useful in the Likelihood Ratio Test.

deviance.test(mod) This function performs a chi-square test for whether the deviance differs from unity. It is a part of the `RFS` package.

pchisq(x) This gives the value of the cumulative distribution function (CDF) under the Chi-squared distribution. The necessary parameter is the number of degrees of freedom, `df=`. By default, it returns the lower-tail probability. Usually, we will want to have the upper-tail probability, thus we will use the `lower.tail=FALSE` parameter.

var.test(x,y) This function performs an F test, which compares the variances of two samples (`x` and `y`) from Normal populations. It can only compare two samples. If you need to compare more than two samples for equality of variance, you will need to perform either a Bartlett test or a Fligner-Killeen test.

PROGRAMMING:

for This command is one of the basic control-constructs in the R language (as in most programming languages). The usual use is

```
for(var in seq) expr,
```

where `var` is the looping variable (the variable that equals the current loop number). The parameter `seq` is a vector of values. Usually, `seq = something like 1:100`, which is a vector of values from 1 to 100. Finally, `expr` is the expression (or series of expressions) that are performed for each value in the `seq` vector.

13.7.2 EXERCISES This section offers suggestions on things you can practice from this chapter. I suggest that you save the scripts in your Chapter 13 folder.

1. Show that the Binomial distribution is a member of the Exponential Class of distributions.
2. Let $Y \sim \text{Bin}(n, \pi)$. Calculate $\mathbb{E}[Y]$ and $\mathbb{V}[Y]$ using the method of Section 11.2.4.
3. Explain why a negative (and statistically significant) slope indicates that the differential invalidation was in favor of that candidate (page 392).
4. Refit the Sri Lanka election data using the Beta-Binomial distribution. Comment on any differences.

13.7.3 THEORY READINGS

- Alan Agresti and Brent A. Coull. (1998). “Approximate is better than ‘exact’ for interval estimation of binomial proportions.” *The American Statistician*, 52(2): 119–126.
- Peter McCullagh and John A. Nelder. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- John A. Nelder and Robert W. Wedderburn. (1972). “Generalized Linear Models.” *Journal of the Royal Statistical Society Series A (General)* 135(3): 370–84.
- Thomas W. Yee (2010). The VGAM Package for Categorical Data Analysis. *Journal of Statistical Software*, 32(10), 1-34. DOI: 10.18637/jss.v032.i10. URL <https://www.jstatsoft.org/article/view/v032i10/>.
- Thomas W. Yee (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA: Springer.
- Thomas W. Yee and C. J. Wild (1996). Vector Generalized Additive Models. *Journal of Royal Statistical Society, Series B*, 58(3), 481-493.

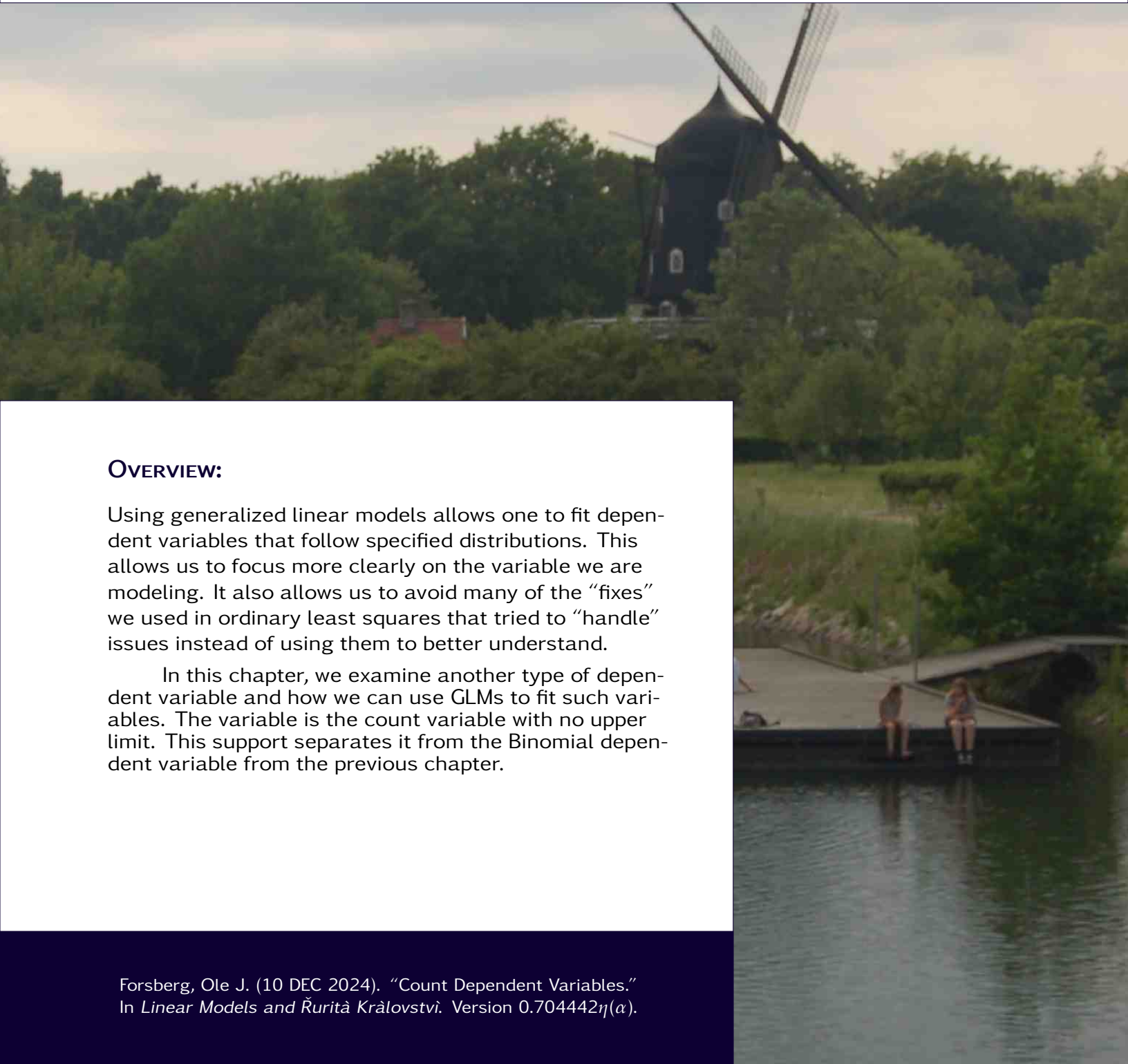
CHAPTER 14:

COUNT DEPENDENT VARIABLES

OVERVIEW:

Using generalized linear models allows one to fit dependent variables that follow specified distributions. This allows us to focus more clearly on the variable we are modeling. It also allows us to avoid many of the “fixes” we used in ordinary least squares that tried to “handle” issues instead of using them to better understand.

In this chapter, we examine another type of dependent variable and how we can use GLMs to fit such variables. The variable is the count variable with no upper limit. This support separates it from the Binomial dependent variable from the previous chapter.



Chapter Contents

14.1	Linear or Poisson Regression?	410
14.2	The Mathematics.	413
14.3	Overdispersion	420
14.4	Full Example: Body Counts	426
14.5	The Bias-Variance Trade-Off.	433
14.6	Conclusion	434
14.7	End of Chapter Materials.	435



Remember that we are examining these many different types of regressions for one primary reason:

The requirements of Ordinary Least Squares are violated by discrete dependent variables.

Rather than seeing this as a problem, we can use it as an indicator that we can model the data better and extract more information from the data.

This marks the next chapter of discrete dependent variables. In Chapter 12, we discussed binary dependent variables — dependent variables that can only take on two values. In the previous chapter, we examined dependent variables that were counts of successes over a known number of trials (or attempts). In this chapter, we examine count dependent variables that have *no upper limit*. Some examples of such count variables include the number of fires in Galesburg in a year, the number of deaths due to terrorist attacks in the world in a month, and the number sorties per day in a battle.



Let us set the stage with an example that we will return to throughout this chapter: The Troubles of Ruritania, a lengthy period of terrorist and counter-terrorist activity in the country, lasted approximately from 1969 until 2002. In that time, over 1800 people died as a result of terrorist actions — by both republican and loyalist groups. Six prime ministers of the Ruritania — on both the political left and right — had to deal with this terrorism. If we assume that the terrorist groups are rational actors, then they will act to maximize their chances of achieving their goals. Because of its hierarchical structure and large size, the Ruritanian Republican Army (RAVŘ) was best able to organize its actions to affect the elections.

The question is whether they did.

Question

Did the RAvŘ adjust its tactics in reaction to the political ideology of the prime minister?

Unfortunately, the extant literature is divided on the direction of the effect. Some research suggests that the RAvŘ became more violent and killed more people when the Conservatives (a.k.a. the Tories) held power. Other research suggests that the RAvŘ became more violent under the Left party. Which is it?



For the unbounded count variables in this chapter, there are three identifying characteristics:

1. the variable can never be negative,
2. it has no theoretic upper bound, and
3. it is discrete.

If Y is this type of count variable, then

$$Y \in \{0, 1, 2, 3, \dots\}$$

If we just do usual linear modeling without taking these three items into consideration, we lose information inherent in the data; we are making assumptions about the data that are incorrect. Performing count data analysis extracts more information from the data you worked so hard to collect. It gives better predictions and explanations of the phenomena under study. It also (usually) means not having to “fix” violations of homoskedasticity or fit.

14.1: Linear or Poisson Regression?

To illustrate some of these observations, let us create a count dataset, fit it with a simple linear model, fit it with a Poisson model, and then compare the results. The data that we will use for this example, `fakepoisson`, was fabricated so that we know the parameters. As such, we can compare the estimates we get from the three modeling techniques to the true parameters. Here is the code I used to create the `fakepoisson` data set:

```
set.seed(370)

n=75
x = sort( runif(n, min=0, max=2) )
beta0 = 0
beta1 = 2
lambda = exp( beta0 + beta1*x )

y = rpois(n, lambda)
```

By this point, you should be able to determine what each line of code does. You should also take note of how the parameter `lambda` is defined and keep this in mind as you read forward.

For this example, the true parameters are $\tilde{\beta}_0 = 0$ and $\tilde{\beta}_1 = 2$. Both of these are in log units (the tildes to serve as reminders of this). Except for those provided for the linear model, it is difficult to compare the estimates of the true value. It is much easier to compare the prediction curves.

OLS MODEL (UNTRANSFORMED): The OLS (untransformed) model can easily be performed. However, it does not fit the data well at all (Figure 14.1). If you decide to perform the three usual numeric tests, you will find all three violated. Yikes!

OLS MODEL (LOG-TRANSFORMED): The transformed OLS model has its own problem. Logically, a log transform would be appropriate (only bounded below). However, the dependent variable takes on a 0 value. This means you should either perform an additional transformation (add 1 to each dependent value) or drop the records with $y = 0$.

If we add 1 to each dependent variable before performing the log transform (that is, we perform the transformation $y^* = \log[y + 1]$.) We see that there is a lingering issue with heteroskedasticity. Is it ignorable? Perhaps, but let us try Poisson regression.

POISSON MODEL: Here, we easily perform Poisson regression and check the requirements. No requirement is violated. As such, we are good to go with this model.

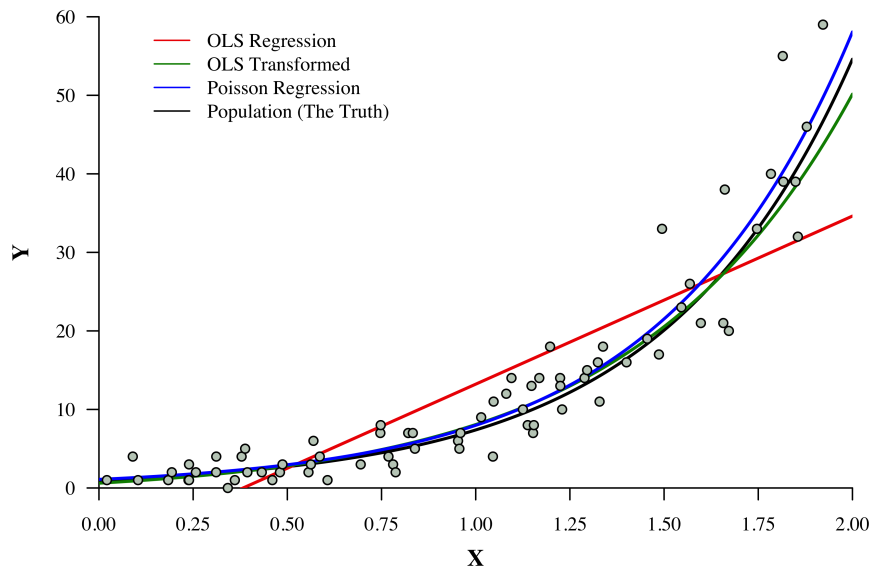


Figure 14.1: Plot of the pseudo data with three regression equations overlaying. The linear regression is in red, the linear regression on the log-transformed data is in green, and the Poisson regression is in blue. The black curve is the “correct” curve.

What was the code?

```
||| model.c <- glm(y ~ x, family=poisson)
||| summary(model.c)
```

That’s all.

At this point, you should be able to guess what the lines do and what information in the first line is missing because it is the default link... as we see shortly.

WHICH IS THE BEST MODEL OF THE THREE?: It may not be clear that these three cannot be directly compared. The linear model makes no adjustments, the log-transformed model does, as does the Poisson regression model. None of the transformations are the same. This means that we cannot use information criteria to compare the models.¹

So, how can we determine which of these models is best? A first step is to determine which is appropriate by checking the assumptions. The untransformed OLS model violates all three requirements. The transformed OLS model is heteroskedastic. The Poisson model, however, violates none of the requirements. Thus, on the basis of meeting assumptions, the third model is the best of these three.

If all we care about is estimates (and not confidence intervals), we could look at the graphic comparing the data and the estimations from the model (Figure 14.1). Numerically, we could also check how much the uncertainty in y has changed. The uncertainty using the null model (predicting $y = \bar{y}$) is 184.96. The uncertainty with the linear model is 43.6 — a reduction of 76%. The log-transformed linear model has an uncertainty of 8.1, which is a reduction of 96%. This is quite much different from the pure linear model. The Poisson model also has an uncertainty of 8.1 and a total reduction of 96%.

Thus, if all you care about is the estimate (which scientists should *not*), finding some adjustment so that the curve fits the data works. If you are a true scientist, then the confidence interval (and p-values) are important. This means assumptions about homoskedasticity are important — if they exist. Some modeling requires homoskedasticity others do not. Poisson regression does not (at least, not *really*).

¹Remember that we can use AIC, BIC, etc. *only when* the y-values are the same. This is not true here, as the y-values are all transformed differently.

14.2: The Mathematics

Count models have dependent variables that can take on only non-negative integers. Back in the time of OLS, we handled the non-negative aspect by taking the logarithm of the dependent variable (perhaps by adding 1 before taking the logarithm if there are values of 0). However, OLS does not allow for discrete dependent variables. The discrete aspect must be handled through Generalized Linear Models (GLMs).

Recall that using GLMs requires that we explicitly specify three things.

1. First, we need to know the linear predictor, η .
2. Second, we need to know the distribution of the dependent variable, conditioned on the independent variables.
3. Finally, we need to know the link function that appropriately connects the two of them.

The linear predictor is the same as always: the weighted sum of our independent variables. The canonical link function is the logarithm function. Finally, the distribution we will use is the Poisson distribution.

The Poisson is *not* the only option for such count dependent variables. The Negative Binomial distribution can also be used, but as the Negative Binomial distribution is a bit more complicated than the Poisson, we will motivate this chapter with the Poisson and save the Negative Binomial for later (Section 14.3.3).

14.2.1 THE POISSON DISTRIBUTION The Poisson distribution has the following probability mass function (pmf):

$$f(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad y \in \{0, 1, 2, 3, \dots\} \quad (14.1)$$

Again, the probability mass function (pmf) is not as important as the expected value of this distribution. Why? Remember that the Generalized Linear Model paradigm models the *expected value*, $\mathbb{E}[Y | X]$, not the actual outcomes.

Calculating the expected value of the Poisson distribution is not as easy as it was for the Binomial; it requires a trick:

$$\mathbb{E}[Y] := \sum_{y=0}^{\infty} y f(y) \quad (14.2)$$

$$= \sum_{y=0}^{\infty} y \frac{e^{-\lambda} \lambda^y}{y!} \quad (14.3)$$

$$= \sum_{y=1}^{\infty} y \frac{e^{-\lambda} \lambda^y}{y!} \quad (14.4)$$

$$= \sum_{y=1}^{\infty} \frac{e^{-\lambda} \lambda^y}{(y-1)!} \quad (14.5)$$

$$= \lambda \sum_{y=1}^{\infty} \frac{e^{-\lambda} \lambda^{(y-1)}}{(y-1)!} \quad (14.6)$$

Let us define $z := y - 1$:

$$= \lambda \sum_{z=0}^{\infty} \frac{e^{-\lambda} \lambda^z}{z!} \quad (14.7)$$

and so, we have

$$\mathbb{E}[Y] = \lambda \quad (14.8)$$

This last step is correct as $e^{-\lambda} \lambda^z / z!$ is the probability mass function for the Poisson, therefore

$$\sum_{z=0}^{\infty} \frac{e^{-\lambda} \lambda^z}{z!} = 1 \quad (14.9)$$

Thus, the expected value of a Poisson random variable is $\mathbb{E}[Y] = \lambda$.

Note: Recall that one of the assumptions of Ordinary Least Squares is that the variance is constant with respect to the (expected value of the) dependent variable. When the outcomes are distributed as Poisson random variables, we can actually prove that the variance is *not* constant with respect to the predicted outcomes. To see this, let $Y \sim \mathcal{P}(\lambda)$. With this, and with the probability mass function above, we can use the definition to calculate the variance of Y . Without proof, the variance of Y is $\mathbb{V}[Y] = \lambda$. Yes, the variance is *the same as* the expected value.

Thus, the variance is a function of the expected value, and the homoskedasticity requirement of OLS is violated.

Note: That the variance is a function of the expected value also creates a problem. Quite often, we will be dealing with data in which the variance

is *not* equal to, but is greater than, the expected value. Such data is termed *overdispersed*. When we encounter it (Section 14.3), we will discuss what it means and what we should do.

By the way, this is the same overdispersion as we discussed in Section 13.3.3. It is caused by the same structures and is tested in the same manner. I encourage you to revisit that section (page 384).

Now that we understand our choice of distribution a bit better, and the resulting expected value, let us examine the third facet: the link function. First, note that λ is bounded; $\lambda \in (0, \infty)$. Thus, we need a function that takes a bounded variable and transforms it into an unbounded variable. We have already met a link function that can handle this — the logarithm function (see Chapter 7).

Note: Again, note that we are modeling $\lambda = \mathbb{E}[Y]$, not the observed count. As λ is continuous and bounded below by zero (but never equal to zero), we can use the logarithm function as our transformation link.

And so, we have the three necessary components to use Generalized Linear Models for count data:

- the linear predictor,

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (14.10)$$

- the conditional distribution of the dependent variable,

$$Y | x \sim \mathcal{P}(\lambda) \quad (14.11)$$

with the formula for the expected value,

$$\mu = \lambda \quad (14.12)$$

- and the link function,

$$\log(\mu) = \eta \quad (14.13)$$

Note: Here is what you need to take away from this section: The distribution must fit the possible outcomes. The link must translate the bounds on the parameter to the lack of bounds on the linear predictor. Both require you to know some distributions.

14.2.2 DERIVING THE CANONICAL LINK In Chapter 11, we mentioned that each distribution has a canonical link. Let us derive the canonical link for the Poisson distribution. The steps to determine the canonical link are the same for the Poisson as it was for the Gaussian (Chapter 11), Bernoulli (Chapter 12), and Binomial (Chapter 13):

1. Write the probability mass function (pmf).
2. Write the probability mass function in the required form.
3. Read off the canonical link.

For this distribution, this results in:

$$\text{pmf: } f(y | \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (14.14)$$

$$= \exp[\log(e^{-\lambda}) + \log(\lambda^y) - \log(y!)] \quad (14.15)$$

$$= \exp[-\lambda + y \log(\lambda) - \log(y!)] \quad (14.16)$$

$$= \exp\left[\frac{y \log(\lambda) - \lambda}{1} + -\log(y!)\right] \quad (14.17)$$

This is in the required form:

$$\exp\left[\frac{y \theta - b(\theta)}{a(\phi)} + c(y, \theta)\right] \quad (14.18)$$

$$\exp\left[\frac{y \theta - b(\theta)}{a(\phi)} + c(y, \theta)\right] \quad (14.19)$$

$$(14.20)$$

Thus, reading off the standard form, we have the following:

- $y = y$
- $\theta = \log(\lambda)$
- $a(\phi) = 1$
- $b(\theta) = \lambda = \exp(\theta)$
- $c(y, \theta) = -\log(y!)$

As such, the canonical link is the log function. I leave it as an exercise to show that $\mathbb{E}[Y] = \lambda$ and $\mathbb{V}[Y] = \lambda$ using the methods of Section 11.2.4.

Example 1

The people in many US states have the ability to formulate binding laws by placing them before the people for a vote. This process is called the Citizens' Initiative. Extant theory suggests that states with a higher population will also use the initiative process more often than states with a lower population. Let us test this hypothesis with data (`crime` datafile).

Solution: As we are performing GLM modeling, we need to determine the three needed components. First, since the dependent variable is a *count* of the number of initiatives placed before the voters, we will assume that the dependent variable has a Poisson distribution:

$$\text{inituse} \mid \lambda \sim \mathcal{P}(\lambda) \quad (14.21)$$

The linear predictor will use our explanatory variable:

$$\eta = \beta_0 + \beta_1 \text{pop90} \quad (14.22)$$

The link function will be the logarithm function:

$$\log(\lambda) = \eta \quad (14.23)$$

With this, we use these commands to load and analyze the data

```
|| vcr = read.csv("http://rur.kvasaheim.com/data/crime.csv")
||
|| m2 = glm( inituse ~ pop90, family=poisson(link=log),
||         data=vcr, subset=(ccode!=93) )
```

Now, `summary(m2)` tells us that there is a statistically significant relationship between the state's population in 1990 and its use of initiatives in the 1990s. Unfortunately, the relationship is negative ($\hat{\beta}_1 = -7.433 \times 10^{-8}$), which is definitely *inconsistent* with the original hypothesis. We have shown that the original hypothesis does not agree with this reality.

Let us now predict the number of initiatives that Utah would have had during the 1990s using the fact that the population of Utah is 1,722,850. We can do this by hand or we can use the `predict` function. In either case, we must remember to back-transform using the inverse of the logarithm function, the exponential function. Using the latter method gives me an *un-transformed* prediction of 2.0, which means the model predicts 7.44 initiatives for Utah in the 1990s. The real value is 3.

```
|| UTAH = data.frame(pop90=1722850)
```

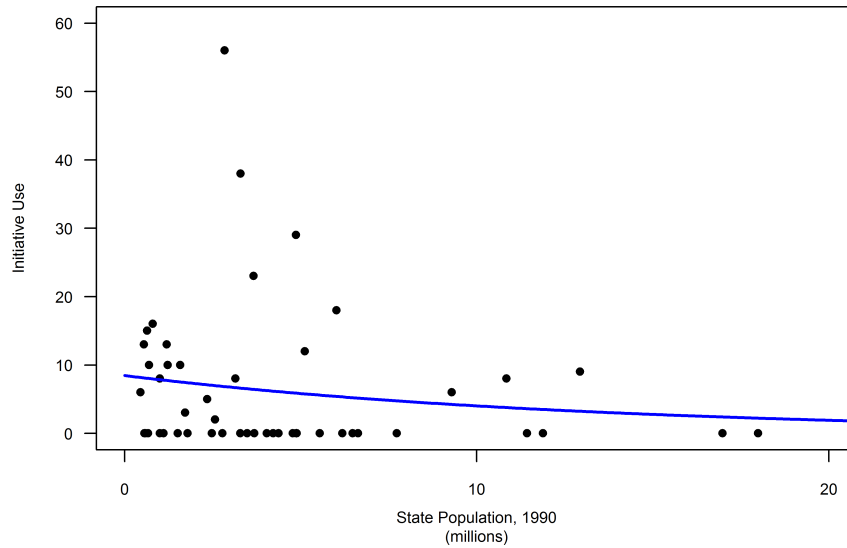


Figure 14.2: A plot of initiative use against the population of the state in 1990, with the Poisson regression curve superimposed.

```
prL = predict(m2, newdata=UTAH)
exp(prL)
```

Note: The `glm` function used here includes an additional parameter that we have not discussed: `subset`. This parameter allows us to explicitly specify which data to include in the analysis. Here, I removed the state with `cocode` equal to 93 (California) from the analysis. The reason I did this is that a plot of the entire data suggested that California was an influential point.

Figure 14.2 is a plot of the data, with the regression curve superimposed. The interesting thing is that the graph visually calls into question the results of the GLM regression above. While the effect direction does definitely appear to be negative, it is hard to believe that this effect has such a high level of significance ($p \ll 0.0001$). There is a lot of variance in the data. What is happening? ♦

The problem is that the model/data are *overdispersed*.

14.3: Overdispersion

Recall that one result of using the Poisson as our distribution of choice is that the residual variance and the expected value are assumed equal, because the probability mass function has $a(\phi) = 1$. As in the case of the Binomial models, overdispersion means that $a(\phi) > 1$.

In essence, this means $\mathbb{V}[Y | X] > \mathbb{E}[Y | X]$. For a Poisson model (and for a Binomial model), the overdispersion measure equals the ratio of the residual deviance to the residual degrees of freedom. We use the ‘residual,’ since overdispersion is a function of the model, as well as the data. For the Initiative Use model (Example 14.2.2), the overdispersion factor is $681.68/48 = 14.2$. In other words, the level of unexplained variance is 14.2 times too high for this model.

Note: Please revisit Section 13.3.3 for one way of *testing* whether 681.68 really is evidence of significant overdispersion. If you create the right test,² you will get a critical value of just 69. Since $681.86 \gg 69$, there is a lot of overdispersion in this data.

Since overdispersion is a function of the model you are fitting to your data, the first solution is to determine if you are missing some important variables (or powers of variables). Frequently, modifying your linear predictor by adding appropriate variables will reduce the overdispersion.

Even though this is the most appropriate method in many ways, there is an extreme danger to using this method: you may need to include *too many* variables and combinations of variables to eliminate the overdispersion. This results in overfitting the data; that is, you are fitting the data and not the data-generating process in which we are actually interested (see Section 14.5).

Thus, if you end up including too many variables before the overdispersion is treated, you may want to consider other options.

The first is to adjust the standard errors by hand. This frequently works acceptably, as the primary effect of overdispersion is to underestimate the standard errors. The second option is to fit the model using a different fitting technique, one that allows you to use the Poisson but also allows you to have a different relationship between the mean and variance. Maximum quasi-likelihood estimation (MQLE) is a common alternative to the usual maximum likelihood methods. Finally, you can fit the model using a different distribution, one that does not require the mean to equal the variance. The Negative Binomial is a common alternative to the Poisson.

²The code I used was `qchisq(c(0.025,0.975), df=48)`.

	Parameter Estimate	Original Std. Error	Adjusted		
			Std. Error	z-value	p-value
Intercept	2.136	0.0814	0.307	6.960	$\ll 0.0001$
1990 Population	-7.433×10^{-8}	1.743×10^{-8}	6.568×10^{-8}	-1.131	0.2578

Table 14.1: The results from the Poisson model, with standard errors adjusted for overdispersion. In the original Poisson model, the residual deviance was 681.68 and the residual degrees of freedom was 48. Thus, the dispersion factor was 14.202. Thus, we adjust the standard errors by multiplying the original estimates by $\sqrt{14.202} = 3.769$. The calculation for the z value is the same: coefficient divided by standard error, with the new p-value based on the adjusted z value.

14.3.1 ADJUSTING THE STANDARD ERRORS This first option adjusts the estimated standard errors to try to compensate for the overdispersion. Recall that the dispersion factor is the ratio of the residual variance to the expected variance. As the standard error is the square root of a variance, it would make sense that we could ‘fix’ the overdispersion by *multiplying* the standard errors by the square root of the dispersion factor.

Table 14.1 presents the original standard error estimate along with the adjusted standard errors, z-values, and p-values. Note that the 1990 population was highly significant in the unadjusted model, but is not significant in the adjusted model ($p = 0.2578$).

The strength of this method is that it is easily performed. The drawback is that the correction is only an approximate estimate. In the era of expensive computational times, this method was commonly used; in the modern era of cheap computing, not-so-much. The next two methods are more appropriate in that their results are more statistically sound than this approximation.

	Parameter	Original	Adjusted		
	Estimate	Std. Error	Std. Error	z-value	p-value
Intercept	2.136	0.0814	0.345	6.103	$\ll 0.0001$
1990 Population	-7.433×10^{-8}	1.743×10^{-8}	7.492×10^{-8}	-0.992	0.3261

Table 14.2: *The results from the Poisson model, with standard errors estimated using Quasi-Likelihood Estimation. Note that the coefficient estimates are the same between the two methods. The differences are due to the re-estimated standard errors. Note, again, that the 1990 population is no longer a statistically significant variable as it was in the original Poisson model.*

14.3.2 MAXIMUM QUASI-LIKELIHOOD ESTIMATION The maximum likelihood estimation method makes assumptions about the relationship between the mean and variance of the underlying distribution. For the Poisson distribution, that relationship is the identity function; that is, $\mathbb{E}[Y | X] = \mathbb{V}[Y | X]$. The presence of overdispersion indicates that this relationship — or this value of $a(\phi)$ — is incorrect.

A different way of estimating the parameters is to use Maximum Quasi-Likelihood Estimation (Section 13.3.3, page 385). This method allows for modeling different relationships between the expected value and variance for the distribution. It effectively includes an additional parameter for $a(\phi)$.

The strength of using MQLE is that you can use the same distributions with which we are familiar, and the interpretation is identical. The weakness is that some statistical programs are not able to model using this method. Thankfully, R can. To model using MQLE in R, we prefix the distribution with the word `quasi`. Thus, we would use

```
|| glm(y ~ x, family=quasipoisson(link=log))
```

to fit this model. This command produces the results in Table 14.2. Note that the coefficient estimates are the same as for the Poisson model. The difference is in the standard errors — they are increased. This reduction causes our z-values to decrease, resulting in increased p-values.

Note: The only two distributions that have the MQLE option in \mathbb{R} are the Poisson (`quasipoisson`) and the Binomial (`quasiBinomial`). These are the only two popular distributions that have a specific numeric value for $a(\phi) = 1$. The rest have a value for $a(\phi)$ that can be estimated from the data; for instance, the Gaussian has $a(\phi) = \sigma^2$.

14.3.3 THE NEGATIVE BINOMIAL FAMILY In the Generalized Linear Model framework, you need to select an appropriate distribution that matches your dependent variables. If that variable is a count, then the sole requirements for that distribution are that the outcomes can only be discrete and non-negative. The Poisson is the usual distribution, but it is not the only one. An alternative distribution is the Negative Binomial. The Negative Binomial family allows for both over- and under-dispersion in the model. It does this by assuming the rate parameter λ in the Poisson is distributed as a Gamma random variable (Venables and Ripley 2004). Specifically, it assumes

$$Y \mid \mu, \theta \sim \text{NegBin}(\mu, \theta) \quad (14.24)$$

where

$$Y \mid W \sim \mathcal{P}(\mu W) \quad (14.25)$$

with

$$W \sim \frac{1}{\theta} \text{GAM}(\theta) \quad (14.26)$$

where $\mathbb{E}[W] = 1$ and $\mathbb{V}[W] = 1/\theta$.

an

	Estimate	Std. Error	z-value	p-value
Constant Term	2.2376	0.4835	4.63	$\ll 0.0001$
Population in 1990	-1.0091×10^{-7}	8.1903×10^{-8}	-1.23	0.2179

Table 14.3: The results table for modeling the initiative use using the Negative Binomial distribution. Note that the population is no longer statistically significant.

exercise

With this formulation, it can be shown that $\mathbb{E}[Y] = \mu$, $\mathbb{V}[Y] = \mu + \mu^2/\theta$, and that the probability mass function for Y is

$$f(y; \theta) = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{\theta+y}} \quad (14.27)$$

The strength of this formulation is that a greater number of variations are able to be fit.

The drawback is that interpreting the results is a bit more difficult. However, since we make the computer do all the heavy lifting, this drawback is minor. It does, however, introduce a new set of possible error messages and parameters that you may have to interpret.

The other drawback is that the Negative Binomial distribution is *not* a member of the exponential family (unless θ is known, which it is not). As such, it cannot be used within the GLM paradigm (strictly speaking). With that said, fitting a model using the Negative Binomial distribution is just as easy as it is for any of the previous distributions.

In R, you will have to load the `MASS` package to use the Negative Binomial family, since it has its own regression function: `glm.nb`. The options for `glm.nb` are similar to those for `glm` — Venables and Ripley designed it that way (or had their grad students design it that way). Thus, the command

```
|| m2n = glm.nb(inituse ~ pop90, data=vcr, subset=(ccode!=93))
```

will perform Negative Binomial regression similar to the regression performed in Section 14.3.2. The first thing to notice is that the overdispersion is no longer relevant. With this, we can have more confidence in the parameter estimates (provided in Table 14.3). The second thing to notice is that the effect of population is still no longer statistically significant. This agrees with our observation in Sections 14.3.1 and 14.3.2. Finally, we notice that there are additional parameters estimated (at the bottom). The `Theta` is the estimated value of θ in the Gamma distribution above.

Note: The *direction* of the coefficient estimate is still directly comparable to the other coefficients estimates we have examined. The magnitudes are also comparable, but only to the other log-linked models. Thus, this model tells

us that there is a negative relationship between the state's population and the level of initiative use (although it is not statistically significant).³

This model estimates that Utah will have had approximately 7.9 initiatives during the 1990s. I leave it as an exercise to determine this.

exercise

³That the direction is comparable is due to choosing a link function that is strictly *increasing*. That the estimates are comparable is due to having the same link function or same transform function.

14.4: Full Example: Body Counts

Using the above information, let us examine the problem of understanding terrorism. This extended example will also allow us to discuss a few things that are becoming important to our analyses, namely the bias-variance trade-off.

Example 2

The Troubles in Ruritania lasted from 1969 until 2002. In that time, over 1800 people died as a result of terrorist actions — both republican and loyalist groups. Six prime ministers — both left- and right-leaning — had to deal with the terrorism. If we assume that the terrorist groups are rational actors, then they will act to maximize their chances of achieving their goals. Because of its hierarchical structure and large size, the Ruritanian Republican Army (RAvŘ) was best able to organize its actions to affect the elections.

The question is whether they did —

Did the RAvŘ react to the political ideology of the prime minister?

Unfortunately, the extant literature is divided on the direction of the effect. Some research suggests that the RAvŘ became more violent and killed more people when the Conservatives held power. Other research suggests that the RAvŘ became more violent under the Left party. Which is it?

The dataset, `terrorism`, contains just three variables of import: `total` (the total number of deaths under that prime minister for the year, or part of the year), `days` (the number of days during the year that the prime minister was in power), and `riteleft` (the level of conservatism of the prime minister). The second variable is necessary to control for the fact that some prime ministers only ruled for a part of the year. The third variable is the research variable. The first variable is the response variable (dependent variable). The basic research model is

$$\text{deaths} \sim \text{riteleft} \quad (14.28)$$

However, we need to deal with `days`, the number of days the premier is in power. If we include `days` as a simple independent variable, we allow the effects of the `days` variable to freely vary to fit the data.

It is almost always better to treat `days` as the divisor for terrorist killings, thus ostensibly creating a variable of `killings per day`. But, this is no longer a count model (non-integer values), nor is it a proportion model (values can be greater than one). What should we do?

Fear not! Through the magic of mathematics, we can handle it.

Recall in Section 14.2 that the link function we used was the logarithm: $\log[\lambda] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$. If, instead of the expected count, λ , we wanted to model the expected ratio, λ/days , we would have:

$$\log\left[\frac{\lambda}{\text{days}}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (14.29)$$

Using one of the properties of logarithms, this is equal to

$$\log[\lambda] - \log[\text{days}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (14.30)$$

This, in turn, is mathematically equivalent to

$$\log[\lambda] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \log[\text{days}] \quad (14.31)$$

As such, we now have a count model (the $\log[\lambda]$ is alone on the left and is a random variable) with an additional factor ($\log[\text{days}]$ on the right as a non-random variable). Note that *there is no parameter* to estimate for $\log[\text{days}]$. This is important in how we set up the model, as `days` is not a typical variable. Let us call it an *offset* variable.

Offset variables do not have parameters to estimate. They are direct effects with no multipliers. One can think of them as being subsumed in the constant term (which would be true if the offset variable were constant). Most statistical programs have an offset option available when you specify the model to be fit. In \mathbb{R} , the offset is specified in the model call by the keyword `'offset'`.

	Poisson	Quasi-Poisson	Negative Binomial
Constant Term	2.2622 (0.1190)	2.2622 (0.7071)	2.0363 (0.6004)
Conservatism	-0.0115 (0.0011)	-0.0115 (0.0065)	-0.0142 (0.0093)
Days Served	0.0050 (0.0004)	0.0050 (0.0021)	0.0058 (0.0019)
AIC	1482.8	—	369.5

Table 14.4: Results of three different “families:” Poisson, quasi-Poisson, and Negative Binomial. The numbers in parentheses, below the coefficient estimates, are the standard errors.

For the `glm` function,

```
|| glm(pira ~ riteleft, offset=log(days), data=terror)
```

For `glm.nb`,

```
|| glm.nb(pira ~ riteleft, offset(log(days)), data=terror)
```

specific

OPTION 1: DAYS AS AN INDEPENDENT VARIABLE: The first option is to treat the `days` variable as just another independent variable. This is not the best answer, as `days` has a specific meaning with respect to the number of terrorist deaths. The better option is to use Option 2 (below). However, for pedagogical purposes, let us first enter `days` as an independent variable. Performing regressions for each of the three count data families, we get the summarized results in Table 14.4.

Note that the direction of each of the effects is the same. This is not always true, especially when the variable has little effect or has no statistical significance. However, if the variable is significant *and* changes effect direction, then there is something severely wrong with your research model. Also note that the effects are the same between the Poisson and the quasi-Poisson families. The only difference is the size of the standard errors. The quasi-Poisson will *always* give a better estimate of the standard errors (and of the statistical significance) than the Poisson.

Note that the Poisson model is severely overdispersed — the residual deviance is much larger than the residual degrees of freedom (the residual deviance is 1298, the residual degrees of freedom is 36, the overdispersion factor is 36.06). As such, the Poisson family would be (very) inappropriate for this model. Thus, either the quasi-Poisson or the Negative Binomial model would be preferable.

If we had just used the Poisson family, we would have concluded that the level of conservatism of the prime minister is *highly* significant. However, looking

	Poisson	Quasi-Poisson	Negative Binomial
Constant Term	-1.8280 (0.0254)	-1.8280 (0.1495)	3.8744 (0.0969)
Conservatism	-0.0106 (0.0011)	-0.0106 (0.0063)	-0.0069 (0.0041)
AIC	1479.6	—	2080.2

Table 14.5: Results of three different families: Poisson, quasi-Poisson, and Negative Binomial. The numbers represent the estimated coefficients. The numbers in parentheses are the estimated standard errors.

at the more-appropriate results of fitting using Maximum Quasi-Likelihood Estimation (or using the Negative Binomial family), we see that the effect of conservatism is non-existent. Since the effect of conservatism on deaths was the purpose of this research question, it is extremely important to reach good conclusions about the effects of this variable.

As our research variable is not statistically significant at the usual level of significance, we will not even bother to predict and graph our predictions here.

OPTION 2: DAYS AS AN OFFSET VARIABLE: The second (and preferred) option uses `days` as an offset (or “exposure”) variable. This makes more sense than allowing it to freely enter the model as a typical independent variable. The results from fitting the data with the three model families are found in Table 14.5.

exposure

According to the results, the Poisson family is not appropriate; the level of overdispersion is very high — on the order of 35. As such, using the MQLE method or the Negative Binomial family would make good substitutes. In the quasipoisson model, the parameter estimates remain the same, but the estimates of the standard errors change to reflect the overdispersion. Thus, while the effect of conservatism was statistically significant in the Poisson model, it was not in the quasipoisson model ($p = 0.1013$).

The Negative Binomial model echoes the qualitative conclusions of the quasipoisson: The level of conservatism has no statistically discernible effect on the level of deaths resulting from RAVR terrorism in Ruritania during the Troubles ($p = 0.0905$).

14.4.1 BETTERING THE FIT* Using the results from both the quasi-Poisson and the Negative Binomial model does offer you the ability to strengthen your conclusions. If one result gave statistical significance and the other did not, then you would realize that your conclusions depended on the assumptions you made about the underlying mechanism that produced the data, and not on the variables you chose to

include (or exclude). It is never a good place to find yourself when your substantive results depend on the choice between two acceptable models.⁴

extreme

Maybe, one should not stop here. Our formula is rather simplistic: it states that one independent variable is all we need to explain the dependent variable. It also assumes that the effect is linear between the independent and the dependent variable. If we believe that *extremist* prime ministers suffer from higher (or lower) levels of terrorist killings, then the research formula we have cannot capture that effect. To capture *that* effect, we will have to use the square (and/or higher powers) of the `riteleft` variable.

In fact, let us examine the effects of conservatism (up to the fourth power), *plus* the effects of having Labour in power, *plus* an interaction between having Labour in power and the level of conservatism in the Labour government. Thus, the research model we wish to fit will be

$$\text{pira} = \beta_0 + \beta_1 \text{riteleft} + \beta_2 \text{riteleft}^2 + \beta_3 \text{riteleft}^3 \quad (14.32)$$

$$+ \beta_4 \text{riteleft}^4 + \beta_5 \text{labour} \quad (14.33)$$

$$+ \beta_6 \text{labour} \times \text{riteleft} \quad (14.34)$$

theory!!

Of course, we would need to have good theory to provide this model, but let's just have fun with this.

In most statistical programs, one would have to create new variables for each of the powers (three new variables) and a new variable for the interaction term (`labour × riteleft`). In R, however, we can just write the formula to reflect what we want without having to worry about the additional step of creating new variables. As such, in R, the formula will be

$$\begin{aligned} \text{pira} \sim & \text{riteleft} + \text{labour} + \text{I(riteleft}\wedge\text{2)} + \\ & \text{I(riteleft}\wedge\text{3)} + \text{I(riteleft}\wedge\text{4)} + \\ & \text{I(labour}\ast\text{riteleft)} \end{aligned} \quad (14.35)$$

The use of `I()` indicates that R should evaluate what is in the parentheses as a new variable. Fitting this model using Maximum Quasi-likelihood Estimation indicates that none of the terms have a statistically significant effect. This should not really surprise us, since there is a lot of correlation among the independent variables in that model. In the presence of high correlation, the standard errors tend to be larger than they should be.

Since nothing was statistically significant, let us pare the model to reduce the effect of correlation and get at some more basic effects. The best first thing to remove

⁴With this said, there is some research into combining estimates from separate models. These estimates require that you are able to specify your personal beliefs in the correctness of the models.

	Quasi-Poisson	Negative Binomial
Intercept	-12.51 (4.478)	-6.980 (2.396)
Labour	-4.742 (1.553)	-4.8430 (0.2856)
Conservatism	1.847 (0.07101)	1.8660 (0.3778)
Conservatism ²	-0.03830 (0.01425)	-0.03833 (0.00750)
Conservatism ³	-0.002585 (0.0009421)	-0.0026070 (0.0005005)
Conservatism ⁴	-0.00007314 (0.00002642)	-0.00007361 (0.00001398)

Table 14.6: Results of two different models: fitting with MQLE and using the Negative Binomial family. The numbers are the parameter estimates; in parentheses, the estimated standard errors.

from the model is the interaction term. Doing this gives us the research model:

$$\begin{aligned} \text{pira} = & \beta_0 + \beta_1 \text{riteleft} + \beta_2 \text{riteleft}^2 + \beta_3 \text{riteleft}^3 \\ & + \beta_4 \text{riteleft}^4 + \beta_5 \text{labour} + \varepsilon \end{aligned} \quad (14.36)$$

Fitting this model using both the quasi-Poisson family and the Negative Binomial family gives us the results in Table 14.6.

Notice that all of our variables are now statistically significant at the $\alpha = 0.05$ level. It turns out that the interaction term was so highly correlated with the other variables that it made it impossible to correctly estimate the effects of the individual research variables.

Now that we have two models that tell us, substantively, the same story, we should *show* the effect of the variables of interest. There are really only two independent variables involved here, with one being dichotomous. As such, we can show the effects on the same graph (one graph for each family), with two prediction curves per graph. Figure 14.3 shows the predictions from both the quasi-Poisson model (Left Panel) and the Negative Binomial model (Right Panel). The upper curve in both cases (red) corresponds to predictions when the Conservatives are in power.

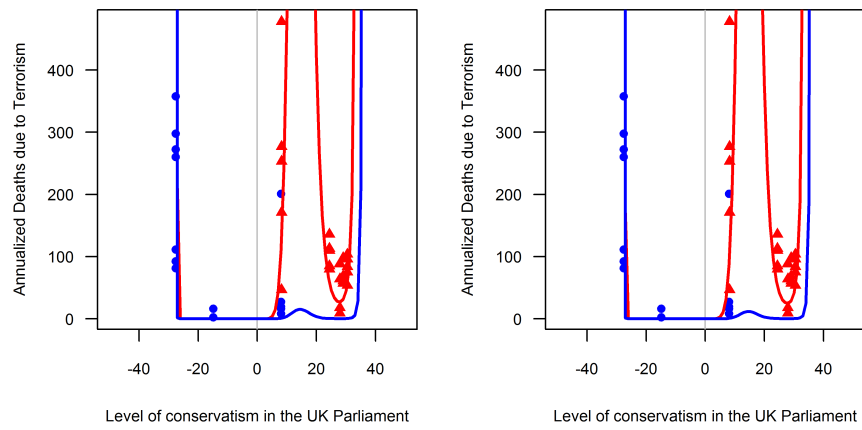


Figure 14.3: Plot of the number of deaths due to terrorism, caused by the Ruritanian Republican Army (RAvŘ), in Ruritania during the Troubles. The points are overlaid with the Troubles. The points are overlaid with the quasi-Poisson model (Left Panel) and the Negative Binomial model (Right Panel). In both cases, the upper curve (red) corresponds to the prediction when the Conservative Party is in power.

14.5: The Bias-Variance Trade-Off

Note that these two models are completely worthless in *explaining* the effects of the variables on the population (or the “data generating process”).⁵ Because we used so many parameters, the model fits the data — noise and all — as opposed to the underlying reality (signal). This is a common problem. Since the goodness of our fit *increases* as we increase the number of variables in our models (see the effect of the number of covariates on the R^2 value), there is a pressure for us to increase the number of variables. However, as in this case, using too many variables (or interactions, or powers) usually tells us too little about the underlying *process* that gave rise to the data, which is the entire purpose of performing a statistical analysis.

Note: Remember that we are only using the data (a sample) to help us better understand the process (model) that gave us the data (population). Fitting the data perfectly actually tells us little about the process we are trying to model. However, not using enough variables may not get at the process, either. This trade-off between increasing the number of variables (which increases the reliance of the parameter estimates on the actual data) and reducing the number of variables (which increases the errors in our model) is termed the Bias-Variance trade-off, and it is a problem we must keep in our minds at all times. On the one hand, we want a good model that fits the population, on the other hand, we only know the sample (the data collected).

In the terrorism example (*v.s.*, Section 14.4), we can see that we used too many explanatory variables in our model. A glance at the graphs in Figure 14.3 suggests that we should have gone with a quadratic model (second power) *at most*, even though the quartic model (fourth power) fit the data better. Avoiding over-fitting the data is as simple as being aware of the dataset and the model predictions (of course, a good graph helps).

⁵Explanation of the relationships is *very* important. Our job as scientists is to use numerical relationships to better understand the data-generating model (how the dependent variable came to being).

14.6: Conclusion

In this chapter, we examined what we can do when our dependent variable is an unbounded count variable. As such variables are non-negative and discrete, nothing we have done thus far can properly handle them. While performing a log transform of the dependent variable as we did in Chapter 7 would allow us to actually make predictions that made sense (provided that there were no zero counts), the resultant model would probably violate one or more of the assumptions of the Classical Linear Model.

Two model families were introduced to handle count data. The Poisson family requires that the mean and the variance be equal (which translates to the residual deviance and the residual degrees of freedom be equal). This is rarely the case. When the residual variance is much larger than the mean, the data are overdispersed. The Negative Binomial family models overdispersed (and underdispersed) data, but it is a bit more difficult to fit with data.

As with Generalized Linear Models in general, the methods in this section model the expected value and not the actual outcome. As the parameters must be non-negative, we use a log link to ensure this condition holds. Note that we are *not* transforming the dependent variable, we are transforming the family parameter (or parameters) — λ , in the case of the Poisson and the quasi-Poisson; λ and θ for the Negative Binomial.

The last point of this chapter was a warning about the Bias-Variance trade-off: Including more variables fits the *data* better, not necessarily the *process* that gave rise to the data. Fewer variables may miss both the data and the underlying process. There is a happy medium — unfortunately, we cannot know what it is.

14.7: End of Chapter Materials

14.7.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

MASS This package is another “book” package — a package created for a specific book. Here, that book is “Modern Applied Statistics with S,” by William N. Venables and Brian D. Ripley (2004).

STATISTICS:

glm(formula) This function performs generalized linear model estimation on the given formula. There are three additional parameters that can (and often should) be specified.

The `family` parameter specifies the distributional family of the dependent variable, options include `gaussian`, `binomial`, `poisson`, `gamma`, `quasiBinomial`, and `quasipoisson`. If this parameter is not specified, R assumes `gaussian`.

The `link` parameter specifies the link function for the distribution. If none is specified, the canonical link is assumed.

Finally, the `data` parameter specifies the data from which the formula variables come. This is the same parameter as in the `lm` function.

glm.nb(formula) As Negative Binomial regression is fit using different methods, it cannot be included in the base `glm` command. To use the `glm.nb` command, you must include the (very helpful) `MASS` package in your script, `library(MASS)`. The output of the `glm.nb` function is similar to that of the normal `glm` command, with the inclusion of an estimate for θ and its standard error. If $\theta = 1$, then the Poisson model may be appropriate.

offset The `offset` function (or function parameter) allows us to include known varying values in our regression. The variable included as an `offset` will not have an effect parameter estimated for it.

predict(model, newdata) As with almost all statistical packages, R has a `predict` function. It takes two parameters, the `model`, and a `dataframe` of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The `dataframe` must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

14.7.2 EXERCISES

1. Show that $\mathbb{E}[Y] = \lambda$ and $\mathbb{V}[Y] = \lambda$ using the methods of Section 11.2.4.
2. Example 14.2.2 mentioned that California was an outlier in this model. First, plot the `initiative` data with California included. Second, appropriately fit the model with California included and interpret the coefficients. Finally, predict the number of initiatives Utah would have (a population of 1,722,850). Save the script as `ext01.R`.
3. In Section 14.3.3, we fit the `initiative` data using the Negative Binomial distribution. I made the statement that this model predicted 7.9 initiatives for Utah in the 1990s. Please graph the data, plot the prediction curve, and predict the number of initiatives Utah will have in the 1990s. Finally compare the results between the model with California and the model without California. Save the script as `ext02.R`.
4. Go back to the last model we fit (Eqn 14.36). Consider the comments about the model made in Section 14.5. Create a better model. Fit it with both the quasipoisson and the Negative Binomial. Plot graphs like those in Figure 14.3. Comment on the differences in the predictions between the two models. Save the script as `ext03.R`.
5. Estimate the number of initiatives that Utah had during the 1990s.
6. Prove Equation 14.27 (the formula for the probability mass function) on Page 424 is true.
7. Given the probability mass function in Equation 14.27, prove $\mathbb{E}[Y] = \mu$ and $\mathbb{V}[Y] = \mu + \mu^2/\theta$.
8. Given the definition of the Negative Binomial distribution (Equations 14.25 and 14.26), prove that an overdispersion of $\theta = \infty$ reduces the Negative Binomial to a Poisson.

14.7.3 APPLIED READINGS

- Richard Berk and John M. MacDonald. (2008) “Overdispersion and Poisson Regression.” *Journal of Quantitative Criminology* 24(3): 269–84.
- M. Katherine Hutchinson and Matthew C. Holtman. (2005) “Analysis of Count Data using Poisson Regression.” *Research in Nursing & Health* 28(5): 408–18.
- Dana Loomis, David B. Richardson, and L. Elliott. (2005) “Poisson Regression Analysis of Ungrouped Data.” *Occupational and Environmental Medicine* 62(5): 325–29.
- Katarina A. McDonnell and Neil J. Holbrook. (2004) “A Poisson Regression Model of Tropical Cyclogenesis for the Australian–Southwest Pacific Ocean Region.” *Weather & Forecasting* 19(2): 440–55.
- Ron Michener and Carla Tighe. (1992) “A Poisson Regression Model of Highway Fatalities.” *American Economic Review* 82(2): 452–56.
- Marta N. Vacchino. (1999) “Poisson Regression in Mapping Cancer Mortality.” *Environmental Research* 81(1): 1–17.
- Weiren Wang and Felix Famoye. (1997) “Modeling Household Fertility Decisions with Generalized Poisson Regression.” *Journal of Population Economics* 10(3): 273–83.
- Lisa A. White. (2009) *Predicting Hospital Admissions with Poisson Regression Analysis*. Masters Thesis. Naval Post-Graduate School.

14.7.4 THEORY READINGS

- Kurt Brannas. (1992) "Limited Dependent Poisson Regression." *Journal of the Royal Statistical Society. Series D (The Statistician)* 41(4): 413–23.
- A. Colin Cameron and Pravin K. Trivedi. (1998) *Regression Analysis of Count Data*. New York: Cambridge University Press.
- Edward L. Frome. (1981) "Poisson Regression Analysis." *The American Statistician* 35(4): 262–63.
- Jie Q. Guoa and Tong Li. (2002) "Poisson Regression Models with Errors-in-Variables: Implication and treatment." *Journal of Statistical Planning and Inference* 104(2): 391–401.
- Alexander Kukush, Hans Schneeweis, and Roland Wolf. (2004) "Three Estimators for the Poisson Regression Model with Measurement Errors." *Statistical Papers* 45(3): 351–68.
- Alfonso Palmer, J. M. Losilla, J. Vives, and R. Jiménez (2007) "Overdispersion in the Poisson Regression Model: A comparative simulation study." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 3(3): 89–99.
- Tsung-Shan Tsou. (2006) "Robust Poisson Regression." *Journal of Statistical Planning and Inference* 136(9): 3173–86.
- William N. Venables and Brian D. Ripley. (2004) *Modern Applied Statistics with S*, 4th edition. New York: Springer.
- Rainer Winkelmann. (2000) *Econometric Analysis of Count Data*. New York: Springer.
- Liming Xiang and Andy H. Lee.(2005) "Sensitivity of Test for Overdispersion in Poisson Regression." *Biometrical Journal* 47(2): 167–76.
- Feng-Chang Xie and Bo-Cheng Wei. (2009) "Diagnostics for generalized Poisson regression models with errors in variables." *Journal of Statistical Computation & Simulation* 79(7): 909–22.

A photograph of a golf course with trees in autumn foliage. The trees are in various stages of fall, with some showing bright orange and red leaves, while others are still green. The grass is a mix of green and brown, suggesting some dryness. The sky is clear and blue. The overall scene is peaceful and scenic.

CHAPTER 15:

NOMINAL AND ORDINAL DEPENDENT VARIABLES

OVERVIEW:

This chapter finishes our time examining various types of dependent variables. In this chapter, we examine dependent variables that are categorical — both nominal and ordinal response variables are covered in this chapter.

This could be variables with an ordering (like a Likert scale) or without (pet type). The type of regression used needs to take into consideration the characteristics of the dependent variable. Thus, this chapter starts with modeling nominal variables (no ordering) and proceeds to ordinal variables (with ordering).

Chapter Contents

15.1	Nominal Dependent Variable	442
15.2	Ordinal Dependent Variable	454
15.3	Extended Example: Cattle Feed	458
15.4	Extended Example: The State University of Ruritania	463
15.5	Conclusion	467
15.6	End-of-Chapter Materials	468



One of the most pervasive research questions in Political Science is to predict a person's vote based on demographic information. In other words, if you know a person's age, gender, income, education, and religion, how well can you predict how that individual will vote in the upcoming parliamentary election?

At first glance, this question appears to be a binary dependent variable problem. After all, there are only two parties, right? Well, *even* if you ignore third parties, there is a third option: abstention. In each Ruritanian parliamentary election, a sizable number of registered voters decide not to vote. For instance, in the 2016 election, while Kuzněcov (of the royalist Král a Země party) received 48% of the vote cast and Ivanović (of the republican Republikánská Strana) received 46%, a full 45.3% of the eligible voters did not vote. Thus, the distribution of votes in this election is 26.3% Kuzněcov, 25.2% Ivanović, 3.2% other, and 45.3% none of the above. As such, conclusions based on those models that assume a binary outcome have definite issues with generalization to the voting public at large. They are ignoring important information.

A better alternative is to specifically add in 'abstention' and model the three possible outcomes at once (or 'abstention' and 'other' and model the four). Such a regression model is called a **nominal regression model** or a multinomial regression model, because there is no inherent ordering among the levels of the dependent variable.

There is a second type of dependent variable that is closely related to the nominal case — the ordinal dependent variable. The difference between the nominal and the ordinal is that the ordinal has more information contained in it. There is no ordering in the nominal case, whereas there *is* an implicit ordering in the ordinal case.¹ Examples of ordinal variables include ratings and indices.

If we just use our logistic regression methods (Chapter 12), we come up with some odd results. If we force a nominal variable into just two categories, we lose information in the data. If we treat ordinal dependent variables simply as nominal,

¹Ordinal is actually a portmanteau for "ordered nominal."

information is also lost. If we treat them as continuous, our conclusions may not match reality.

Thus, both nominal and ordinal dependent variables need their own modeling methods. This chapter examines how to model both the nominal dependent variable and the ordinal dependent variable more properly.²

Note: This chapter sits uneasily here. From the standpoint of the dependent variable type, this is its proper place. However, these are not generalized linear models (GLMs). They are particular expansions to the GLM paradigm. As such, if you are looking at the GLM modeling method as being the unifying theme to this part of the book, this chapter should not exist.

paradigm

But, it does.

²The study of statistics emphasizes both estimating the value (expected value) and the variance of that estimate (confidence interval).

15.1: Nominal Dependent Variable

A nominal variable is a categorical variable where there does not exist a meaningful ordering in the categories. Examples may include job type, presidential vote (and non-vote), and beer brand choice. These variables are categorical — not numeric — and the categories have no inherent ordering. White Collar is not ‘greater than’ Professional. Voting *monarčista* is not ‘more than’ voting *republikán*. *Widmer* is not ‘more than’ *Coors*.³ How do we model such dependent variables?

There are a couple of ways of doing this. The first, easiest, and most understandable is to model the variable as a series of binary dependent variables. We already understand how this works, the testing of the model is already conceptually understood, and it works (*not really, but close?*).⁴ There are just a couple things to clarify.

15.1.1 MATHEMATICAL MODEL As with the simply binary dependent variable case, let us layout the mathematical background to the nominal dependent variable case. As in the binary dependent variables case, we are actually modeling the underlying probabilities of each of the outcomes. Also, as in the binary case, there are five requirements for the random variable to follow a Multinomial distribution (*cf.* Section 13.1):

1. the number of trials, n , is known;
2. each trial has J possible outcomes;
3. the success probability for each trial, $\{\pi_1, \pi_2, \dots, \pi_J\}$, is constant;
4. each trial is independent from the others; and
5. the random variable is the number of each type of outcome in those n trials.

Thus, if we let π_j be the probability that category j is selected, then the following two conditions must hold:

$$0 < \pi_j < 1 \quad \text{for all } j \in \{1, 2, \dots, J\} \quad (15.1)$$

$$\sum_{j=1}^J \pi_j = 1 \quad (15.2)$$

³Of course, there may be a time when you are predicting *republikán* vote by examining an underlying level of conservatism. In such a case, *monarčista*–*republikán* would be ordered. Thus, it really depends on what you are predicting (as always).

⁴Usually. Nothing in statistics *always* is best. As you have seen by now, there are always methods that work better, but with trade-offs. The science here is to be aware of the strengths with the weaknesses and balance them to get closer to the true process you are trying to model.

Condition (15.1) must hold because we are dealing with probabilities bounded by 0 and 1, and Condition (15.2) holds because one of the J possible outcomes *must* happen. In the binary case, our two probabilities were π and $1 - \pi$, which satisfies the second condition by default and the first because it makes no sense to study phenomena that always or never occurs.

When we generalize the binary case, we need to select an appropriate probability distribution — one that can model J possible outcomes with J different probabilities. That distribution is called the multinomial distribution.⁵ The probability density function for the multinomial distribution in the general case is

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{n!}{x_1!x_2!\cdots x_J!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_J^{x_J} \quad (15.3)$$

Here, x_i are non-negative integers and $\sum x_j = n$. The expected value of this distribution for a specified outcome is

$$\mathbb{E}[X_j] = n\pi_j \quad (15.4)$$

and the covariance between two outcomes is

$$\text{Cov}[X_i, X_j] = -n\pi_i\pi_j \quad (15.5)$$

Be aware that \mathbf{X} is a vector. So, if $n = 1$ and $J = 4$, the following could be outcomes from the Multinomial distribution:

$$x = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}; \quad x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \quad x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

In the first example, a 2 came up; in the second, a 1; in the third, a 3. Note that in each case, the sum of the entries is n and the number of entries is J .

Now, if $n = 4$ and $J = 3$, the following could be outcomes from a Multinomial distribution:

$$x = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}; \quad x = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}; \quad x = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix}$$

In the first example, three 2s and a 3 came up; in the second, two 1s, a 2, and a 3 came up; in the last, four 3s came up.

Note: Be aware that the sum of the entries in each outcome vector is n and that the number of entries is J .

⁵Recall that the distribution in the *binary* case was the *binomial* distribution.

If the random variable \mathbf{X} follows a Multinomial distribution with $n = 3$ and $\boldsymbol{\pi} = [0.1, 0.5, 0.4]'$, then we could write it as

$$\mathbf{X} \sim \text{Multi} \left(n = 3, \boldsymbol{\pi} = \begin{bmatrix} 0.1 \\ 0.5 \\ 0.4 \end{bmatrix} \right) \quad (15.6)$$

and the expected value of \mathbf{X} would be

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} 0.3 \\ 1.5 \\ 1.2 \end{bmatrix} \quad (15.7)$$

The expected value of X_3 would be $\mathbb{E}[X_3] = 1.2$.

Note: Make sure you see that this is just an extension of the binomial distribution, where

$$f_X(x) = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x} \quad (15.8)$$

with

$$\mathbf{X} = \begin{bmatrix} x \\ n-x \end{bmatrix} \quad \text{and} \quad \boldsymbol{\pi} = \begin{bmatrix} \pi \\ 1-\pi \end{bmatrix} \quad (15.9)$$

Example 1

Let us illustrate the multinomial distribution with a typical “rolling a die example.” Assuming that the die is fair, then the probability of rolling each of the six outcomes is $\frac{1}{6}$. If we roll a fair die 3 times, what is the probability the outcome is $[1, 0, 1, 0, 0, 1]'$ (that is, a 1, a 3, and a 6 come up)? What is the expected value of \mathbf{X} ?

Solution: This is a multinomial experiment. There are a fixed number of possible outcomes (six), the probabilities of each outcome are constant (they do not change as we roll the die), and the probabilities sum to one. As such, we know the probability mass function is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{3!}{x_1! x_2! x_3! x_4! x_5! x_6!} \left(\frac{1}{6}\right)^{x_1} \left(\frac{1}{6}\right)^{x_2} \left(\frac{1}{6}\right)^{x_3} \left(\frac{1}{6}\right)^{x_4} \left(\frac{1}{6}\right)^{x_5} \left(\frac{1}{6}\right)^{x_6} \quad (15.10)$$

Thus,

$$\mathbb{P} \left[\mathbf{X} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right] = \frac{3!}{1! 0! 1! 0! 0! 1!} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 \quad (15.11)$$

$$= 6 \left(\frac{1}{6}\right)^3 \quad (15.12)$$

$$= \frac{1}{36} \quad (15.13)$$

Thinking through the problem should get us to the same point.

Finally, we know the expected value is

$$\mathbb{E}[\mathbf{X}] = n\boldsymbol{\pi} = 3 \begin{bmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} \quad (15.14)$$



As we have a formula for our expected value, we have our mechanism for estimating the several π_j : in an experiment (or set of data), count the number of times outcome j occurred and divide by the total number of trials (or records). This is actually the maximum likelihood estimator for π_j . Thus, our linear predictor is

$$\text{logit}(\pi_j) = \beta_{j,0} + \beta_{j,1}x_1 + \beta_{j,2}x_2 + \cdots + \beta_{j,k}x_k \quad (15.15)$$

Notice that this linear predictor has $k + 1$ parameters to estimate for each of the j categories. Thus, you will need more than $j(k + 1)$ pieces of data to fit it. There are ways to reduce the dimensionality of the problem (reduce the number of parameters in need of estimation); however, these are beyond the scope of this book.

We need the logit link (or something just like it) to force our linear predictions to be in the range $\pi_j \in (0, 1)$. As any link that maps $g : \mathbb{R} \rightarrow (0, 1)$ is acceptable, we could use the log-log link, the complementary log-log link, the probit link, or any of an infinite number of others. . . in theory. As before, the choice of the link function is largely a matter of tradition. If you deviate from tradition, the burden of proof is on you to justify the selection. Furthermore, the differences are usually slight. If the differences are large, then there is something wrong with your research model. Because of this, it would behoove you to fit your research model using a couple different (appropriate) link functions to help determine the stability (robustness) of your results.

robustness

Note: Thus, there are two things that you need to take away from this discussion: First, we are able to fit the entire model at once because we have a distribution that can produce the necessary nominal results. Second, we model the underlying probabilities (like in the binary case), not the actual outcomes, as usual.

To see this in action, let us look at an extended example.

Example 2

The General Social Survey (GSS) at the University of Chicago conducts an extensive survey of adult Americans every year. The data is freely available from NORC. In this small subset of the data, `gssocc`, I would like to predict a person's occupation category (`occ`) based on race (`white`), years of education (`ed`), and years of experience (`exper`).

Before getting started, let us examine the variables involved.⁶ The race variable is binary, with a '1' representing the person identifying as 'white' and a '0' otherwise. As a side note, this is a race variable, not an ethnicity variable. Thus, Hispanics may self-identify as either white or non-white. Also note that this is a self-identification variable; that is, the individual being surveyed decided his or her reported race. Looking at a frequency count, a full 91.69% of the respondents stated they were white. This is significantly higher than the population at large, where approximately 80% of Americans were white when the survey was conducted. When we do the final analysis, we need to keep this in mind, as it is not necessarily representative of the nation as a whole.

The median number of years of education in the sample is 12 years, which corresponds to graduating from high school. The mean number of years is 13.09, which indicates the sample is right skewed (the Hildebrand ratio is +0.37). Furthermore, it is interesting to note that 51.0% of the sample only graduated from high school. Additionally, 23.4% of the sample received a bachelor's degree or more, which is close to the population (27% have received a bachelor's degree or higher). Finally, 18.7% of the sample did not graduate from high school, which is close to the 15% estimate of the population. From this, it appears as though the sample is representative of the population in terms of educational attainment.

The third independent variable is the years of experience in the job. There are no general statistics for the population, so we will have to make a large assump-

⁶The raw — and current — data can be accessed from <http://www.norc.uchicago.edu/GSS+Website/>.

	White	Education	Experience
White	1.0000	0.0243	-0.0794
Education	0.0243	1.0000	-0.2740
Experience	-0.0794	-0.2740	1.0000

Table 15.1: Correlation matrix for the three independent variables in the example, from *gssocc* data.

tion that the sample represents the population.⁷ In the sample, the years of experience varies widely, from 2 to 66 years. The median is 17 years and the mean is 20.5 years. Thus, the sample is also right skewed. This makes sense as this is a count variable. Count variables tend to be right skewed as they cannot take on negative values. In fact, there is nothing in the distribution of the experience variable that looks wrong. With that said, however, one still needs to mention the caveat.

Looking at the correlations amongst the independent variables can help us avoid any unpleasantness and surprises due to collinearity and multicollinearity. The correlation matrix (Table 15.1) does not show any hint of multicollinearity. In fact, this correlation matrix suggests that these three variables are *effectively* independent of each other.⁸

Finally, let us note that there *may* be an inherent ordering in some of the jobs (White Collar greater than Blue Collar), but not for all five of the categories. As such, this is definitely a candidate for nominal regression.

⁷This was a safe assumption with respect to the education variable, but not with respect to the white variable. As such, it needs to be mentioned that you are unable to check the representativeness of the experience variable.

⁸Pearson's product-moment correlation test indicates that the correlation between education and experience is statistically significant at the $\alpha = 0.05$ level ($t = -5.2152, df = 335, p \ll 0.0001$). However, the coefficient of -0.2740 is a low level of correlation.

	Estimate	Std. Error	z-value	p-value
Intercept	3.1036	1.0110	3.07	0.0021
White	0.7090	0.6213	1.14	0.2538
Years of education	-0.3721	0.0640	-5.81	0.0000
Years of experience	-0.0259	0.0113	-2.30	0.0215

Table 15.2: Results from the GLM (using the binomial family and the logit link) predicting whether or not a person is a blue collar worker. The AIC for this model is 304.75.

NOMINAL REGRESSION: Now, let us model the outcome variable with the three independent variables. Actually, we need to step back and really think about what we mean by ‘model the outcome’. Do I want to predict the probability that a person will be Blue Collar given the input variables? Or: Do I want to predict the job category given the input variables? These are different questions. They require slightly different methods.

The first question actually asks a binary question: What is the probability that a person will be Blue Collar (compared to *all* of the other job categories)? This is very much like the questions asked in Chapter 12. Here, the dependent variable takes on values 1 (Blue Collar) and 0 (not Blue Collar).

To answer *this* question, we need to create a variable called `bluecol` as an indicator variable for Blue Collared-ness. Thus, the model we fit will be

```
bluecol ~ white + ed + exper
```

We would fit it using a generalized linear model, a binomial family, and a logit link. The results of the regression are in Table 15.2. From this model, we can perform all of the goodness of fit measures from Chapter 12.

Looking at the results from running the model, we see that greater levels of education and greater levels of experience are associated with a lower probability of being a blue collar worker. For Bob, an individual who responded that he was white, had 20 years of education, and 10 years of experience in their current job, the probability of being a blue collar worker is approximately 2% (as compared to not being a blue collar worker).

Note: This last part is subtle, but extremely important. Here is why: What is the probability that Bob is a white collar worker? If we do the same steps above, we get that the probability that Bob is a white collar worker (as compared to not being a white collar worker) is 13.1%. Similarly, if we continue performing separate logistic regressions, the probability that Bob is a professional is 96.9%; menial, 2.3%; and craft, 7.9%.

Note that all of these probabilities add up to **more than 100%**. There is something wrong here, since the probability that Bob holds one of these five job types cannot be greater than 100%.

important

moral of the story

The problem is that we kept changing the base category. In Chapter 12, we never mentioned the need to specify the base category since it always defaulted to the opposite of what we were modeling. In other words, we were actually measuring the probability of an event as compared to the probability of ‘not the event’ (a.k.a. the odds of the event). This ensured that the probabilities always added up to 100%. Within each of the above five regressions, if we added the probability of the event that Bob holds job type X with the probability that Bob holds job type not X , we always get 100%.

the lesson

The lesson: Comparing probabilities of events is not as easy as when we were only working in the binary realm. It is doable — easily so, with one small change. We need to select a base category that does not change throughout our analysis. The choice is up to you, as all choices are equally acceptable from a statistics standpoint.

Since we can select any job type as our base, let us select Blue Collar, since it is the first level according to the alphabet. (We will see again shortly how to switch between the bases.)

To perform this modeling, you will have to load the `nnet` package. Since this comes with your base distribution of \mathbb{R} , there is no need to install it. Once loaded with the `library(nnet)` command, to fit the better model, use the \mathbb{R} command

```
|| multinom(occ ~ white + ed + exper)
```

Because of the large amount of output, the regression table is structured slightly different. The coefficients (in logit units) and the standard errors are still presented. The statistical significance is not. However, a quick rule of thumb is that the variable is statistically significant (at the $\alpha = 0.05$ level) if the parameter estimate is more than twice the standard error. Table 15.3 presents the output from modeling the data in the form given in the output.

Note that one of the five job types is missing: Blue Collar. This is because all of the probabilities are measured *with respect to* Blue Collar. Thus, these percentages are directly comparable (after transforming from logit units).

\mathbb{R} is nice in that if you `predict` on a multinomial model, it will give you the category with the highest probability, by default. Thus, according to this model, Bob will most likely be a Professional (which was our conclusion above). If we want the probabilities for each of the possible job types for Bob, we need to add a `type="probs"` parameter to our function call:

```
|| predict(model.mn1, newdata=BOB, type="probs")
```

Such a call gives us the following probabilities (which sum to one, as they should):

Coefficients:				
	Constant Term	White	Education Level	Experience
Craft	-1.8328	-0.7642	0.1933	0.0230
Menial	-0.7412	-1.2365	0.0994	-0.0074
Prof	-12.2595	0.5376	0.8783	0.0309
WhiteCol	-6.9800	0.3349	0.4526	0.0299
Std. Errors:				
	Constant Term	White	Education Level	Experience
Craft	1.1861	0.6324	0.0775	0.0126
Menial	1.5195	0.1996	0.1023	0.0174
Prof	1.6681	0.7996	0.1005	0.0144
WhiteCol	1.7144	0.9340	0.1023	0.0153

Table 15.3: Results of the multinomial regression. Note that the p -values are not provided. To determine which independent variables are statistically significant for predicting the dependent variable levels, divide the coefficient estimate by the standard error. If that ratio is greater than 2, then the variable is statistically significant at the $\alpha = 0.05$ level.

BlueCol	Craft	Menial	Prof	WhiteCol
0.0020	0.0091	0.0020	0.9565	0.0304

Base switching: If you wish to switch your base category, there are two options. First, you can subtract the parameter estimates of the new base from all the other bases. Thus, if we want to change the base from Blue Collar to Professional, we would subtract the Professional parameter estimates from the other parameter estimates. So, for example, the White Color estimates with Professional as the base will be $-6.9800 - -12.2595 = 5.2795$. Unfortunately, the standard errors are not so easily calculated — or at all reasonably calculable by hand.

Also unfortunately, most statistical programs require you to physically reorder the data to select a different base; most programs use the level of the first data point as the base category. R does allow you to switch among the bases without having to physically alter the data. Unfortunately, the method is rather arcane. Fortunately, the `RFS` package has a function, `set.base` that allows you to change the bases much more easily.

Thus, to set `craft` as the base, we would use the command

```
|| occ = set.base(occ, base="craft", data=gssocc)
```

I leave it as an exercise to rerun the analysis with `craft` as the base. Check that the parameter estimates follow the above observation.

INTERPRETATION: The interpretation of the coefficients (parameter estimates) is the same as for the binary dependent variable case. Just remember that the coefficients are in logit units. In R, however, this library does not require you to back-transform your predictions. To remember this, just look at the output — it is in proportions already (a quick check is that they sum to one).

GOODNESS OF FIT: The first check of the goodness of the model is the relative accuracy (see also Section 12.5). The accuracy is the number of correct predictions divided by the number of cases. The relative accuracy divides this number by the accuracy of always selecting the modal category (**the null model**). For this dataset, the modal category is Professional, with 140 out of 337 cases belonging to Professionals. Thus, the relative accuracy is $\frac{169}{337} / \frac{140}{337} = 1.207$. Thus, this model improves accuracy by 21% over the null model. Is this good? It depends on your other models.

As Maximum Likelihood Estimation is used, the Akaike Information Criteria score is also reported. For this model, $AIC = 885$. Is this good? Again, it depends on your other models. In other words, model comparison needs another model. I leave it as an exercise to see that the null model has $AIC = 1027$. Thus, our model is much better than the null model.

Now that we have looked at our model, let us look at the parameter estimates. According to our model, Whites have a higher probability of being Professionals and White Collar workers than they are to be Craft or Menial laborers. As for education, higher levels of education are associated with higher odds of being a Professional or a White Collar worker (both of these are statistically significant) than being a Blue

Collar worker. Finally, years of experience are not a statistically significant predictor of job type, as none of the coefficients are statistically significant (coefficient / standard error > 2).⁹

So, we have a picture of Professionals and White Collar workers, when compared to Blue Collar workers: they are White and well educated. Not an earth-shattering conclusion, but it is encouraging to see that our conclusions do seem to reflect reality.

⁹This rule of thumb comes from the fact that in a Normal distribution, the ratio needs to exceed 1.96 to be statistically significant at the $\alpha = 0.05$ level. These parameter estimates are not guaranteed to be Normally distributed. As such, the rule of thumb is to be more conservative. Even with the rule of thumb, do not bet the farm.

15.2: Ordinal Dependent Variable

Another variety of categorical dependent variables is ordinal. A variable is ordinal if it is categorical *and* the categories have an underlying order to them. Examples include movie ratings (number of stars), hurricane intensity, and so forth.

There are actually at least four ways of handling ordinal dependent variables:

1. Treat them as nominal. This allows us to fit ordinal data using previous techniques. Unfortunately, it is inefficient as it ignores important aspects of the data itself.
2. Treat their cumulative level as nominal. If the ordinal variable takes on values 1 – 5, then create nominal variables corresponding to Level 1, Levels 1 and 2, Levels 1–3, Levels 1–4, and Levels 1–5. This preserves much of the underlying information *and* allows us to fit it using a previous method.
3. Assume that there is an underlying continuous process that you wish to fit. The ordinal nature is just several threshold values along the possible values. This reduces to a pseudo-OLS, where you also need to fit the threshold values, not just the slopes and intercepts. Using Maximum Likelihood methods, this is trivial to solve.
4. Pretend that the ordinal values are continuous and fit it using ordinary least squares or one of its offsprings. This has the advantage of being easily fit.

Three of these ways have already been discussed, and you are quite adept at using them (Options 1, 2, and 4). Only the third option is completely new to you. This chapter focuses on how to fit Option Three.

15.2.1 OPTION THREE Let us assume that there is an underlying continuous process. We only experience (observe) this process through the ordinal variable. This is very similar to how we first looked at binary variables: underlying process exhibited only in the 0/1 outcomes (see Figure 12.2). Here, there is more than just the one threshold (which traditionally defaulted to 0.500). Thus, we have two sets of parameters to fit. The first is the parameters which describe the process (the β s). The second is the position of those threshold values (the τ s).

Without going into the details, we will use Maximum Likelihood Estimation as our fitting method because it has many nice properties. Thus, our underlying process is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (15.16)$$

Our thresholding process is illustrated in Figure 15.1. The line represents the underlying continuous process that you are trying to model. The A, B, C, and D represent the observed ordinal values. The threshold values, τ_1 , τ_2 , and τ_3 are the values of η that separate the observed ordinal values.

This model is very straight forward and understandable. Using R to obtain the fitting is also straight forward. The results presented are also relatively straight forward.

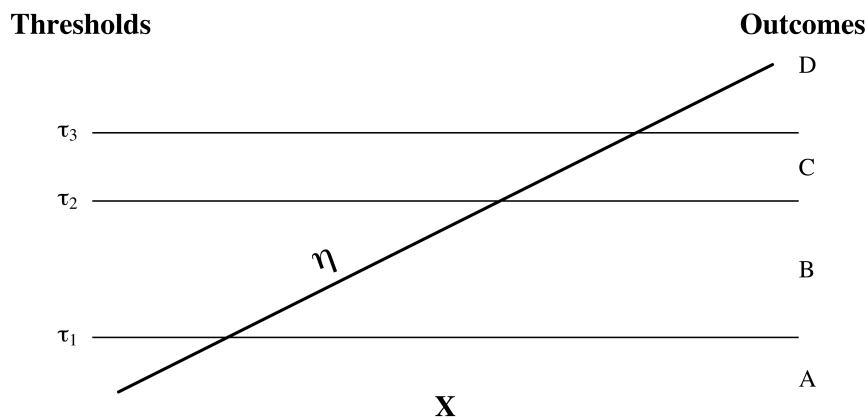


Figure 15.1: Schematic diagram of the thresholding process. The line represents the linear continuous process. The τ s represent the threshold values. A, B, C, and D represent the ordinal outcomes.

Variables:		Value	Std. Error	t-value
	Woman	0.743	0.078	9.50
	White	-0.400	0.118	-3.39
	Age	-0.020	0.0024	-8.17
	Years of Education	0.098	0.013	7.52
Thresholds:				
	SD — D	-1.700	0.237	-7.18
	D — A	0.111	0.233	0.48
	A — SA	1.979	0.236	8.37

Table 15.4: Result of ordinal regression in R. Note that the women tend to view President Obama in a more favorable light; whites, less; older, less; and higher educated, more. All of these agree with multiple surveys throughout his tenure as President.

Example 3

Let us use some more data from the GSS. This data explores the ‘warmth of feeling’ the respondent has for President Obama. The demographic information is the gender (`male`), the race (`white`), the age, and the number of years of education (`ed`). The response variable has four ordered levels: Strongly Disagree (SD), Disagree (D), Agree (A), and Strongly Agree (SA). Our goal is to explain a person’s feelings toward the president based solely on demographic information.

Solution: Let us fit this data with ordinal regression. The function in R is `polr`, which stands for “proportional odds logistic regression” (although the `probit` is an option as a link function). This function requires the `MASS` package. Thankfully, since `MASS` also comes with the base distribution of R, there is no need to install it, only to load it via the `library(MASS)` command.

The actual command to fit this model using ordinal regression is

```
|| polr( warm ~ male + white + age + ed )
```

This command will give the coefficients of the underlying linear regression and the threshold values separating the four categories. From Table 15.4, we see that the equation for the underlying linear process is

$$\eta = 0.743 \times \text{Woman} + -0.400 \times \text{white} + -0.020 \times \text{age} + 0.098 \times \text{ed}$$

The thresholds are also listed. The threshold between Strongly Disagree and Disagree is at $\tau_1 = -1.700$. The threshold between Disagree and Agree is $\tau_2 = 0.111$. The threshold between Agree and Strongly Agree is $\tau_3 = 1.979$. Thus, to calculate our prediction, we calculate the prediction based on the linear model, η , and compare that value to the intervals described by the thresholds. Thus, for Bob, who is Male, White, 40 years old and has 20 years of education, we have

$$\eta = 0.740 \times 0 + -0.400 \times 1 + -0.020 \times 40 + 0.098 \times 20 = 0.76$$

As $\eta = 0.76$, we have our prediction that Bob agrees with the president. If we actually want probabilities that Bob Strongly Disagrees, Disagrees, Agrees, or Strongly Agrees, we would have to back-transform using the inverse of the logit function and calculate each probability using integral calculus... or we could just ask the computer to do it for us:

back-transform

```
||| BOB = data.frame(male="Men", white="White", age=40, ed=20)
||| predict(model.o11, newdata=BOB, type="probs")
```

This gives the probabilities as

SD	D	A	SA
0.0785	0.263	0.429	0.229

Thus, it is far from certain that Bob agrees (or strongly agrees) with the president, although that probability is rather high: $0.429 + 0.229 = 0.658$. ♦

ACCURACY: Finally, let us look at the accuracy of the model. I leave it as an exercise to show that the relative accuracy is 1.105, which indicates that the model is about 10.5% better than the null model (the modal category is “Agree”). This is not a fantastic increase in accuracy, but we do know how certain demographics feel about the president: Whites tend to disagree, Males tend to disagree, older people tend to disagree, and lesser educated people tend to disagree.

Of course, we could have added in a quadratic education term to the model to see if both the more-educated and the less-educated both support the president. I also leave this as an exercise to show that there is no evidence of this. Thus, we have no evidence that the relationship between education and presidential support is anything other than linear.

15.3: Extended Example: Cattle Feed

Now that we have been introduced to these two new types of regression, let us deal with an example of each. This example tries to predict the feed type used for a cow. Such a question would arise if there is missing data in your data file and you wanted to estimate the missing value instead of throwing out the entire record.

Example 4

Previously, we attempted to model the weight of cattle based on a few factors. Let us try something different. Let us predict the brand of food used by the cattle based on the ranch, age, and weight.

Specifically, let's first model feed type. Then, let's say that the RUR ranch sent a 21-year-old cow to slaughter at 1197 pounds. Which food brand was most likely used? What are the probabilities of each brand being used?

Solution: Since the food brand is a nominal variable, we will use multinomial regression. The data file is `cattleData`. Let's load it and look at some summary statistics on it.

```
library(nnet)
cowz = read.csv("http://rur.kvasaheim.com/data/cattleData.csv")
attach(cowz)

summary(cowz)

cor.test(weight, age)
table(ranch, feedType)
```

Note that there is (as expected) a strong correlation between age and weight. If we are doing model selection, we will need to keep this in mind as this multicollinearity will decrease the statistical significance of those two variables.

Note from the cross-tabulation that the EVA ranch only used Purina and the TCL ranch only used Rangeland (in this sample). That fact would make it really easy to predict the feed type for those ranches. The other ranches use a combination of all of the brands.

With this information to guide us, we fit the model

```
cowModel = multinom(feedType ~ weight + ranch + age)
summary(cowModel)
```

The first line fits the model. Note that the model did converge, so we can pay attention to the results. If it had not converged, we should first change the link function, then realize that the multicollinearity is a problem. Dropping one or more variables would be an appropriate action in that case.

The results of the `summary(cowModel)` command gives some insight into the relationships. First, note that the coefficient estimate for `ranchEVA` for estimating `Purina` is 20.37. This is extremely high, meaning it is almost guaranteed that a cow from the EVA ranch will use Purina.

But, from the cross-tabulation above, we already knew this.

Similarly, the coefficient estimate for `ranchTCL` for Rangeland is a huge 22.32. This indicates a cow from the TCL ranch will most certainly use Rangeland food. Again, we knew this from our cross-tabulation.

Note from the regression table that `Accuration` is missing. All feed measurements are taken with respect to that level. This is important to keep in mind if we do this by hand. It is just something to note if we are using the computer to do our calculations.



So, let's estimate the food used by our mystery cow. First, let's define it:

```
|| mysteryMoo = data.frame(weight=1197, age=21, ranch="RUR")
```

Now, let's predict the probabilities it used each of the feed types:

```
|| predict(cowModel, mysteryMoo, type="prob")
```

The results tell us that the mystery cow most likely used Steakmaker. In fact, the probability it used Steakmaker was 79%. The second most likely feed type was Accuration (13%). ◆

15.3.1 GRAPHICS Let us talk about graphics for a bit. A two-dimensional scatter plot looks at two numeric variables. We can therefore easily plot a prediction curve when dealing with only a single dependent variable and single independent variable.

If there is a second independent variable, we can plot several curves, one for each level in that second independent variable.

Once we move beyond two independent variables, graphics are more difficult to do. A simple regression model like the one above may require dozens of graphics to illustrate each aspect.

However, we can simplify things by focusing on only a couple independent variables at a time. The choice depends on the story you are trying to learn (or tell).

GRAPHIC: FEED TYPE VERSUS WEIGHT: For this first graphic, I am consciously making the decision to plot the predicted probability on the y-axis, the cattle weight on the x-axis, and have a prediction curve for each feed type. This will allow me to see the effect of weight on the predicted food type.

This means I need to select values for the other two independent variables. For the numeric age, I would typically use its mean or median, whichever was the "typical" age for these cattle.

For the selected value of the ranch, I would either select the ranch to which the mystery cow belonged (to continue that story) *or* the most popular ranch (to try to generalize the story). It is best to do separate graphics for all ranches so that you, the researcher, can better understand the effect of ranch on the probabilities. It is always better to do more to understand.

So, here is the code to create the predictions:

```

theWeights = seq(1019,1579, length=1e4)
theAge = median(age)
prRUR = predict(cowModel, newdata=data.frame(weight=theWeights
, age=theAge, ranch="RUR"), type="probs")

```

The `prRUR` variable contains 10,000 rows (one for each weight) and 5 columns (one for each feed type). The entries are the probabilities.

Now, we just plot the data and these predictions:

```

par(family="serif", las=1)
par(xaxs="i", yaxs="i")
par(mar=c(4,4,0,0)+0.5)
par(cex.lab=1.2, font.lab=2)

plot.new()
plot.window( xlim=c(1000,1600), ylim=c(0,1))

axis(1); axis(2)
title(xlab="Weight [lb]")
title(ylab="Probability at RUR Ranch")

lines(theWeights,prRUR[,1], col=1) # Accuration
lines(theWeights,prRUR[,2], col=2) # Purina
lines(theWeights,prRUR[,3], col=3) # Rangeland
lines(theWeights,prRUR[,4], col=4) # Steakmaker
lines(theWeights,prRUR[,5], col=5) # Wind and Rain

legend("topright", bty="n", col=1:5, lwd=2,
      legend=c("Accuration", "Purina", "Rangeland", "Steakmaker", "Wind
and Rain")
)

```

Note that this graphic includes a legend that lets the reader know which probability curve belongs to which feed type. Legends are rather important to include on a graphic. Remember that graphics should be stand-alone with their caption. Because a legend contains so much information, it requires a large function. To see all a legend can do, run `?"legend"` in R.

Figure 15.2 is the resulting graphic. Note that the predicted feed type tends to be either Steakmaker, for light cows, or Purina, for heavy cows. When the weight of the cow is middling, there is great uncertainty in which feed type it used.

It is interesting that this analysis gives us additional insight on how we can create big cows for slaughter. This suggests we should use Purina brand. This conclusion, however, is based only on the RUR ranch and a middle-aged cow.

More importantly, this conclusion assumes that the data are representative of the population of interest. As this data was originally collected in conjunction with a dissertation in Animal Science, I tend to think it is representative.

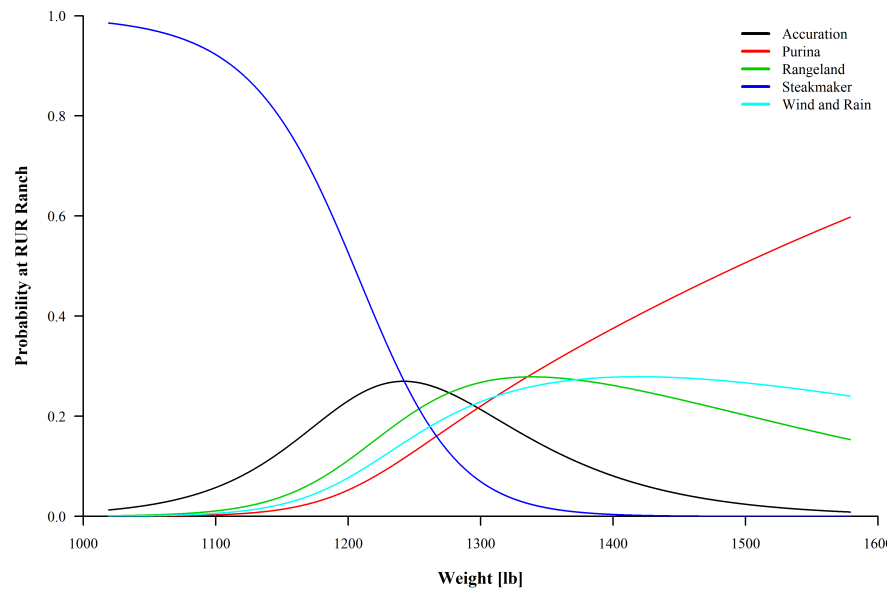


Figure 15.2: Graphic of the probability for each feed type at the RUR ranch. The probabilities vary with the cow's weight. The age is held at the median, 21.

From a strictly statistical standpoint, the additional insight is limited. However, if you are hired by RUR ranch to determine the best feed type, this graphic would be *very* persuasive for you, the decision-maker.

While we could create a similar graphic for feed type against age, I am not convinced it would be helpful. Age is not something one would like to optimize like weight. In other words, I am not sure what story I would tell about it.

Note: Don't make graphics just for fun. Make sure you create them knowing how to interpret them.

15.4: Extended Example: The State University of Ruritania

A second example will try to model the level of the student given some information about the student. Again, this may be interesting for imputation (filling in missing data).

impute

Example 5

Previously, we modeled the grade point average of students at the State University of Ruritania (*Státní Univerzita v Ruritánii*). Let us turn this around and model the student's class (Freshman, Sophomore, Junior, Senior) given only the gender and the current GPA of the student.

Let us also predict the class of Eliska, a female student with a 3.33 GPA.

Solution: As usual, the first step is to import the data and look at a summary, including a cross-tabulation of our categorical independent variable and the dependent variable:

```
library(MASS)

suvrData = read.csv("http://rur.kvasaheim.com/data/suvr.csv")
summary(suvrData)
```

Let us pause here. Note that the `class` variable is an ordinal variable. We need to let R know this:

```
suvrData$class = ordered(suvrData$class, levels=c("Non-
  Matriculated", "Freshman", "Sophomore", "Junior", "Senior")
)
summary(suvrData)
```

There we go, the levels for the `class` variable are in the right order. Let's continue.

```
attach(suvrData)
table(gender, class)
```

Note that none of the non-matriculated students are female. This is just something to know and remember as we get results.

Now, we can fit our model and look at the summary results:

```
suvrModel = polr(class ~ gender + gpa, data=suvrData)
```

```
|| summary(suvrModel)
```

check!

A quick check that you have ordered the levels correctly is to look at the second table in the summary output. The rows should describe subsequent levels.

The AIC of this model is 1547. The AIC of the null model

```
|| suvrNullModel = polr(class ~ 1, data=suvrData)
|| summary(suvrNullModel)
```

is 1566. Thus, our model is an improvement.

The model predicts that Eliska is a Junior (44.7%) or a Senior (36.5%):

```
|| eliska = data.frame(gender="Female", gpa=3.33)
|| predict(suvrModel, eliska, type="prob")
```

Here are the (abbreviated) raw results

```
|| Non-Mat    Fresh    Soph    Junior    Senior
|| 0.00211    0.02356  0.16293  0.44656  0.36483
```

Thus, we do have an estimate for Eliska's class level, but there is a second option which is rather close. I'm not sure I would bet any money on where to put Eliska.

Regardless, it is highly unlikely for Eliska to be either non-matriculated or a Freshman. Those probabilities, while non-zero, are *very* low. ♦



Warning: *Beware! Remember that the data are not representative of the population. The distribution of the classes is quite similar to the probabilities predicted for Eliska. This is not surprising. The effect of the independent variables on the dependent variable are not statistically significant. Thus, these probabilities are essentially the relative proportions of the classes in the sample.*

GRAPHIC: CLASS AGAINST GPA: As for a graphic, we need our dependent variable to be the probability of each class. Since there is only one numeric independent variable, GPA, that will be the variable we graph along the x-axis.

The ultimate question is: What do we do with the gender variable?

One option is to plot the effect of gender on the same graphic. That means we will have 5×2 curves on the same plot (the number of levels by the number of genders recorded). That *may be* problematic as it may overwhelm the graphic. Figure 15.3 is this figure. Note that it does allow us to compare everything at once. However, you may find it overwhelming... or not.

For the higher GPA values, it is most likely that the student is a Junior, regardless of the gender. At no place is it likely the student is either a Freshman or non-matriculated. This is supported by the data, as the number of non-matriculated students is just 2 and the number of Freshman is just 22 — out of a sample size of $n = 661$.

We can also use this graphic to estimate the various probabilities for Eliska. Remember she has a GPA of 3.33. Since Eliska is female, we look at the dashed lines. Going to 3.33 on the x-axis and move vertically, we see that Eliska is most likely a member of the cyan level — Junior — with a close second being the magenta level — Senior. This conclusion agrees with our prediction above.

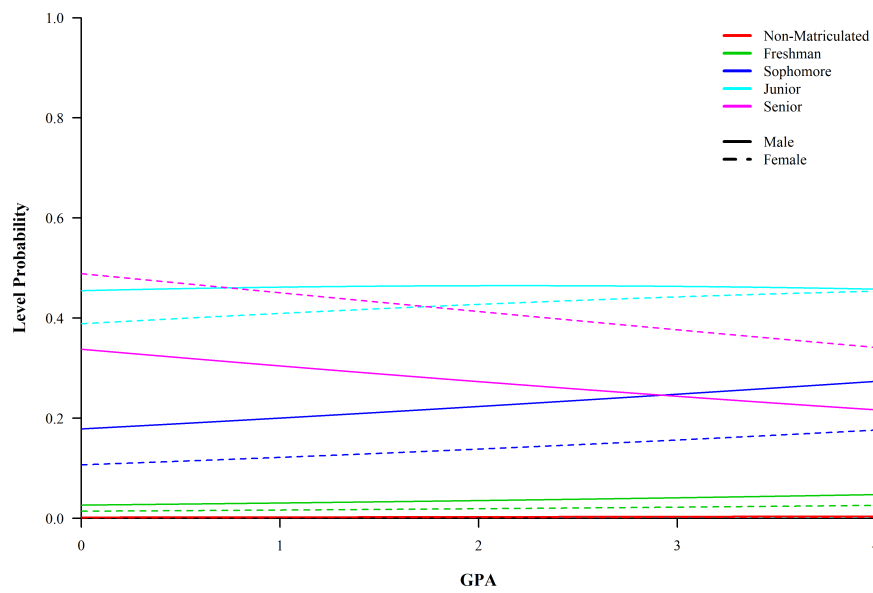


Figure 15.3: Graphic of the probability for each class level for each gender. Note that the non-matriculated and the Freshman levels uniformly have low probability. This is due to the nature of the data; only 2 non-matriculated and 22 Freshmen are in the sample of size $n = 661$. This limits what we can say about the population, unless the level distributions are similar to the population.

15.5: Conclusion

In this chapter, we examined the special issues behind fitting dependent variables that are either nominal or ordinal. Nominal dependent variables are still basically fit with a series of logistic (or other link) regressions. The alteration comes about because we need to keep the same base category throughout in order to make our results comparable.

The ordinal dependent variable can be fit using a technique similar to the previous chapter: fit an underlying linear function, then create thresholds to divide a constant function into an ordinal response.

In both cases, predictions in \mathbb{R} follow the typical structure, with the addition of being able to just predict the outcome category or being able to predict the probabilities associated with the case fitting in each bin.

15.6: End-of-Chapter Materials

15.6.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

RFS This is a “book package,” that is not yet complete. In lieu of installing this package and loading it with `library(RFS)`, you will activate all of its important parts by running

```
source("http://rfs.kvasaheim.com/rfs.R").
```

MASS This package is also a “book package,” a package created for a specific book. Here, that book is “Modern Applied Statistics with S.”

nnet This package contains many functions dealing with neural networks. For this chapter, we use it to fit multinomial models.

STATISTICS:

multinom() This modeling function allows you to fit nominal dependent variables. Its structure is standard in that its main argument is the formula. In order to use the `multinom` function, you must load the `nnet` library.

polr() This modeling function allows you to fit ordinal dependent variables when there is an underlying linear function that drives the process. In order to use the `polr` function, you must load the `MASS` package.

predict(model, newdata) As with almost all statistical packages, R has a `predict` function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

set.base() This allows one to change the base category from which all other levels are estimated. It is a member of the `RFS` package.

15.6.2 EXERCISES This section offers suggestions on things you can practice from this chapter.

1. In Section 15.1.1, we fit a multinomial model to the `gssocc` data. The base used was 'Blue Collar.' Refit the model using 'Craft' as the base category.
2. Determine the AIC of the null model in Section 15.1.1.
3. As mentioned in Section 15.2.1, calculate the relative accuracy of the model of Example 15.2.1.
4. As mentioned in Section 15.2.1, add a quadratic education term to the model of Example 15.2.1 to see if both the highly educated and the lesser educated both support the president.

15.6.3 APPLIED READINGS

- Paul D. Allison and Nicholas A. Christakis. (1994) “Logit Models for Sets of Ranked Items.” *Sociological Methodology* 24: 199–228.
- John Fox and Robert Andersen. (2006) “Effect Displays for Multinomial and Proportional-Odds Logit Models.” *Sociological Methodology* 36: 225–55.
- Daniel Carson Johnson. (1997) “Formal Education vs. Religious Belief: Soliciting New Evidence with Multinomial Logit Modeling.” *Journal for the Scientific Study of Religion* 36(2): 231–46.
- Mark R. Killingsworth and Cordelia W. Reimers. (1983) “Race, Ranking, Promotions, and Pay at a Federal Facility: A Logit Analysis.” *Industrial and Labor Relations Review* 37(1): 92–107.
- Alan B. Lowther and John R. Skalski. (1998) “A Multinomial Likelihood Model for Estimating Survival Probabilities and Overwintering for Fall Chinook Salmon Using Release: Recapture Methods.” *Journal of Agricultural, Biological, and Environmental Statistics* 3(2): 223–36.
- Christopher Winship and Robert D. Mare. (1984) “Regression Models with Ordinal Variables.” *American Sociological Review* 49(4): 512–25.
- Judith E. Zeh, Daijin Ko, Bruce D. Krogman and Ronald Sonntag. (1986) “A Multinomial Model for Estimating the Size of a Whale Population from Incomplete Census Data.” *Biometrics* 42(1): 1–14.

15.6.4 THEORY READINGS

- B. R. Bhat and N. V. Kulkarni. (1966) "On Efficient Multinomial Estimation." *Journal of the Royal Statistical Society. Series B (Methodological)* 28(1): 45–52.
- Zhen Chen and Lynn Kuo. (2001) "A Note on the Estimation of the Multinomial Logit Model with Random Effects." *The American Statistician* 55(2): 89–95.
- Jean-Yves Dauxois and Syed N. U. A. Kirmani. (2003) "Testing the Proportional Odds Model under Random Censoring." *Biometrika* 90(4): 913–22.
- Byung Soo Kim and Barry H. Margolin. (1992) "Testing Goodness of Fit of a Multinomial Model Against Overdispersed Alternatives." *Biometrics* 48(3): 711–19.
- Bercedis Peterson and Frank E. Harrell, Jr. (1990) "Partial Proportional Odds Models for Ordinal Response Variables." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 39(2): 205–17.
- G. A. F. Seber and S. O. Nyangoma. (1974) "Residuals for Multinomial Models." *Biometrika* 87(1): 183–91.
- M. Stone. (1974) "Cross-Validation and Multinomial Prediction." *Biometrika* 61(3): 509–15.
- Y. K. Tse. (1987) "A Diagnostic Test for the Multinomial Logit Model." *Journal of Business & Economic Statistics* 5(2): 283–86.

Part IV

The Appendices

Appendix M	The Appendix of Matrices	475
Appendix R	Experimenting with R	507
Appendix S	The Appendix of Statistics	533

APPENDIX M:

THE APPENDIX OF MATRICES

OVERVIEW:

Before we can start talking about regression, we need to cover the necessary mathematical background. For this book, I expect that you have successfully complete a course in matrix algebra.

This means that you can add and multiply matrices, that you understand matrices are linear transformations, that you have worked with eigenvalues and eigenvectors, and that you can calculate the rank and the trace of a matrix.

All of these topics are important in better understanding the mathematics underlying linear models. Thus, this appendix reviews some of the important parts from your matrix algebra class, and adds some items that you may not have had.

Chapter Contents

M.1	Matrix Basics	476
M.2	Addition	479
M.3	Multiplication	481
M.4	Other Matrix Terms and Operations	493
M.5	Consequences	499
M.6	Statistics in Matrices	502
M.7	End-of-Appendix Materials	506



The purpose of this appendix is to provide the necessary matrix background for this book. Everything here is important at some point in this test. There will be new things here, in which case you need to learn them. If nothing here is new, then you merely need to review them to keep them fresh in your mind.

You should treat this appendix as necessary background knowledge. Gain it now, if necessary. At the very least, know what is here so that you can refer to it as you work through the main part of the text.

M.1: Matrix Basics

matrix

A matrix is just a rectangular array of scalars. It is used to simplify many mathematical calculations. Throughout this book, I will use it in such a sense. The following is an example of a matrix:

$$\mathbf{A} = \begin{bmatrix} 3 & 5 & 2 \\ a & 1 & 18 \end{bmatrix} \tag{M.1}$$

size

Because a matrix is a *rectangular* array, it has a dimension. The matrix \mathbf{A} above has dimension 2×3 because there are 2 rows and 3 columns. We could also write this as

$$\mathbf{A} \in \mathcal{M}_{2 \times 3} \tag{M.2}$$

This can be read as “ \mathbf{A} is a matrix with dimension 2×3 ” or as “ \mathbf{A} is an element of the set of 2×3 matrices.” Note that the symbol \in means “is an element of” and $\mathcal{M}_{2 \times 3}$ is “the set of all matrices of dimension 2×3 .”

Also note that the dimension order is very important and is always written as “rows \times columns.” $\mathcal{M}_{2 \times 3}$ and $\mathcal{M}_{3 \times 2}$ are entirely different sets of matrices.

square

A matrix is square if the number of rows equals the number of columns. That is, \mathbf{B} is square if

$$\mathbf{B} \in \mathcal{M}_{n \times n} \tag{M.3}$$

for some number $n \in \mathbb{Z}^+$. If a matrix is square, the set is often denoted simply by \mathcal{M}_n . The matrix \mathbf{A} above is not square because the number of rows does not equal the number of columns.

Note: In applied statistics, think of the matrix as being the thing that contains your data in the computer. The columns of a matrix represent the variables (and their values). The rows represent the records (e.g., people). Thus, for an applied statistician, an $n \times p$ matrix represents p variables measured for a sample of size n .

M.1.1 REPRESENTATION The next sections cover the algebra of matrices. To ease the notation, let me show you two ways of representing matrices. First, here is matrix \mathbf{A} written out.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1c} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2c} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & a_{r3} & \cdots & a_{rc} \end{bmatrix} \quad (\text{M.4})$$

Note that the subscripts can also use a comma to separate the values. That is only done, however, when you get to double digits and ambiguity ensues. For instance, does a_{242} represent $a_{2,42}$ or $a_{24,2}$? Perhaps it actually represents $a_{2,4,2}$ in a tensor. Who knows when ambiguity ensues?

Note: When we stick to single digits for the indices, commas are usually dropped.

Note that every element in the \mathbf{A} matrix is represented by a lowercase a and its r, c position in the matrix. This allows us to simplify representation at times:

$$\mathbf{A} = [a_{ij}] \quad (\text{M.5})$$

Here, i is the row index and j is the column index.

Example 1

Create the data matrix where the first column contains heights in inches and the second column contains favorite number. Measure these two variables on three people.

Solution: My answer will probably differ from yours. Here is the matrix I obtained:

$$\mathbf{A} = \begin{bmatrix} 65 & 1 \\ 75 & 7 \\ 73 & 7 \end{bmatrix} \quad (\text{M.6})$$

Note that each row is a record corresponding to a different person. Thus, the first person I asked was 65 inches tall and had a favorite number of 1.

Assuming I selected a random sample, the row-order does not matter. Thus, this is the same data matrix:

$$\mathbf{A} = \begin{bmatrix} 75 & 7 \\ 65 & 1 \\ 73 & 7 \end{bmatrix} \quad (\text{M.7})$$



M.2: Addition

Matrix addition is closed. This means that the sum of two matrices will always give you another matrix... as long as it makes sense to add two matrices. Two matrices can be added if they have the same dimension.¹

Let $\mathbf{A} \in \mathcal{M}_{r \times c}$ and $\mathbf{B} \in \mathcal{M}_{r \times c}$ for some values of r and c . \mathbf{A} and \mathbf{B} are commensurate and can be summed. Matrix addition is done element-by-element (element-wise) addition. Thus,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} & \cdots & a_{1c} + b_{1c} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} & \cdots & a_{2c} + b_{2c} \\ a_{31} + b_{31} & a_{32} + b_{32} & a_{33} + b_{33} & \cdots & a_{3c} + b_{3c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{r1} + b_{r1} & a_{r2} + b_{r2} & a_{r3} + b_{r3} & \cdots & a_{rc} + b_{rc} \end{bmatrix} \quad (\text{M.8})$$

This can also be symbolized (shortened) as

$$\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}] \quad (\text{M.9})$$

Matrix addition has a zero (additive identity). It is the commensurate matrix with all elements equal to zero:

$$\mathbf{0} = [0_{ij}] \quad (\text{M.10})$$

How does it work in addition? Just as you would expect:

$$\mathbf{A} + \mathbf{0} = \begin{bmatrix} a_{11} + 0 & a_{12} + 0 & a_{13} + 0 & \cdots & a_{1c} + 0 \\ a_{21} + 0 & a_{22} + 0 & a_{23} + 0 & \cdots & a_{2c} + 0 \\ a_{31} + 0 & a_{32} + 0 & a_{33} + 0 & \cdots & a_{3c} + 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{r1} + 0 & a_{r2} + 0 & a_{r3} + 0 & \cdots & a_{rc} + 0 \end{bmatrix} = \mathbf{A} \quad (\text{M.11})$$

This can also be symbolized (shortened) as

$$\mathbf{A} + \mathbf{0} = [a_{ij} + 0] = [a_{ij}] \quad (\text{M.12})$$

I leave it as an exercise to prove $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$.

Matrices also have an additive inverse. As with scalar arithmetic, a matrix plus its additive inverse equals the zero matrix; that is, if \mathbf{B} is the additive inverse of \mathbf{A} , then $\mathbf{A} + \mathbf{B} = \mathbf{0}$.

Two things about the additive inverse: First, it is commensurate with the original matrix. Second, it is unique (just as in scalar arithmetic).

¹When the matrices have the correct dimension to perform the mathematical operation, they are called “commensurate.” For addition, commensurate matrices have the same dimension. For multiplication, the requirement is much different (see M.3).

elementwise

zero

exercise

To calculate the additive inverse of \mathbf{A} , just negate each element of \mathbf{A} . Thus, if $\mathbf{B} = [-a_{ij}]$ then \mathbf{B} is the additive inverse of \mathbf{A} .

Finally, as with all elementwise operations, matrix addition is both commutative and associative:

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

Note: Thus, in conclusion matrix addition behaves like scalar addition, as long as the matrices are commensurate.

M.3: Multiplication

There are many, many, *many* types of multiplication with matrices. The one selected depends on the intention (the need). They include scalar product, matrix product, Hadamard product (a.k.a. Schur product), and Kronecker product. As only the first two are typically seen in an undergraduate linear models course, we will only discuss those two here.

M.3.1 SCALAR PRODUCT As in arithmetic, the scalar product arose from needing to repeatedly add a matrix to itself. Thus, instead of writing $\mathbf{A}+\mathbf{A}+\mathbf{A}+\mathbf{A}+\mathbf{A}+\mathbf{A}+\mathbf{A}+\mathbf{A}$, one could write $8\mathbf{A}$, where 8 is a scalar. This was quickly generalized to non-integer values for the scalar multiple, just as $3 \times a$ was quickly generalized to things like $4.25 \times a$.

Scalar multiplication is defined as

$$c\mathbf{A} = [ca_{ij}] \quad (\text{M.13})$$

Scalar products are commutative. That is, if c is a scalar and \mathbf{A} is a matrix, then $c\mathbf{A} = \mathbf{A}c$. This will come in handy later, so be aware of it. Note that c does not need to be a natural number.

commutative

Scalar products are also associative. That is, if c is a scalar, then the following are equivalent:

associative

- $c\mathbf{A}\mathbf{B}$
- $(c\mathbf{A})\mathbf{B}$
- $c(\mathbf{A}\mathbf{B})$
- $\mathbf{A}c\mathbf{B} = \mathbf{A}\mathbf{B}c$

Scalar multiplication is also distributive over matrix addition. Thus:

distributive

$$c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B} \quad (\text{M.14})$$

$$\begin{matrix} \mathbf{A} & \mathbf{B} & = & \mathbf{AB} \\ r_1 \times c_1 & r_2 \times c_2 & & r_1 \times c_2 \\ & \text{same} & & \end{matrix}$$

Figure M.1: A schematic designed to illustrate commensurability with matrix multiplication. Note that the “inner” dimensions of the factors must be equal and that the product dimension is the “outers” of the two factors.

M.3.2 MATRIX PRODUCT The matrix product is the multiplication that is (usually) meant when one just says “matrix multiplication.” Its definition arises from linear algebra and repeated linear transformations. It has many nice properties. Calculation is not one of them.

Let us define two matrices **A** and **B** such that the number of columns of **A** equals the number of rows of **B**. Their product is defined as

$$\mathbf{AB} = [ab_{ij}] = \left[\sum_k a_{ik} b_{kj} \right] \quad (\text{M.15})$$

where k ranges between 1 and the number of columns of **A**. The dimension of the product is the number of rows of **A** by the number of columns of **B**. That is, let $\mathbf{A} \in \mathcal{M}_{r_1 \times c_1}$ and $\mathbf{B} \in \mathcal{M}_{r_2 \times c_2}$. Then, one can multiply **A** and **B** if $c_1 = r_2$. The dimension of the product is $r_1 \times c_2$. Figure M.1 illustrates this.

Example 2

Here is an example of matrix multiplication. Let us define our two matrices as

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad (\text{M.16})$$

and

$$\mathbf{B} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & m \end{bmatrix} \quad (\text{M.17})$$

Let us find the product **AB**.

commensurate

Solution: The first step is to check that multiplying these matrices can be done. Do the number of columns of \mathbf{A} equal the number of rows of \mathbf{B} ? Note that $\mathbf{A} \in \mathcal{M}_{2 \times 3}$ and $\mathbf{B} \in \mathcal{M}_{3 \times 3}$. Because the “inners” of \mathbf{AB} are equal to each other, the matrix multiplication \mathbf{AB} makes sense.

Second, we determine the dimension of the product. It is the number of rows of \mathbf{A} by the number of columns of \mathbf{B} : 2×3 , the “outers.”

Third, since we know the dimension of the product, we just have to fill in the blanks in the matrix:

$$\mathbf{AB} = \begin{bmatrix} - & - & - \\ - & - & - \end{bmatrix} \quad (\text{M.18})$$

The top-right element in the product matrix is element 1, 1. Thus, by our definition, it equals

$$ab_{11} = \left[\sum_k a_{1k} b_{k1} \right] \quad (\text{M.19})$$

$$= [a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}] \quad (\text{M.20})$$

$$= [1a + 2d + 3g] \quad (\text{M.21})$$

The top-middle element, ab_{12} , is

$$ab_{12} = \left[\sum_k a_{1k} b_{k2} \right] \quad (\text{M.22})$$

$$= [a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32}] \quad (\text{M.23})$$

$$= [1b + 2e + 3h] \quad (\text{M.24})$$

Note what is happening here. The elements of the “top-right” cell is the inner product of the top row and the right column. Similarly, the bottom-center element is the inner product of the bottom row and the center column.

inner product

I leave it as an exercise for you to finish the multiplication. Here is the final product:

exercise

$$\mathbf{AB} = \begin{bmatrix} 1a + 2d + 3g & 1b + 2e + 3h & 1c + 2f + 3m \\ 4a + 5d + 6g & 4b + 5e + 6h & 4c + 5f + 6m \end{bmatrix} \quad (\text{M.25})$$



except

In scalar arithmetic, we have a multiplicative identity, multiplicative inverse, and multiplication is commutative, associative, and distributes over addition. All of these hold for matrix multiplication — *except* commutativity. In general, $\mathbf{AB} \neq \mathbf{BA}$, even when both of the multiplications make sense.

identity

The multiplicative identity, which we will symbolize by \mathbf{I} , has the property that $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$. Note that \mathbf{I} is a square matrix.²

inverse

If a multiplicative identity exists, then a multiplicative inverse also exists in this system. The inverse of a matrix \mathbf{A} , denoted \mathbf{A}^{-1} , is a matrix that satisfies these two requirements:

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \quad (\text{M.26})$$

invertible

Not all matrices have inverses. Those that do *not* are called **singular**. Those that *do* are called **invertible**.

full rank

Only square matrices can be invertible. However, not all square matrices *are* invertible. From linear algebra, a matrix is invertible if and only if it is square and is of **full rank**.

determinant

A consequence of this is that a matrix is invertible if its **determinant is non-zero**. In general, the calculation of the determinant and the inverse are computationally intensive. However, they are rather straight-forward for 2×2 matrices.

Let $\mathbf{A} \in \mathcal{M}_{2 \times 2}$. Then, the determinant of \mathbf{A} is $\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21}$. Note that the determinant is a *scalar*, not a matrix.

Example 3

Show that the matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 3 & 0 \end{bmatrix} \quad (\text{M.27})$$

is full rank.

Solution: Later in this appendix, we will see the direct way to find the rank of the matrix. However, since the determinant of \mathbf{A} is not 0 (it is -6), we know \mathbf{A} is full rank. \blacklozenge

²Technically, this statement is only true if \mathbf{A} is square. If it is not square, then the two \mathbf{I} matrices will have different dimension. We will restrict ourselves to square matrices. The “generalized inverse” is beyond the scope of this text.

Lemma M.1 (The 2×2 Inverse). *Let \mathbf{A} be a 2×2 invertible matrix. The inverse of \mathbf{A} is*

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \quad (\text{M.28})$$

Proof. To prove this, we will show $\mathbf{AA}^{-1} = \mathbf{I}$ and $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

$$\mathbf{AA}^{-1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \quad (\text{M.29})$$

$$= \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \quad (\text{M.30})$$

$$= \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & -a_{11}a_{12} + a_{11}a_{12} \\ a_{21}a_{22} - a_{21}a_{22} & -a_{12}a_{21} + a_{11}a_{22} \end{bmatrix} \quad (\text{M.31})$$

$$= \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & -a_{12}a_{21} + a_{11}a_{22} \end{bmatrix} \quad (\text{M.32})$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (\text{M.33})$$

$$= \mathbf{I} \quad (\text{M.34})$$

Thus, we have shown that $\mathbf{AA}^{-1} = \mathbf{I}$. This is one-half of the proof. I leave it as an exercise for you to prove the second half: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. \square

exercise

Note: The reason I provide the mathematics for 2×2 matrices is that in our study of simple linear regression, many of the important calculations will be done with 2×2 matrices. (See Chapter 3.)

Note: Formulas exist for 3×3 matrices, too. However, once we move beyond that, hand calculations are time-prohibitive. At the end of this section, I provide some examples of performing these calculations in \mathbb{R} .

Now that we have a mechanism to calculate a multiplicative inverse, let us see that not all square matrices have one.

Example 4

Let $\mathbf{A} \in \mathcal{M}_{2 \times 2}$ be defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \quad (\text{M.35})$$

Calculate \mathbf{A}^{-1} .

Solution: Let us calculate \mathbf{A}^{-1} using our formula,

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \quad (\text{M.36})$$

Applying this formula is straight-forward:

$$\mathbf{A}^{-1} = \frac{1}{0} \begin{bmatrix} 6 & -3 \\ -2 & 1 \end{bmatrix} \quad (\text{M.37})$$

Yes, the determinant of \mathbf{A} is $\det \mathbf{A} = 1 \times 6 - 3 \times 2 = 6 - 6 = 0$. Since the determinant is 0, the inverse does not exist (one cannot divide by 0). \blacklozenge

singular

Note: What is it about the \mathbf{A} matrix that makes it singular? Note that the second column is just 3 times the first column (or the second row is twice the first). This means the matrix is not full rank. The columns are not linearly independent.

independent

When we get to using these matrices with real data, we will interpret this as the second column giving us no knowledge about the world that is not already contained in the first column. The second column is **redundant information**.

redundant

M.3.3 RANK Now that we have mentioned the rank of a matrix a few times, let us formally define it.

Definition M.2. Rank

The rank of a matrix is the greatest number of columns that are linearly independent.

So, to calculate rank, one must know what it means for a set of column vectors (a matrix) to be linearly independent.

Definition M.3. Linearly Independent

A set of column vectors $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r$ are linearly independent if

$$\sum_{i=1}^r a_i \mathbf{c}_i = \mathbf{0} \quad (\text{M.38})$$

implies $a_1 = a_2 = \dots = a_r = 0$.

Example 5

Show that the rank of \mathbf{A} is 2.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \\ 9 & 3 & 12 \end{bmatrix}$$

Solution: To show that \mathbf{A} has rank of 2, we first show it has rank of at most 2, then show it has rank of *at least* 2.

Part 1. Note that the third column is the sum of the first two columns. Thus,

$$1\mathbf{c}_1 + -1\mathbf{c}_2 + -1\mathbf{c}_3 = \mathbf{0} \quad (\text{M.39})$$

Since the coefficients $\{1, -1, -1\}$ are not all 0, we have shown that the rank of \mathbf{A} is at most 2.

Part 2. To show that the rank of \mathbf{A} is *at least* 2, we can find a submatrix that has a non-zero determinant:

$$\mathbf{A}_c = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$$

The determinant of \mathbf{A}_c is 1. Since this is not 0, we know that \mathbf{A} has rank of at least 2.

Conclusion. Combining these two results proves that \mathbf{A} has rank of 2. ♦

Note: Technically, you calculated the *column* rank in the previous example. However, one important result from linear algebra is that the column rank, the row rank, and the rank are all equal.

I leave it as an exercise for you to determine the values in the coefficient vector $\{a_1, a_2, a_3\}$ to show that the row rank is at most 2.

So far, we have seen the multiplicative identity and the multiplicative inverse. It is time to note that matrix multiplication is not commutative.

Theorem M.3.1

Matrix multiplication is not commutative. That is, there exist matrices **A** and **B** such that $\mathbf{AB} \neq \mathbf{BA}$.

Note: If **A** and **B** are not square with the same dimension, then this statement is trivially true.

Proof. The proof is simple. Since we need to prove existence, we simply need to provide a counter-example. To wit, let

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 7 \end{bmatrix} \tag{M.40}$$

and

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{M.41}$$

Note that $\mathbf{AB} \neq \mathbf{BA}$.

Since we have found a counter-example, we have shown that matrix multiplication is not commutative, in general. \square

Note: While technically correct, this proof leaves us feeling a little empty. So we found one counter-example. Cool beans. But we learned precious little about commutativity with matrix multiplication. Let us explore a bit and see if we can determine *when* (under what conditions) matrix multiplication is commutative.

In doing this, we may learn something interesting.

First, let us assume **A** and **B** are square and commensurate. This ensures that **AB** and **BA** can be calculated. For instance, let both be 2×2 matrices. Then, **AB** is

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \quad (\text{M.42})$$

$$= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \quad (\text{M.43})$$

and **BA** is

$$\mathbf{BA} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad (\text{M.44})$$

$$= \begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & b_{11}a_{12} + b_{12}a_{22} \\ b_{21}a_{11} + b_{22}a_{21} & b_{21}a_{12} + b_{22}a_{22} \end{bmatrix} \quad (\text{M.45})$$

$$= \begin{bmatrix} a_{11}b_{11} + a_{21}b_{12} & a_{12}b_{11} + a_{22}b_{12} \\ a_{11}b_{21} + a_{21}b_{22} & a_{12}b_{21} + a_{22}b_{22} \end{bmatrix} \quad (\text{M.46})$$

By comparing the two product matrices, **AB** and **BA**, we can determine an instance where multiplication is commutative.

For instance, if $a_{12} = a_{21} = 0$ and $b_{12} = b_{21} = 0$, then the two product matrices are the same. In other words, if both **A** and **B** are diagonal matrices, multiplication will be commutative. That's an interesting consequence we would have missed if we just stopped with our proof.

In fact, it can be proven that multiplication of diagonal matrices is commutative in general. Kewl!

Cool beans!

Theorem M.3.2

Let **A** and **B** be diagonal matrices of the same size. The product is commutative; that is, $\mathbf{AB} = \mathbf{BA}$.

Proof. Since **A** and **B** are diagonal and of the same shape, they can be represented as

$$\mathbf{A} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n \end{bmatrix} \quad (\text{M.47})$$

and

$$\mathbf{B} = \begin{bmatrix} b_1 & 0 & 0 & \cdots & 0 \\ 0 & b_2 & 0 & \cdots & 0 \\ 0 & 0 & b_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b_n \end{bmatrix} \quad (\text{M.48})$$

Their product is

$$\mathbf{AB} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 & 0 & 0 & \cdots & 0 \\ 0 & b_2 & 0 & \cdots & 0 \\ 0 & 0 & b_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b_n \end{bmatrix} \quad (\text{M.49})$$

$$= \begin{bmatrix} a_1 b_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 b_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 b_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n b_n \end{bmatrix} \quad (\text{M.50})$$

Similarly, the product \mathbf{BA} is

$$\mathbf{BA} = \begin{bmatrix} b_1 & 0 & 0 & \cdots & 0 \\ 0 & b_2 & 0 & \cdots & 0 \\ 0 & 0 & b_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b_n \end{bmatrix} \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n \end{bmatrix} \quad (\text{M.51})$$

$$= \begin{bmatrix} b_1 a_1 & 0 & 0 & \cdots & 0 \\ 0 & b_2 a_2 & 0 & \cdots & 0 \\ 0 & 0 & b_3 a_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b_n a_n \end{bmatrix} \quad (\text{M.52})$$

Since scalar multiplication (in the cells) is commutative, we have

$$= \begin{bmatrix} a_1 b_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 b_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 b_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n b_n \end{bmatrix} \quad (\text{M.53})$$

$$= \mathbf{AB} \quad (\text{M.54})$$

Thus, we have shown $\mathbf{AB} = \mathbf{BA}$ for two diagonal matrices of the same size. \square



exercise

associative

Finally, I leave it as an exercise for you to prove that matrix multiplication is **associative** (when the multiplication can be done). That is, if \mathbf{ABC} can be calculated, then it can be calculated as either $(\mathbf{AB})\mathbf{C}$ or as $\mathbf{A}(\mathbf{BC})$.

M.4: Other Matrix Terms and Operations

There exist other helpful operations on matrices. Already, we have come across the determinant as being especially helpful in determining if a matrix is invertible or singular.

Another useful function is the **trace**. It is just the sum of the diagonal elements. That is:

$$\text{tr } \mathbf{A} = \sum_{i=1}^r a_{ii} \quad (\text{M.55})$$

trace

The formula is simple... deceptively so. In fact, one may wonder what the trace actually tells us about a matrix. Well, in general, you will need to revisit your matrix algebra class notes. In linear models, however, the trace is used to calculate the number of degrees of freedom (see Section 3.2).

The **transpose** of a matrix is just the matrix where the rows and columns are switched. Thus, if \mathbf{B} is the transpose of \mathbf{A} , then $b_{ij} = a_{ji}$. In symbols, we indicate the transpose as

$$\mathbf{B} = \mathbf{A}' \quad (\text{M.56})$$

transpose

A matrix is **symmetric** if it is equal to its transpose, $\mathbf{A} = \mathbf{A}'$.

symmetric

$$\begin{bmatrix} a_{ij} \end{bmatrix} = \begin{bmatrix} a_{ji} \end{bmatrix} \quad (\text{M.57})$$

Note that only square matrices can be symmetric. Symmetric matrices have some nice properties with respect to calculations.

In all cases, $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ both exist and are symmetric. Furthermore,

$$\text{rank } \mathbf{A}'\mathbf{A} = \text{rank } \mathbf{A}\mathbf{A}' \quad (\text{M.58})$$

Also note that one can “**symmetrize**” any square matrix. That is, one can form symmetric matrix \mathbf{X} from a square matrix \mathbf{A} as

symmetrize

$$\mathbf{X} = \frac{\mathbf{A} + \mathbf{A}'}{2} \quad (\text{M.59})$$

I leave it as an exercise to prove that \mathbf{X} is symmetric.

exercise

One important feature of symmetric matrices is that they can be transformed into a diagonal matrix. If \mathbf{A} is symmetric, then there exists a \mathbf{Q} such that \mathbf{AQ} is diagonal. Why is this helpful? First, remember that multiplication of diagonal matrices is commutative. Second, as you will see in the text, diagonal covariance matrices indicate independence.

This means that any set of variables can be linearly transformed into a set of independent variables. This fact is the basis for a procedure called “principal component analysis.”

j vector

The \mathbf{j} vector is a vector of 1s. It is used to calculate row sums (if pre-multiplying) or column sums (if post-multiplying). The matrix \mathbf{J} is a matrix of 1s. It does what \mathbf{j} does, but puts the sums in a matrix.

e_i vector

The \mathbf{e}_i vector is a vector of 0s, with a 1 in the i^{th} position. It is used in proofs, as it can be used to select an individual row, column, or element of a matrix.

Example 6

What are \mathbf{j}_2 , \mathbf{J}_2 , and \mathbf{e}_1 ?

Solution: The first two are

$$\mathbf{j}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (\text{M.60})$$

and

$$\mathbf{J}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (\text{M.61})$$

The third depends on context for the actual length of the vector. Assuming we are working in three dimensions, then we have

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (\text{M.62})$$



The **eigenvalues** of a matrix \mathbf{A} are those values ξ that solve the equation $\mathbf{A}\mathbf{v} = \xi\mathbf{v}$. The vectors \mathbf{v} corresponding to each of the eigenvalues are called the **eigenvectors**.

Recall from your linear algebra course that the determinant of \mathbf{A} is just the product of its eigenvalues. Also, the trace of \mathbf{A} is the sum of the eigenvalues.

Example 7

Determine the eigenvalues of

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} \quad (\text{M.63})$$

Solution: The trace and determinant of \mathbf{A} are 7 and 10, respectively. Thus, if ξ_1 and ξ_2 are the two eigenvalues, we can solve the system of equations

$$\begin{cases} \xi_1 + \xi_2 = 7 \\ \xi_1 \times \xi_2 = 10 \end{cases} \quad (\text{M.64})$$

Thus, we see that the two eigenvalues are 2 and 5. \blacklozenge

idempotent

A matrix \mathbf{A} is **idempotent** if $\mathbf{A}\mathbf{A} = \mathbf{A}$. The trace of an idempotent matrix equals its rank. This is rather important in studying linear models, since the rank is also the degrees of freedom.

Example 8

Show that this matrix is idempotent

$$\mathbf{A} = \begin{bmatrix} 4 & 6 \\ -2 & -3 \end{bmatrix} \quad (\text{M.65})$$

Solution: Because I already have R open, let's use it to show it is idempotent:

```
|| A = matrix( c(4,-2,6,-3), ncol=2)
|| A %*% A
```

The output is the same as the matrix \mathbf{A} , so we have shown it is idempotent.

Now that we have shown \mathbf{A} is idempotent, what is its determinant?

```
|| det A
```

Thus, \mathbf{A} is rank deficient and is singular (non-invertible). ♦

Note: Since determinants multiply, then for \mathbf{A} to be idempotent, it must either have determinant 0 or 1.

Why?

Matrices \mathbf{A} and \mathbf{B} are orthogonal, $\mathbf{A} \perp \mathbf{B}$, if $\mathbf{A}'\mathbf{B} = \mathbf{0}$. That is, if the inner products of the columns of \mathbf{A} and \mathbf{B} are orthogonal, then the matrices themselves are orthogonal.

A matrix \mathbf{P} is a projection matrix if it is idempotent. The purpose of projection matrices is to project a higher space onto a subspace. If \mathbf{P} is also symmetric, it is called an orthogonal projection matrix. This means it projects the larger space orthogonally (perpendicularly) onto the subspace. Think of shining a flashlight on a plant. If you put the flashlight directly over the plant, it will project the plant orthogonally onto the floor. If you do it at an angle, then the projection is called oblique.

The key in both instances is that you are simplifying a complicated reality (3-D object) onto a simpler model (2-D shadow).

projection matrix

oblique projection

M.4.1 POSITIVE DEFINITE MATRICES This is an important section, so let us start off with a definition.

Definition M.4. *Positive Definite*

A matrix \mathbf{A} is a **positive definite** (pd) if $\mathbf{q}'\mathbf{A}\mathbf{q} > 0$ for all non-zero vectors \mathbf{q} .

It is usually difficult to determine if a matrix is positive definite (pd). However, once you know it is, there are some important properties, which we look at in the next section (Sections M.5 and M.5.1).

positive definite

Example 9

Determine if this matrix is positive definite (pd).

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \quad (\text{M.66})$$

Solution: Often, it is easier to prove it is not. So, a few quick checks.

- Are the diagonal elements all positive?
If so, then it may be pd. Otherwise, it cannot be.
- Is the determinant positive?
If so, then it may be pd. Otherwise, it cannot be.

Since \mathbf{A} passes the first check, but not the second, it is not pd.

Directly applying the definition is also an option. In fact, it is the only option if \mathbf{A} passes all easy checks.

$$\mathbf{q}'\mathbf{A}\mathbf{q} = \begin{bmatrix} q_1 & q_2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \quad (\text{M.67})$$

$$= \begin{bmatrix} 2q_1 + 2q_2 & 3q_1 + 1q_2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \quad (\text{M.68})$$

$$= (2q_1 + 2q_2)q_1 + (3q_1 + 1q_2)q_2 \quad (\text{M.69})$$

$$= 2q_1^2 + 5q_1q_2 + q_2^2 \quad (\text{M.70})$$

This is not necessarily positive. For instance, if $q_1 = 1$ and $q_2 = -1$, then the polynomial equals -2 .

Thus, we have directly shown that \mathbf{A} is not pd; that is, we have determined a vector \mathbf{q} such that $\mathbf{q}'\mathbf{A}\mathbf{q} \leq 0$. That matrix is

$$\mathbf{q} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (\text{M.71})$$

In fact, we have shown that it is not even positive *semi*-definite (psd). ♦

Definition M.5. Positive Semi-Definite

We say that a matrix is **positive semi-definite** (psd) if $\mathbf{q}'\mathbf{A}\mathbf{q} \geq 0$ for all non-zero vectors \mathbf{q} .

M.5: Consequences

With these definitions, there are a lot of consequences. Many of which are important in the study of linear models. This section covers many of them.

First, when taking the transpose of a product, you switch the order of the multiplication: $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. A similar result holds with inverses. The only difference is that all three inverses must exist: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Lemma M.6. *For any matrix \mathbf{X} , the matrix $\mathbf{X}'\mathbf{X}$ is symmetric.*

If you know the determinant of a matrix, you can easily calculate the determinant of a scalar multiple of that matrix.

Lemma M.7. *If $c \in \mathbb{R}$ and $\mathbf{A} \in \mathcal{M}_n$, then $\det c\mathbf{A} = c^n \det \mathbf{A}$.*

There is a similar result with the trace of a matrix.

Lemma M.8. *If $c \in \mathbb{R}$ and $\mathbf{A} \in \mathcal{M}_n$, then $\text{tr } c\mathbf{A} = c \text{tr } \mathbf{A}$.*

M.5.1 POSITIVE DEFINITE MATRICES Positive definite matrices are very important in applied statistics. So, let us break these results into a separate subsection.

Lemma M.9. *The diagonal elements of a pd matrix are all positive. That is, let $\mathbf{A} \in \mathcal{M}_n$ be positive definite, then $a_{ii} > 0, \forall i \in \{1, 2, \dots, n\}$.*

This can easily be shown by letting the \mathbf{q} vector be \mathbf{e}_i . The quadratic form $\mathbf{q}'\mathbf{A}\mathbf{q}$ would therefore equal the diagonal element at position i . Since \mathbf{A} is positive definite, that element must be greater than 0.

Note that the converse is not true. Just because the diagonal elements are all positive does not mean that the matrix is positive definite. For an example, note

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{M.72}$$

is not positive definite. To see this, let \mathbf{q}' be the vector $[1 \ -1]$. Then

$$\mathbf{q}'\mathbf{A}\mathbf{q} = [1 \ -1] \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{M.73}$$

$$= [0 \ 0] \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{M.74}$$

$$= 0 \tag{M.75}$$

Since this is not greater than 0, \mathbf{A} is not positive definite.

Note that the determinant of this \mathbf{A} is 0. This suggests a second consequence: The determinant of a positive definite matrix is positive. This means all pd matrices are invertible. Similarly:

Lemma M.10. *The inverse of a pd matrix is also pd. That is, if \mathbf{A} is positive definite, then so is \mathbf{A}^{-1} .*

thoughts

To see this, ask yourself: What is the determinant of the inverse of a matrix? How can you use that to show that the inverse of a positive definite matrix is also positive definite?

All of the eigenvalues of a pd matrix are positive. And, since all of the diagonal elements of a pd matrix are positive, then the trace is positive.

Lemma M.11. *If \mathbf{X} is a full column rank matrix, even if not square, $\mathbf{X}'\mathbf{X}$ is positive definite.*

How would you prove this? It requires understanding what it means to be full column rank. Sketch out a proof here:



And, *most importantly*, the covariance matrix is positive definite if the design matrix, \mathbf{X} , is full rank. Otherwise, it is positive semi-definite and has a determinant of zero.

covariance

M.6: Statistics in Matrices

In this section, we rewrite some of the equations you learned in your previous statistics course in terms of matrices. This section is useful for matrix practice and for figuring out what the computer is actually doing (it always operates using matrices). In all of the following, let \mathbf{Y} be a column vector of length n .

Lemma M.12. *The sample mean using matrices: $\bar{Y} = \frac{1}{n}\mathbf{j}'\mathbf{Y}$.*

Proof.

$$\frac{1}{n}\mathbf{j}'\mathbf{Y} = \frac{1}{n} [1 \ 1 \ 1 \ \cdots \ 1] \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \quad (\text{M.76})$$

$$= \frac{1}{n} (1Y_1 + 1Y_2 + 1Y_3 + \cdots + 1Y_n) \quad (\text{M.77})$$

$$= \frac{1}{n} \sum_{i=1}^n Y_i \quad (\text{M.78})$$

$$= \bar{Y} \quad (\text{M.79})$$

□

Lemma M.13. *The sum of squared values using matrices: $\mathbf{Y}'\mathbf{Y}$.*

Proof.

$$\mathbf{Y}'\mathbf{Y} = [y_1 \ y_2 \ y_3 \ \cdots \ y_n] \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad (\text{M.80})$$

$$= (y_1 y_1 + y_2 y_2 + y_3 y_3 + \cdots + y_n y_n) \quad (\text{M.81})$$

$$= \sum_{i=1}^n y_i^2 \quad (\text{M.82})$$

□

Lemma M.14. *The sample variance using matrices:*

$$s_y^2 = \frac{1}{n-1} (\mathbf{Y} - \bar{y}\mathbf{j})' (\mathbf{Y} - \bar{y}\mathbf{j}) \quad (\text{M.83})$$

Proof.

$$\frac{1}{n-1} (\mathbf{Y} - \bar{y}\mathbf{j})' (\mathbf{Y} - \bar{y}\mathbf{j}) \quad (\text{M.84})$$

$$= \frac{1}{n-1} [y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}, \dots, y_n - \bar{y},] \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_3 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{bmatrix} \quad (\text{M.85})$$

$$= \frac{1}{n-1} (y_1 - \bar{y})(y_1 - \bar{y}) + (y_2 - \bar{y})(y_2 - \bar{y}) + \cdots + (y_n - \bar{y})(y_n - \bar{y}) \quad (\text{M.86})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{M.87})$$

□

Lemma M.15. *The sample covariance using matrices:*

$$s_{xy} = \frac{1}{n-1} (\mathbf{Y} - \bar{y}\mathbf{j})' (\mathbf{X} - \bar{x}\mathbf{j}) \quad (\text{M.88})$$

Proof. This proof echoes the previous proof.

$$\frac{1}{n-1} (\mathbf{Y} - \bar{y}\mathbf{j})' (\mathbf{X} - \bar{x}\mathbf{j}) \quad (\text{M.89})$$

$$= \frac{1}{n-1} [y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}, \dots, y_n - \bar{y},] \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \dots \\ x_n - \bar{x} \end{bmatrix} \quad (\text{M.90})$$

$$= \frac{1}{n-1} (y_1 - \bar{y})(x_1 - \bar{x}) + (y_2 - \bar{y})(x_2 - \bar{x}) + \dots + (y_n - \bar{y})(x_n - \bar{x}) \quad (\text{M.91})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \quad (\text{M.92})$$

□

synonym

Note that this notation leads to the synonym $s_y^2 = s_{yy}$. It also leads to a nice proof that the covariance matrix is symmetric.

Definition M.16. *The Variance-Covariance Matrix*

Let \mathbf{Y} be a random vector (a column vector whose elements are random variables). The quantity $\mathbb{V}[\mathbf{Y}]$ is called the **variance-covariance matrix** of \mathbf{Y} . It is often called just the covariance matrix of \mathbf{Y} .

With this definition, we have the following lemmas that you may wish to prove:

Lemma M.17. *Let $\mathbf{Y} \in \mathcal{M}_{r,c}$. If $\mathbf{Y}' = [\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \dots, \mathbf{Y}_r]$, then the elements of $\mathbb{V}[\mathbf{Y}]$ are $[\sigma_{ij}]$, where σ_{ij} is the covariance between Y_i and Y_j and $\sigma_{i,i}$ is the variance of Y_i .*

Lemma M.18. *Covariance matrices are symmetric.*

Lemma M.19. *Correlation matrices are symmetric.*

Lemma M.20. *If \mathbf{Y} is a random vector and \mathbf{X} is not, then $\mathbb{V}[\mathbf{X}'\mathbf{Y}] = \mathbf{X}'\mathbb{V}[\mathbf{Y}]\mathbf{X}$, assuming the multiplication makes sense (i.e., that the matrices are commensurate).*

M.7: End-of-Appendix Materials

M.7.1 EXERCISES

1. Prove $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$.
2. Prove matrix addition is commutative.
3. Prove matrix addition is associative.
4. Prove that scalar multiplication is associative.
5. Prove that scalar multiplication is distributive over addition.
6. Using a counterexample, prove that matrix multiplication is not commutative when only one of the matrices is diagonal (thus showing that *both* must be diagonal).
7. Let \mathbf{A} be any square matrix. Show that $\frac{1}{2}(\mathbf{A} + \mathbf{A}')$ is symmetric.
8. Determine the determinant of \mathbf{J}_{10} .
9. Determine the rank of \mathbf{J}_{10} .
10. Prove that $\mathbf{j}_{10} \mathbf{j}'_{10}$ is not positive definite.
11. Prove that $\mathbf{j}'_{10} \mathbf{j}_{10}$ is positive definite.
12. Prove Lemma M.6.
13. Prove Lemma M.7.
14. Prove Lemma M.8.
15. Prove Lemma M.17.
16. Prove Lemma M.18.
17. Prove Lemma M.20.



APPENDIX R:

EXPERIMENTING WITH R



OVERVIEW:

One of the keys to gaining a better understanding of statistics and randomness is to experiment with them. This chapter shows how to utilize the R Statistical Environment to perform simulations that allow you a glimpse into the wonder that is statistics.

Chapter Contents

R.1	Installing R	509
R.2	R Packages	510
R.3	R Functions.	512
R.4	Programming Practice	517



The purpose of this appendix is to

R.1: Installing R

The first step in using R is to install it on your computer. The process is relatively straight-forward. You download the installation program, then run the installation program.

You can find the installation program at the CRAN website:

```
https://cran.r-project.org/
```

Once there, download the appropriate installation program for your specific type of computer. Once it is downloaded, find the installation program and run it. Selecting all of the default options is fine. It's what I do.

The specifics are dependent on the type of computer operating system, so I will not get into the specifics. Let me point you in the direction of an Internet search engine and/or a site with a lot of videos. Learning how to access the information in these sites is a skill in itself.

So, enjoy the exploration and the learning.

Some prefer to install both R *and* an integrated developer environment (IDE) called R Studio. I do not. Except when doing certain niche projects, none of which we are doing in this book, R Studio just provides nothing (at best) or worse.

R.2: R Packages

While the base R that you installed has many statistical features, it is important to learn how to extend that base R installation. One way to do this is to install and load packages.

This book relies heavily on these packages:

- `car`
- `lawstat`
- `lmtest`
- `MASS`
- `randtests`
- `snpar`

Note that this book is not a part of the so-called `tidyverse`. Perhaps a future edition will be, but not now.

R.2.1 INSTALLING PACKAGES The overall scheme for installing packages is the same for each operating system. The specifics are slightly different to take advantage of what the operating systems allow. That scheme is

1. Run the `install.packages` function
2. Select a place to download the package from
3. ????
4. Profit

For instance, if I want to install the `car` package, I would run the following code in the Console window

```
|| install.packages("car")
```

Running this will download the `car` package from the Internet (the mirror site you selected) and install it on your computer. [If you have already specified the mirror site, then step two will not happen.]

It will also download all packages that are needed to run the `car` functions (called “dependencies”). You can see what other packages are installed by watching the Console window.¹

¹The script window is where you type your analysis script... it is where you “show your work.” The Console window is used for the quick work that is not a part of your analysis. Things like help queries and one-time-only work is done in the Console window.

R.2.2 LOADING PACKAGES The installation needs to be done just once for your computer. However, if you need to use the package in a script, you will need to load it into the working memory. Since this is something tied to the script, you should have this line in your script if you are using the `car` package:

```
|| library("car")
```

Once that line is run, R is able to access all functions (and/or data) in the package. Usually, there are a lot of functions in a package. To see what is available in the package, run this in the Console window:

```
|| help(package="car")
```

After running this, a page will pop up showing all available functions and links to their help pages. This particular package has a lot of functions associated with it. Some other packages have just a couple. The number of functions depends on the purpose of the package.

R.3: R Functions

There are a plethora of functions available in R. The key is to use functions to become familiar with them. Also, because you will always need to use new functions, it is very important to be familiar with the R help files for the functions.

R.3.1 BASIC FUNCTIONS The following are basic functions in R. You will use these frequently.

- `source` run external R script
- `head` show top 6 elements
- `tail` show last 6 elements
- `length` number of elements
- `seq` sequence of numbers
- `summary` six-number summary
- `mean` arithmetic average
- `median` median
- `sum` sum
- `sd` standard deviation
- `var` variance

Please know what each does. If necessary, use the help file for the given function. The more you familiarize yourself with the help files, the more they will tell you.

R.3.2 MATRIX FUNCTIONS R can also do arithmetic on matrices. Since the internal computer calculations are all based on matrices, it is important to be familiar with matrix operations to make sure you know what R is doing (can check the output).

ELEMENT-WISE FUNCTIONS:

- `+` usual matrix addition
- `*` Hadamard product (element-wise multiplication)
- `^` element-wise exponentiation

ADDITIONAL MATRIX FUNCTIONS:

- `%*%` usual matrix product
- `t` matrix transpose
- `solve` matrix inverse

R.3.3 PROBABILITY FUNCTIONS At the core of statistics is probability and probability distributions. These will be important in helping you better understand the effects of randomness on the estimates... and the effects of violating procedure requirements on those estimates.

- `set.seed` specify random number seed
- `sample` random sample from a vector

PROBABILITY FUNCTION NAMING LOGIC: Each probability function in R can be parsed into two parts, the stem and the prefix. The **stem** specifies the probability distribution, whereas the **prefix** specifies what aspect of the distribution you wish to access.

This is an exhaustive list of the prefixes:

- **d**
specifies the likelihood value. If the distribution is discrete, then this will also be the probability, otherwise it is the density.
- **p**
specifies the *cumulative* probability, $\mathbb{P}[X \leq x]$. This prefix will rarely be available for multivariate functions.
- **q**
specifies the quantile, the value of x that produces the probability. This prefix will rarely be available for multivariate functions.
- **r**
specifies a random variate. In simulation, this is the most important prefix, as it produces a random sample of a given size from the specified distribution.

The second part is the stem. This specifies the distribution involved. Here is a list of some of the more interesting stems available:

- **binom**
Binomial distribution. One needs to specify the number of trials, `size`, and the success probability, `prob`.
- **cauchy**
Cauchy distribution. Optionally, one can specify the location and the spread. The default is the standard Cauchy with `location` of 0 and `scale` of 1.
- **exp**
Exponential distribution. One needs to specify the `rate`, λ .
- **f**
Snedecor's F distribution. One needs to specify both degrees of freedom, with the numerator preceding the denominator degrees of freedom, `df1` and `df2`.
- **gamma**
Gamma distribution. One needs to specify the `shape` parameter, α . The `rate` parameter has a default value of $\sigma = 1$.
- **norm**
Normal (Gaussian) distribution. Optionally, one can specify the mean `m` and standard deviation `s` of the Normal. The default is mean 0 and standard deviation 1.
- **pois**
Poisson distribution. One needs to specify the expected value, `lambda`.

- `t`
Student's t distribution. One needs to also specify the number of degrees of freedom, `df`.
- `unif`
Continuous Uniform distribution. Optionally, one can specify the minimum and maximum value. The default is the standard Uniform with `min` of 0 and `max` of 1.

R.3.4 TESTING FUNCTIONS Because R is a statistical program, it is able to perform all of the basic statistical tests and procedures. These are the related functions.

- `aov` Analysis of Variance procedure
- `lm` OLS regression
- `summary.lm` summary of a linear model
- `summary.aov` summary of an ANOVA
- `residuals` calculated residuals from a model
- `confint` confidence interval for estimated parameters
- `predict` estimation and prediction of a value
- `runs.test` the runs test
- `hetero.test` univariate test of heteroskedasticity
- `fligner.test` test of heteroskedasticity across groups
- `bptest` Breusch-Pagan test of heteroskedasticity

R.3.5 CONTROL FUNCTIONS

- `for` for-loops
- `if` if-then logic
- `numeric` creates a vector in memory

R.3.6 GRAPHICAL FUNCTIONS R is a full-fledged graphical system. In fact, this is what set R apart from its competitors (and still does!). Every pixel of a graphic can be modified in R. This book relies on the basic R graphic engine. There are two other graphical engines (metaphors): `grid` and `ggplot2`. Base-R graphics will always serve you. `ggplot2` is the modern graphics engine for R. It serves as a wrapper for the basic graphics, making some graphics much easier to create.

- `qqnorm` Q-Q plot for a Normal target
- `qqline` plots the diagonal line in a Q-Q plot
- `barplot` bar chart
- `boxplot` box plot
- `hist` histogram
- `histogram` histogram that can be more easily modified
- `overlay` histogram with a overlaid density function
- `plot` basic scatter plot
- `par` specifies a graphical parameter
- `plot.new` starts a new plot
- `plot.window` specifies the viewing window
- `axis` draws an axis
- `title` writes a title on the graphic
- `lines` draws lines
- `points` draws points

R.4: Programming Practice

This section provides a series of practical examples to help you apply statistical concepts using the R Statistical Environment, one of the most versatile programming languages for statistical analysis and data visualization. R is particularly well-suited for solving a wide range of statistical problems, from basic descriptive statistics to advanced modeling techniques. As you progress through these examples, you will gain hands-on experience with R's powerful functions, libraries, and data structures, enabling you to approach statistical challenges with confidence.

I designed each example in this section to demonstrate the practical application of key statistical methods while also showcasing the capabilities of R. To help you fully understand the solutions, each example includes clear explanations, annotated code snippets, and discussions of the outputs. Even if you are new to R, the examples are structured to quickly build your proficiency with the language.

By working through these examples, you will not only deepen your understanding of statistical methods but also start developing a toolkit of practical R skills. Whether you aim to analyze data for research, business, or personal projects, these examples will equip you with the knowledge and techniques needed to leverage R effectively. Take your time to experiment with the code, explore variations, and see how the results change — this active learning approach will enhance both your statistical intuition and your programming expertise.

Note that there is a lot of white space in this section. It is there so that you can take notes directly on the examples.

Example 1

Produce a basic density plot of a Cauchy distribution between -3 and +3, where the Cauchy is centered at $x = 2$ and has an IQR of 3.

Solution: Since we are working with a probability distribution, let us refer to Section R.3.3. The function to calculate the density of the Cauchy centered at 2 with IQR 3 is `dcauchy(x, 2, 1.5)`. Thus, some code to produce a basic plot between -3 and +3 is

```
|| x = seq(-3, 3, length=1000)
|| y = dcauchy(x, location=2, scale=1.5)
|| plot(x, y)
```

With a little bit of work, you can make a graphic like Figure R.1. The fill color is the green in a palette of three colors designed to be safe for those with the usual color blindness, #1b9e77. When possible, it is best to accommodate those with color blindness and those who print out the graphic in shades of grey. ♦

The three safe colors are Green (#1b9e77), Orange (#d95f02), and Blue (#7570b3). These colors are from the `colorbrewer2.org` site.

Extension: Plot the pdf of a standard Normal distribution from -3 to +3 on the same graphic as a standard Cauchy. Looking at the two distributions, which has a higher variance?

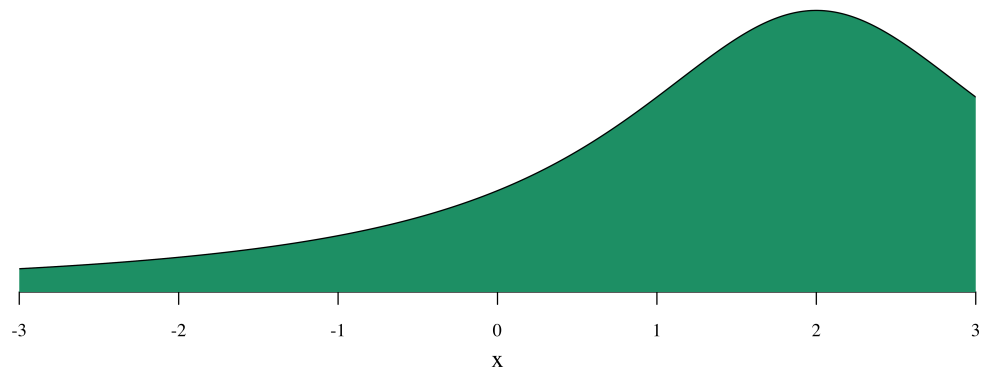


Figure R.1: The probability density function (pdf) of a Cauchy(2,1.5) distribution.

Example 2

Estimate a density graph of the volume of a cylinder with radius following a standard Uniform distribution and a height following a Normal distribution with mean 10 and standard deviation 0.1.

Solution: Since we are working with a probability distribution, I again refer you to Section R.3.3.

```
radius = runif(n=1e6)
height = rnorm(n=1e6, m=10, s=0.1)
volume = pi * radius^2 * height

hist(volume)
```

This is an example where using simulation easily obtains the approximate distribution. After using my R skills, I obtained the histogram in Figure R.2. See how close you can come to this. ♦

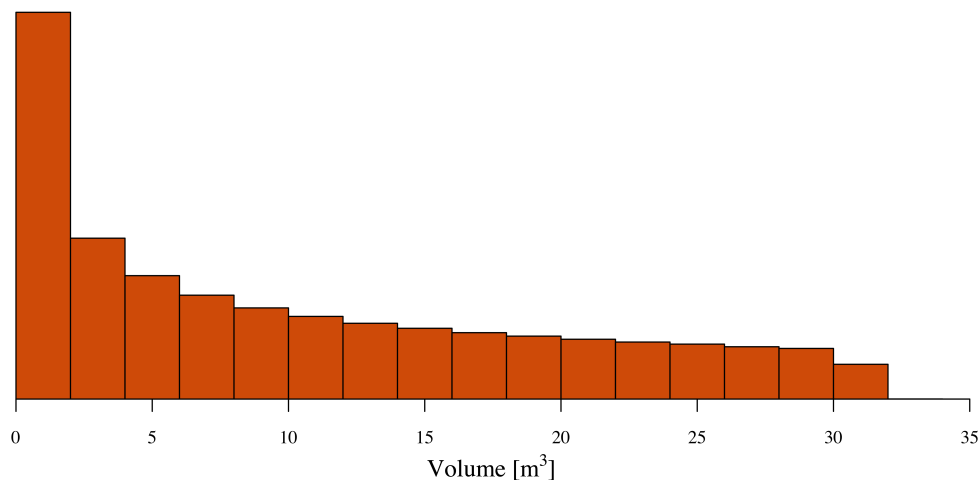


Figure R.2: The estimated probability density function (pdf) of the volume of a cylinder.

Example 3

Determine the effect of rounding on the appropriateness of the one-sample t-test.

Rounding can significantly affect statistical decisions because it alters the precision of numerical data (the inputs), which can lead to changes in calculated values such as means, variances, or p-values (the outputs). In hypothesis testing, for instance, rounding may cause a test statistic to fall on the border of a critical value, potentially shifting the conclusion about rejecting or failing to reject the null hypothesis. Similarly, in regression analysis, rounding predictor or response variables can affect the accuracy (and precision) of parameter estimates and the overall model fit.

These impacts are particularly pronounced in datasets with small sample sizes or values that are close to decision thresholds. Therefore, careful consideration of rounding practices is essential to ensure that statistical conclusions remain valid and reliable. This example explores the effects of rounding on the one-sample t-test.

Solution: This leaves a lot of decisions to us. The one-sample t-test requires data being generated from Normal distribution. So, let's generate data from a $\mathcal{N}(5,1)$ distribution. We need to round the data, so we will need to use the `round` function. Finally, we will need to determine if the distribution of the resulting p-values is standard Uniform (Section S.6.1).

The following code does this.

```
pval = numeric()

for(i in 1:1e4) {
  x = rnorm(10, m=5, s=1)
  y = round(x)
  pval[i] = t.test(y, mu=5)$p.value
}

ks.test(pval, "punif")
binom.test(sum(pval<0.05), n=length(pval), p=0.05)
```

The Kolmogorov-Smirnov test indicates that the distribution of the p-values is not standard Uniform. Thus, rounding in this situation breaks the t-test. Note that if we only care about $\alpha = 0.05$, we would use the Binomial test results. Given that p-value, I would still conclude that I should not use the t-test in this situation.

What about increasing the sample size from 10 to 50? The Kolmogorov-Smirnov test still indicates that the test is no longer acceptable. Note that if all we

care about is $\alpha = 0.05$, then the t-test *does* appear to be appropriate under these conditions.

What about increasing the variability in the data? Having a wider spread to the data may make the rounding less important. Let's change the standard deviation from 1 to 10 (and return the sample size to 10). From my run, the distribution of the p-value is still not standard Uniform, but the rejection rate for $\alpha = 0.05$ is close enough to 0.05.

What if we increase the sample size back to 50? In this situation, the distribution of the p-values is close enough to standard Uniform that the rounding does not affect the quality of the t-test conclusions. ♦

Example 4

If the inter-arrival time is Exponentially distributed with average time of 20 minutes, then what is the distribution of people who show up in an 8 hour period?

Algorithmic thinking is crucial in statistics as it enables a structured approach to solving complex problems by breaking them into smaller, logical steps. This mindset is particularly valuable when designing workflows for data analysis, from data preprocessing and visualization to modeling and interpretation. By thinking algorithmically, statisticians can systematically address challenges such as cleaning messy data, optimizing computational efficiency, or automating repetitive tasks.

Moreover, algorithmic thinking facilitates the translation of statistical concepts into code, allowing for reproducibility, scalability, and adaptability in analysis. In a field increasingly reliant on computational tools, developing this skill ensures that statisticians can efficiently tackle problems and adapt to new methods or technologies. This example illustrates algorithmic thinking to arrive at an interesting result.

Solution: This is a tough question but let's break it down into its parts, then simulate it.

The inter-arrival time is the time between arrivals.

```
|| iat = rexp(100, rate=3) ### 'iat' is in hours
```

The total time between the first and 20th arrival would be the sum of 20 of those inter-arrival times.

```
|| iat = rexp(100, rate=3)
|| sum(iat[1:20])
```

On the other hand, the number of arrivals in an hour would be

```
|| iat = rexp(100, rate=3)
|| iatCS = cumsum(iat)
|| max(which(iatCS <= 1))
```

So, the number of arrivals in eight hours would be

```
|| iat = rexp(100, rate=3)
|| iatCS = cumsum(iat)
```

```
|| max(which(iatCS <= 8))
```

To get the *distribution* of those “number of arrivals in eight hours,” you just need to repeat the code many times, saving the number of arrivals each time:

```
|| arrNum = numeric()  
||  
|| for(i in 1:1e6) {  
||   iat = rexp(100, rate=3)  
||   iatCS = cumsum(iat)  
||   arrNum[i] = max(which(iatCS <= 8))  
|| }  
||
```

The histogram (Figure R.3) shows the estimated distribution of the number of customers arriving in those 8 hours.

Closely examining the histogram *and* deeply thinking about the distributions you should have already learned, it is clear that the number of arrivals in 8 hours follows this distribution

$$\text{Number of Arrivals} \sim \mathcal{P}(\lambda = 24) \quad (\text{R.1})$$

To make it even more manifest, overlaying the histogram with a graph of that Poisson distribution illustrates this, Figure R.3.

```
|| hist(arrNum, freq=FALSE, breaks=seq(1,50)-0.5)  
|| points(1:50, dpois(1:50, lambda=24), pch=16)
```

Again, I used some R programming skills to obtain the graphic at the bottom. ♦

Note that this is not *proof* of the relationship between the two distributions. It merely suggests the relationship. Your probability theory course will give you the tools to actually prove the relationship.

Extension 1: Change all of the time measurements in the previous example from hours to minutes. Make sure that the conclusions are the same.

Extension 2: Change the inter-arrival time to 10 minutes. What is the distribution of the number of arrivals in three hours? Show it using the histogram.

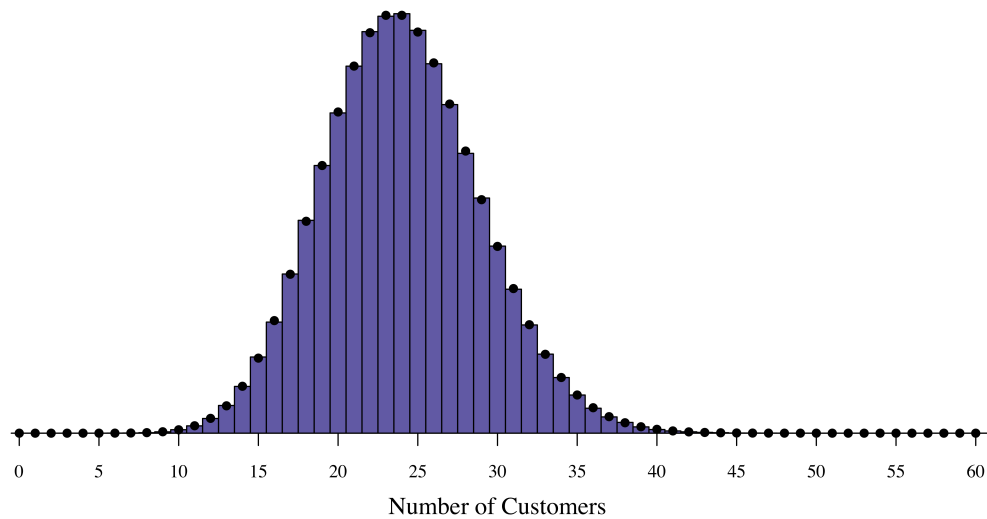


Figure R.3: The estimated probability mass function (pmf) of the number of customers arriving in eight hours. Note that it closely follows the $\mathcal{P}(\lambda = 24)$ distribution. Using the techniques of probability theory, one can prove this relationship.

Example 5

A flashlight uses five batteries. The lifetime of each battery is independent and follows a Gamma distribution with mean 50 days and standard deviation 5 days (`shape = 100`, `scale = 0.5`). The flashlight will show light until the first battery dies.

What is the expected time the flashlight will work after receiving five new batteries?

If we know the distribution of the lifetimes of each component, does this mean we know the distribution of the minimum lifetime? Sometimes. Note that sometimes we may be curious about the distribution of a *function* of random variables... which is also a random variable.

But, it goes beyond simple curiosity. Calculating the distribution of functions of random variables is essential because it allows us to understand how transformations or combinations of random variables behave and how their uncertainty propagates. This knowledge is fundamental in many applications, such as deriving the sampling distribution of an estimator, which forms the basis for hypothesis testing and constructing confidence intervals. Additionally, understanding the distribution of these functions enables probabilistic modeling in scenarios where direct distributions are unavailable, such as in operations research or risk management. It also helps in determining the likelihood of complex events, optimizing decision-making, and evaluating reliability in systems involving random inputs. By studying these distributions, statisticians can make more accurate predictions and draw meaningful inferences in both theoretical and applied contexts.

Solution: This is a great place for you to think through the problem (algorithmic thinking). What is the first step modeling this physical event? What is the second? Etc.? Use the next page to write out the steps... and the code. If done correctly, you will obtain the graphic at the bottom of the page, Figure R.4.

The question actually asked for the expected lifetime of the flashlight. The expected lifetime (mean) of the flashlight is 44.3 days, with a standard deviation of 3.1 days (sd). It is nice to know that 90% of the flashlights survive between 39.1 and 49.2 days (quantile).

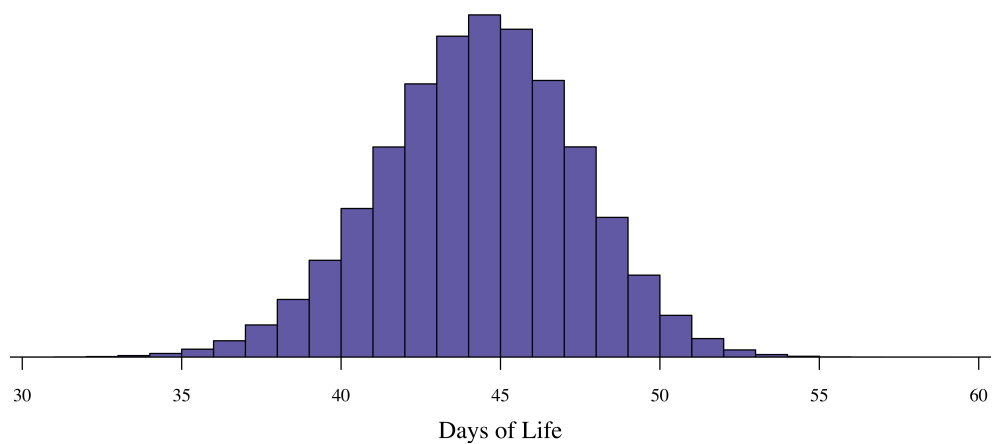


Figure R.4: The estimated probability density function (pdf) of the lifetime of a flashlight.

Example 6

A sample $n = 10$ counts is drawn from a population with unknown distributional characteristics. From the collected data, estimate a 95% confidence interval for the population mean.

4, 3, 4, 5, 4, 4, 1, 4, 9, 8, 3

The **bootstrap** is a powerful and versatile tool in statistics because it provides a non-parametric method for estimating the sampling distribution of a statistic without requiring strong assumptions about the underlying population distribution. By repeatedly resampling with replacement from the observed data, the bootstrap allows statisticians to approximate the variability of estimators, construct confidence intervals, and conduct hypothesis tests.

This approach is especially valuable when theoretical distributions are difficult to derive, such as with complex estimators or small sample sizes. Furthermore, the bootstrap is widely applicable across diverse statistical methods, making it a critical tool for robustness and flexibility in real-world data analysis. Its ability to leverage computational power to provide insights where traditional methods may falter has made the bootstrap indispensable in modern statistics.

Solution: Note that the sample size is small in this example. As such, we cannot rely on the Central Limit Theorem to assume the sample mean is Normally distributed. We do not even know if the population has a finite variance. The only thing we know is that we collected that particular sample. Thus, we will use the bootstrap.

That is, we will treat the sample as being representative of the population. Without this assumption, there is absolutely nothing we can do. With that assumption, we can effectively (or essentially) “recreate” the population as being an infinite repetition of this data. Then, to estimate the variability in the population, we simply redraw a sample of the same size from that population.

This is called the **non-parametric** bootstrap because it does not assume a specific (named) distribution of the data.

Here is the code to accomplish this once.

```
theData = c(4, 3, 4, 5, 4, 4, 1, 4, 9, 8, 3)
newData = sample(theData, replace=TRUE)
mean(newData)
```

This will give us one more sample mean. It takes thousands of them to understand the distribution (centers and variabilities). Thus, the non-parametric bootstrapping code will be

```
theData = c(4, 3, 4, 5, 4, 4, 1, 4, 9, 8, 3)
newMeans = numeric()

for(i in 1:1e6) {
  newData = sample(theData, replace=TRUE)
  newMeans[i] = mean(newData)
}

mean(newMeans)
sd(newMeans)
quantile(newMeans, c(0.025,0.975))
```

From this code, I estimate the population mean is 4.45, with an estimated 95% confidence interval from 3.27 to 5.82. ♦

The non-parametric bootstrap is used when all you know about the population is that you randomly selected the sample. It is a *very* flexible procedure that should be a part of your statistical toolbox.

However, statistically, it tends to be of low power (confidence intervals are too wide; p-values are too high). The reason for this lower power is that you are making fewer assumptions on the population. If you know more about the population and are able to use it, then that procedure will have higher power (all things being equal).

Figure R.5 shows the histogram of the sample means from the simulation. The triangle on the x-axis represents the mean of these sample means. The thick bar along the axis represents the estimated 95% confidence interval.

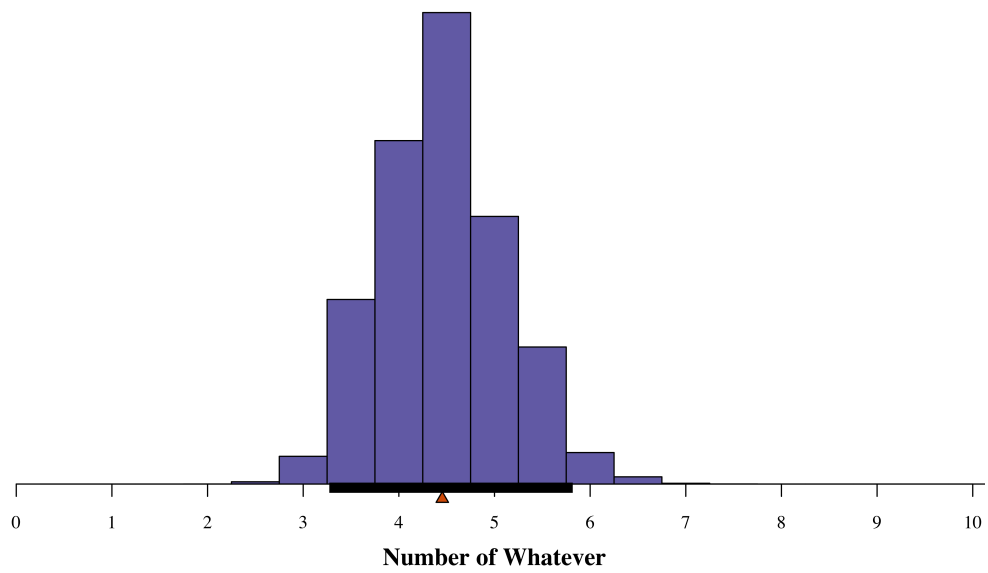


Figure R.5: The estimated probability density function (pdf) of the sample means from the unknown population. The mean of the sample means is denoted by the triangle on the axis; the 95% confidence intervals, thick bar on the axis.

Example 7

How good is the non-parametric bootstrap in terms of covering the population mean? For this investigation, let's assume the population really is standard Normal.

Solution: The first step is to draw a sample of size $n = 10$ from that population, then do the non-parametric bootstrap, then determine how frequently the population mean is in the confidence interval. It should happen 95% of the time.

Here is the code:

```
estLCL = numeric()
estUCL = numeric()

for(j in 1:1e6) { # The loop testing the NPB

  theData = rnorm(10)
  newMeans = numeric()

  for(i in 1:1e6) {
    newData = sample(theData, replace=TRUE)
    newMeans[i] = mean(newData)
  }

  estLCL[j] = quantile(newMeans,0.025)
  estUCL[j] = quantile(newMeans,0.975)
}
```

Note that this will take quite some time to run. There are a total of 1,000,000,000,000 iterations taking place. On a new laptop, one should probably expect it to run overnight. On an older one, it may take an entire day. However, the results will be rather precise.

Let me take this opportunity to (re-) introduce you to a measure of the quality of a confidence interval: **coverage**. Coverage is defined as the proportion of the time that the confidence interval contains the true mean.

For our example, the true mean is 0. So, the coverage will be the proportion of the time that 0 is between the two estimated confidence bounds.

```
|| mean( 0>estLCL & 0<estUCL )
```

When I run this, I obtain 0.896, which is quite different from the hoped-for 0.95. It means I am rejecting at a rate of 10.4% instead of the claimed 5%. This is not good.



Having too low of a coverage (as here) is problematic because it increases the risk that the true parameter value lies outside the confidence interval or prediction range. Coverage probability reflects the proportion of intervals that, over repeated sampling, are expected to contain the true value. If the coverage is too low, it means that the interval is overly narrow or poorly calibrated, leading to an underestimation of uncertainty. This can result in overconfidence in statistical conclusions, which may cause critical errors in decision-making or policy formulation, particularly in fields like medicine, finance, or engineering where risks are high.



APPENDIX S:

THE APPENDIX OF STATISTICS

OVERVIEW:

This appendix covers most of the things you need to remember from your introductory statistics course — and more. Treat this as more than just a review. It includes all you need from your statistics course, *plus some additional* items that you may find helpful. Starred sections, as usual, are optional for understanding linear models at a basic level.

Note that I *do* include a section on moment generating functions (MGFs) and a proof on the Central Limit Theorem (CLT). Neither tend to be included in introductory statistics courses. Also, neither proof is overly important for this course. Both are offered up on the altar of “this is kinda interesting.”

Chapter Contents

S.1	Importantly Confusing Points	535
S.2	Sample Statistics	537
S.3	Population Parameters.	545
S.4	Probability Distributions	556
S.5	Distributions of Sample Statistics	576
S.6	Other Topics	580
S.7	End-of-Appendix Materials	603



At the very center of statistics... and of *understanding* statistics... is the **random variable**. There are many ways of defining a random variable. This is important, because the random variable is central to our understanding of probability and statistics. The “official” definition of a random variable is

rv

Definition S.1. *Random Variable*

A random variable $X : \Omega \rightarrow E$ is a measurable function from a set of possible outcomes Ω to a measurable space E .

While this definition has the advantage of being mathematically precise, it is not necessarily helpful in terms of understanding what a random variable *really is*. For me, a random variable is just an unknown outcome of an experiment.

The “variable” you were introduced to in high school algebra is a value that is unknown until more information is available. A *random variable* is a value that is unknown until that experiment is performed. Repeated experiments may produce different values.

Here, I am using the term “experiment” broadly. It is any action, including simple observation. Thus, my height right now is not a random variable. However, my height in three years *is* a random variable. I will not know its value until I perform the experiment (measure it in three years).

Random variables have distributions. Nothing else has a distribution in the sense we are using it here — *nothing* else. Since random variables have distributions, they also have expected values, variances, minimums, maximums, medians, and many other measures on the variable.

Samples are drawn from random variables to better understand how they behave (a.k.a. their distribution, Section S.5). Thus, if I want to understand the rela-

relationship between engine displacement (a fixed/set variable) and mileage (a random variable) for automobiles in general, I would measure engine displacements and mileages from a sample of automobiles. That sample would give me information about the relationship between those two variables.

Note: There is a difference between the variable (what we measure) and the values (results from our measurements). The values constitute our sample. This means that the engine displacement is a variable and some values are {1.76, 2.25, 1.8, 2.5, 0.75} liters. Similarly, mileage is a variable and {25.3, 17.8, 34.4, 14.5, 30.5} are some values.

To help with notation, I *try* to follow these rules: Observations (data, values) are denoted with lowercase Latin letters like x , y , and z ; random variables (unobserved), with uppercase Latin letters like X , Y , and Z ; population parameters in need of estimating, with Greek lowercase letters like μ , σ , and π ; and the set of possible parameter values, with Greek uppercase letters like M , Σ , and Π .

S.1: Importantly Confusing Points

So, here is something that is extremely confusing because we use the same word to mean/imply different things. A set of numbers has a variance. A sample from a random variable has a variance. A distribution has a variance.¹ All three measures are termed “variance;” however, they mean different things.

Here is a set of numbers I specify (i.e., they are not randomly generated): {1, 2, 3, 4, 5, 6, 7}. They have a mean and a variance. The mean is 4. The variance is 4.667. Those measures only tell me about the numbers I specified. Since they are not realizations of a random variable, it does not make sense to use those values for anything other than mathematical calculations on those particular values.

Here is a set of numbers generated from a Poisson distribution (see Section S.4.3) with mean $\lambda = 1$: {1, 2, 1, 2, 0}. These values (this sample) also have a mean ($\mu = 1.20$) and a variance ($\sigma^2 = 0.70$). Since these numbers are a realization of (an observation from) a random variable, these two values can be called “statistics” (descriptive or sample statistics). They are functions of observed random variables. Because they are statistics, the mean should be close to the expected value of the random variable, and the variance should be close to the variance of the random variable.

observations

statistics

¹One can also say that a random variable has a variance. While a random variable is not a distribution (it *follows* one, however), we will often conflate the random variable and its distribution.

population
parameters

Since these statistics are functions of a random sample, those statistics are also random variables. As such, they have distributions. In fact, understanding and knowing the distribution of sample statistics is one of the goals of statistics. This study led you to confidence intervals and test statistics back in your previous statistics course.

Here is a Chi-square distribution with parameter “4 degrees of freedom”: $\chi^2_{\nu=4}$. It has an expected value ($\mu = 4$) and a variance ($\sigma^2 = 8$). These are (population) parameters and are not a function of a sample (hence they are represented with lowercase Greek letters). These are genuine numbers that have no distribution.²

They are also numbers that we do not know when we are applied statisticians. These are, however, numbers that we *want* to know — or at least obtain a good estimate of. To get those estimates, we obtain a sample from the distribution and measure the corresponding sample statistics. While the sample statistics will not (usually) exactly equal the population parameters, they will tend to be close. How close depends on the sample size and the statistic measured.

Note: This “issue” with the variance is not unique to the variance. The lesson to take away from this is that you need to be aware of the symbols and what they mean. This is especially true when it comes to values, random variables, and distributions.

²They have no distribution because they do not vary. Our *understanding* of them may vary, but the value does not.

S.2: Sample Statistics

A statistic is a function of the data. Because the data are random variable, so are statistics. There are several sample statistics that you should have seen in your previous statistics course. Here are the definitions of those important to linear models.

S.2.1 SAMPLE MEAN The sample mean is the usual “average” that we have calculated many times in the past. If we want to use just one number to summarize the data, the mean is the usual value.

Definition S.2. *Sample Mean*

Let Y_i be a random sample from a distribution. The sample mean is defined as

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad (\text{S.1})$$

Note that this definition is equivalent to

$$\sum_{i=1}^n Y_i = n\bar{Y} \quad (\text{S.2})$$

This form is useful in many proofs, and you will see one later. Before we get to that, however, here is an elementary property of the sample mean that needs to be provided.

Lemma S.3. *The sample mean is a linear functional.*

Proof. This means we have to show that the mean of $(aY + b)$ is $a\bar{Y} + b$ for scalars a and b . The proof is rather straight forward. We just find the mean of $aY + b$ using

the definition.

$$\overline{aY + b} = \frac{1}{n} \sum_{i=1}^n (aY_i + b) \quad (\text{S.3})$$

$$= \frac{1}{n} \sum_{i=1}^n aY_i + \frac{1}{n} \sum_{i=1}^n b \quad (\text{S.4})$$

$$= a \frac{1}{n} \sum_{i=1}^n Y_i + \frac{1}{n} nb \quad (\text{S.5})$$

$$= a\overline{Y} + b \quad (\text{S.6})$$

Thus, through the transitive property, the expected value of a linear combination of the variable is just a linear combination of the expected values. \square

Note: You should go through each of the steps in the previous proof and give a reason the step is mathematically correct. This will give you some practice in the mathematics underlying linear models.



Now, we show that the average deviance from the mean is zero.

Lemma S.4 (The Mean Deviance Lemma). *Let Y_i be a sample of size n from a random variable.*

$$\sum_{i=1}^n (Y_i - \overline{Y}) = 0$$

exercise

I leave the proof as an exercise.

While it seems as though this lemma is not important, you will see its application over and over again. It will serve you well to learn it by sight as quickly as possible.

S.2.2 SAMPLE VARIANCE If we use the mean to summarize the entire data set using a single number, the measure of spread is used to see how well that number

represents each element of the data. There are many measures of spread that you probably have come across. These include the variance, standard deviation, and interquartile range. Here we will focus on the variance.

Definition S.5. *Sample Variance*

Let Y_i be a random sample of size n from a distribution. The sample variance is defined as

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (\text{S.7})$$

Here are some elementary properties of the sample variance that need to be provided.

Lemma S.6. *The sample variance is a quadratic functional.*

Proof. This means we need to show that the variance of $(aY + b)$ is $a^2 S_y^2$. This is also a relatively straight forward application of the definition.

Thus, let Y_i be a random sample of size n from a distribution. With this, we have the following

$$S_{aY+b}^2 = \frac{1}{n-1} \sum_{i=1}^n ((aY_i + b) - \overline{(aY + b)})^2 \quad (\text{S.8})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (aY_i + b - a\bar{Y} - b)^2 \quad (\text{S.9})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (aY_i - a\bar{Y})^2 \quad (\text{S.10})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (a(Y_i - \bar{Y}))^2 \quad (\text{S.11})$$

$$= \frac{1}{n-1} \sum_{i=1}^n a^2 (Y_i - \bar{Y})^2 \quad (\text{S.12})$$

$$= a^2 \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (\text{S.13})$$

$$= a^2 S_y^2 \tag{S.14}$$

□

Note: Proofs S.3 and S.6 both exemplify one important method for proving statements: substitute and simplify. It also provides to us an important result. Go over it a few times. Realize the importance of the scalar multiple on the variance, as well as the scalar addend on the variance.

Also, think about why it makes sense for a translation to have no effect on the variance, but a scale change to have a significant effect.

What is so important about the previous lemma? Think about the variance of heights of students in this room. First, measure in feet, then in inches, then in centimeters. Why does the variance change in the three measurements?

Now, have everyone stand on the same chair while measuring them in inches. Why does the variability when measured to the floor does not depend on whether they are standing on the chair or not?

S.2.3 SAMPLE COVARIANCE The variance measures the variability of a single value. The *covariance* measures the variability of the relationship between two numeric variables.

Definition S.7. *Sample Covariance*

Let X_i be a random sample from a distribution. Let Y_i be a random sample from a distribution (same as X or different). The sample covariance is defined as

$$S_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \quad (\text{S.15})$$

The covariance is also denoted by $\text{Cov}[X, Y]$. However, using this symbol may lead to confusion about whether you are talking about a sample or a population. It is safer to stay with $S_{X,Y}$ as the symbol.

Here are some elementary properties of the sample covariance.

Lemma S.8. $S_{X+Y,Z} = S_{X,Z} + S_{Y,Z}$.

Proof. This proof is also a direct application of the definition and of the proof style of Theorem S.3.³

$$S_{X+Y,Z} = \frac{1}{n-1} \sum_{i=1}^n ((X_i + Y_i) - (\bar{X} + \bar{Y}))(Z_i - \bar{Z}) \quad (\text{S.16})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X} + Y_i - \bar{Y})(Z_i - \bar{Z}) \quad (\text{S.17})$$

³Remember that proving things is important. As you work through this material, you need to be able to prove things mathematically. In statistics, however, not all things can be proven mathematically. Some things need simulation to give provisional results. However, proving things mathematically is preferred.

$$= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Z_i - \bar{Z}) + \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z}) \quad (\text{S.18})$$

$$= S_{X,Z} + S_{Y,Z} \quad (\text{S.19})$$

□

exercises

I leave the following as exercises for you.

Lemma S.9. $S_{X,Y} = S_{Y,X}$

Lemma S.10. $S_{aX,bY} = ab S_{X,Y}$

Lemma S.11. $S_{X,X} = S_X^2$

S.2.4 SAMPLE CORRELATION Like the covariance, the correlation measures the strength of the relationship between two numeric variables. However, where the covariance cannot be meaningfully compared across variables, the correlation can. A correlation value of 0.25 indicates the same exact thing whether you are looking at the relationship between height and weight or between population and GDP. A covariance of 100 means something different if one is looking at the relationship between height and weight or if one is looking at the relationship between population and GDP.

Definition S.12. Sample Correlation

Let X_i be a random sample of size n from a distribution with finite variance. Let Y_i be a random sample from a distribution (same as X or different) with finite variance. The sample correlation is defined as

$$R_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (\text{S.20})$$

The correlation is also symbolized using $\text{Cor}[X, Y]$. However, this may be confusing as this symbol is also used for the population correlation.

Alternate formulas for the correlation include:

$$R_{X,Y} = \frac{S_{X,Y}}{\sqrt{S_X^2 S_Y^2}} \quad (\text{S.21})$$

$$R_{X,Y} = \frac{S_{X,Y}}{S_X S_Y} \quad (\text{S.22})$$

I leave the following as exercises for you.

exercises

Lemma S.13. $R_{X,Y} = R_{Y,X}$

Lemma S.14. $R_{aX,Y} = R_{X,Y}$

Lemma S.15. $R_{aX+b,Y} = R_{X,Y}$

Lemma S.16. $-1 \leq R_{X,Y} \leq 1$

Note: The Lemma S.15 result is very important. Do not just think about them in terms of the mathematics. Think about what these mean in terms of the

reality: If we apply a linear transformation to the two variables, how does that affect the correlation? What does that *really* mean?

The Lemma S.16 result is also important. Why? It gives a meaning to the correlation that can be compared across data sets.



bias

In each of the above definitions, the numeric divisor is chosen to ensure that the sample statistic is an unbiased estimator of the population parameter. In other words, we divide by $n - 1$ so that this equation is true:

$$\mathbb{E}[\text{statistic}] = \text{parameter} \tag{S.23}$$

df

It is interesting that those denominators are the “degrees of freedom” for that statistic. This gives one helpful “definition” for the term **degrees of freedom**.

S.3: Population Parameters

Let us spend a section examining population parameters. A population parameter is a measurement on the *population*, analogous to the sample statistic being a measurement on the sample.

Note that it is the population parameters we seek to know. The sample statistics are only used to give us information about them. Some usual population parameters of interest are the mean, variance, covariance, and correlation.

S.3.1 POPULATION MEAN The population mean is a measure of center (or “typicalness”) for the entire distribution (or random variable). It is known by a few different names: population mean, expected value, and first moment. It is represented by either μ or $\mathbb{E}[Y]$.

If the distribution is discrete, the formula for the expected value is

$$\mathbb{E}[Y] = \sum_{i \in \mathcal{S}} y_i \mathbb{P}[Y = y_i] \quad (\text{S.24})$$

If the distribution is continuous, the formula for the expected value is

$$\mathbb{E}[Y] = \int_{\mathcal{S}} y f(y) dy \quad (\text{S.25})$$

Note the similarities between the two formulas. The difference is due solely to the mathematics underlying Equation S.25. Those mathematics tend to be covered in an undergraduate real analysis course.

Note: Another interpretation of the expected value is the “long-run average” of the outcomes. I frequently find this useful in checking that my results match my expectations.

Example 1

Let Y be the number of heads in one flip of a fair coin. What is the expected number of heads?

Solution: The random variable Y is discrete with sample space $\mathcal{S} = \{0, 1\}$. Thus, the formula for the expected value gives us

$$\mathbb{E}[Y] = \sum_{i \in \mathcal{S}} y_i \mathbb{P}[Y = y_i] \quad (\text{S.26})$$

$$= 0 \mathbb{P}[Y = 0] + 1 \mathbb{P}[Y = 1] \quad (\text{S.27})$$

$$= 0(0.500) + 1(0.500) \quad (\text{S.28})$$

Thus, the expected number of heads is $\mathbb{E}[Y] = 0.500$. ♦

Note: The expected value of a random variable does *not* need to be an element of the sample space (as here). Thus, the interpretation as a **long-run average** helps in understanding the mean.

Example 2

Let Y be the time I spend at a stoplight. If the light has a 120-second cycle, spending 55s on green, 5s on yellow, and 60s on red, calculate the expected time I wait, *given* that I have to stop.

Solution: The random variable Y is continuous with sample space $\mathcal{S} = [0, 60]$. Without any information beyond knowing that there is a lower and an upper bound, we should assume the random variable follows a Uniform distribution.⁴

⁴For those interested in understanding why the Uniform is the distribution, please read up on “maximum entropy distributions.”

Thus, the formula for the expected value gives us

$$\mathbb{E}[Y] = \int_S y f(y) \, dy \quad (\text{S.29})$$


$$= \int_0^{60} y \frac{1}{60} \, dy \quad (\text{S.30})$$

$$= \frac{1}{60} \int_0^{60} y \, dy \quad (\text{S.31})$$

$$= \frac{1}{60} \left. \frac{y^2}{2} \right|_{y=0}^{60} \quad (\text{S.32})$$

$$= \frac{1}{60} \left(\frac{60^2}{2} - \frac{0^2}{2} \right) \quad (\text{S.33})$$

$$= 30 \quad (\text{S.34})$$

Thus, the expected time I spend at the stop light is $\mathbb{E}[Y] = 30$ seconds. Why does that not surprise me? 

Example 3

Let Y be the time I spend at a stoplight. If the light has a 120-second cycle, spending 55s on green, 5s on yellow, and 60s on red, calculate the expected time I wait.

Solution: This is quite different from the previous example. There, we knew the light was red. Now, we do not know if it is green (no time stopped), yellow (no time stopped), or red.

One can follow the above procedure to solve the problem. However, I would like to use a more helpful procedure. To motivate the procedure, note that the mean has been calculated as the sum of each value times its probability (Eqn. S.24). Thus, we could calculate the expected time to wait as

$$\begin{aligned}\mathbb{E}[\text{wait}] &= \mathbb{P}[\text{green}]\mathbb{E}[\text{green}] + \mathbb{P}[\text{yellow}]\mathbb{E}[\text{yellow}] + \mathbb{P}[\text{red}]\mathbb{E}[\text{red}] \\ &= \frac{55}{120} \cdot 0 + \frac{5}{120} \cdot 0 + \frac{60}{120} \cdot 30 \\ &= \frac{1800}{120}\end{aligned}$$

Thus, *without* the additional information that I actually stopped at the stoplight, the expected wait time is only 15s. \blacklozenge

Now that we have gone through the calculations, make sure that the logic makes sense.



With formulas S.24 and S.25 above, one can find the expected value of any *function* of a random variable, too. For instance, $\mathbb{E}[Y^2]$ is the expected value of the square of the random variable. It is calculated as

$$\mathbb{E}[Y^2] = \sum_{i \in \mathcal{S}} y_i^2 \mathbb{P}[Y = y_i] \tag{S.35}$$

or, if Y is continuous,

$$= \int_S y_i^2 f(y) dy \quad (\text{S.36})$$

This will come in handy in terms of notation.

Example 4

Let $Y \sim \text{Bern}(\pi = 0.500)$. Calculate $\mathbb{E}[Y]$, $\mathbb{E}[Y^2]$, and $\mathbb{E}[Y^3]$.

Solution: Recall that the probability mass function for Y is given by

$$f(y) = \begin{cases} 0.500 & y = 0 \\ 0.500 & y = 1 \end{cases} \quad (\text{S.37})$$

Thus, we have

$$\mathbb{E}[Y] = \sum_{y \in S} p(y) y \quad (\text{S.38})$$

$$= (0.500) 0 + (0.500) 1 \quad (\text{S.39})$$

$$= 0.500 \quad (\text{S.40})$$

and

$$\mathbb{E}[Y^2] = \sum_{y \in S} p(y) y^2 \quad (\text{S.41})$$

$$= (0.500) 0^2 + (0.500) 1^2 \quad (\text{S.42})$$

$$= 0.500 \quad (\text{S.43})$$

and

$$\mathbb{E}[Y^3] = \sum_{y \in S} p(y) y^3 \quad (\text{S.44})$$

$$= (0.500) 0^3 + (0.500) 1^3 \quad (\text{S.45})$$

$$= 0.500 \quad (\text{S.46})$$

Clearly, that these three moments are equal is a feature of the Bernoulli distribution. Such will not necessarily be true for any other distribution. \blacklozenge

Lemma S.17. If $Y \sim \text{Bern}(\pi)$, then $\mathbb{E}[Y^n] = \pi$ for any positive, finite-valued n .

Proof. Recall that the probability mass function for Y is given by

$$f(y) = \begin{cases} 1 - \pi & y = 0 \\ \pi & y = 1 \end{cases} \quad (\text{S.47})$$

Thus, we have

$$\mathbb{E}[Y^n] = \sum_{y \in \mathcal{S}} p(y) y^n \quad (\text{S.48})$$

$$= (1 - \pi) 0^n + (\pi) 1^n \quad (\text{S.49})$$

$$= (1 - \pi) 0 + (\pi) 1 \quad (\text{S.50})$$

$$= \pi \quad (\text{S.51})$$

Thus, it is proven. \square

Note: When discussing the random variable, one will tend to refer to $\mathbb{E}[Y]$. When discussing the *distribution* of the random variable, one will tend to refer to μ . Since random variables “follow” (or “have”) a distribution, the two terms tend to be treated as being interchangeable, especially since $\mathbb{E}[Y] = \mu$.

Formula S.25 for the expected value is the one we all see in our introductory statistics course. There is another formula that may be useful to know. Given $F(y)$ is the cumulative distribution function (CDF) for Y , this is true:

$$\mathbb{E}[Y] = \int_0^{\infty} (1 - F(y)) dy \quad (\text{S.52})$$

Note that this formula *only works* if the support of the random variable is non-negative.

Example 5

Let Y be the time I spend at a stoplight. If the light has a 120-second cycle, spending 55s on green, 5s on yellow, and 60s on red, calculate the expected time I wait, given that the light is red.

Solution: This is the same problem as above (S.3.1). However, let us use equation S.52. We know that the CDF for the above Uniform is

$$F(y) = \begin{cases} 0 & y \leq 0 \\ y/60 & 0 < y \leq 60 \\ 1 & y > 60 \end{cases} \quad (\text{S.53})$$

Thus, $\mathbb{E}[Y]$ is

$$\mathbb{E}[Y] = \int_0^{\infty} (1 - F(y)) \, dy \quad (\text{S.54})$$

$$= \int_0^{\infty} \begin{pmatrix} 1 - y/60 & 0 < y \leq 60 \\ 0 & y > 60 \end{pmatrix} \, dy \quad (\text{S.55})$$

$$= \int_0^{60} 1 - y/60 \, dy \quad (\text{S.56})$$

...

I leave the rest as an exercise for you.



exercise

S.3.2 POPULATION VARIANCE The population variance is a measure of uncertainty (or of spread or uncertainty) for the distribution. It is symbolized as σ^2 and as $\mathbb{V}[Y]$. When discussing the random variable, one will tend to refer to $\mathbb{V}[Y]$; however, when discussing the distribution of the random variable, one will tend to refer to σ^2 , as above.

The typical population variance formulas are

$$\mathbb{V}[Y] = \sum_{i \in S} (y_i - \mu)^2 \mathbb{P}[Y = y_i] \quad (\text{S.57})$$

$$= \int_S (y_i - \mu)^2 f(y) dy \quad (\text{S.58})$$

Note that these are merely

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mu)^2] \quad (\text{S.59})$$

moment

In other words, the variance is defined as the second central moment.

This last definition may be more helpful in our understanding of the variance, especially as it leads to a different formula for the variance:

$$\mathbb{E}[Y^2] = \sigma^2 + \mu^2 \quad (\text{S.60})$$

Lemma S.18. *Let Y be a random variable with finite mean and variance.*

$$\mathbb{E}[Y^2] = \mathbb{V}[Y] + \mathbb{E}[Y]^2$$

Proof. The proof follows from the definition and algebra:

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] \quad (\text{S.61})$$

$$= \mathbb{E}[Y^2 - 2Y\mathbb{E}[Y] + \mathbb{E}[Y]^2] \quad (\text{S.62})$$

$$= \mathbb{E}[Y^2] - 2\mathbb{E}[Y\mathbb{E}[Y]] + \mathbb{E}[\mathbb{E}[Y]^2] \quad (\text{S.63})$$

Since $\mathbb{E}[Y]$ is a population parameter, it is its own expected value. It is also independent of the random variable. These two facts give us

$$\mathbb{V}[Y] = \mathbb{E}[Y^2] - 2\mathbb{E}[Y]\mathbb{E}[Y] + \mathbb{E}[Y]^2 \quad (\text{S.64})$$

$$= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \quad (\text{S.65})$$

Using familiar symbols, this is just

$$\sigma^2 = \mathbb{E}[Y^2] - \mu^2 \quad (\text{S.66})$$

Rearranging the terms gives our result

$$\mathbb{E}[Y^2] = \mathbb{V}[Y] + \mathbb{E}[Y]^2 \quad (\text{S.67})$$

And, this is what we wished to prove. □

Note: While this proof is worded in terms of population parameters, it also works with sample statistics:

$$\mu'_2 = s^2 + \bar{y}^2 \quad (\text{S.68})$$

where

$$\mu'_2 = \frac{1}{n} \sum_i y_i^2 \quad (\text{S.69})$$

Note that the subscript of '2' indicates the second moment and the prime indicates it is the sample version. Thus, μ'_2 is the second sample moment, and $\mu_2 = \mathbb{E}[Y^2]$ is the second population moment.

This is not important for the mathematics underlying linear models. However, you may find it interesting that statisticians really are trying to make our symbols follow consistent rules.

ANOTHER VARIANCE FORMULA: The variance formulas given above in Equations S.57 and S.58 are the typical ones provided in undergraduate statistics. They depend on the probability mass (or density) function (pmf or pdf). There is a formula for the population variance that relies on the cumulative distribution function (CDF) instead. Here it is.

$$\sigma^2 = 2 \int_0^{\infty} y(1-F(y)) dy - \left(\int_0^{\infty} (1-F(y)) dy \right)^2 \quad (\text{S.70})$$

Note that this formula *only works* for a non-negative random variable... that is, for random variables whose possible values must be positive. Also note that this formula should only be used if the CDF is easier to work with than the pdf. This is a rare event, which illustrates why few see this formula.

Example 6

Show that the variance of a standard Uniform distribution is

$$\mathbb{V}[Y] = \frac{1}{12}$$

Use both the pdf and the CDF method.

Solution: I leave this as an exercise for you. ♦

S.3.3 POPULATION COVARIANCE AND CORRELATION The formula for the population covariance is

$$\sigma_{xy} = \mathbb{E}[(X - \mu_x)(Y - \mu_y)] \quad (\text{S.71})$$

Note this reduces to $\mathbb{V}[Y] = \text{Cov}[Y, Y]$ and $\sigma_y^2 = \sigma_{yy}$.

The formula for the population correlation is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (\text{S.72})$$

The Greek letter ρ is “rho.” It is not a “p.”

Example 7

What is the covariance between a variable and a constant?

Solution: I leave this as an exercise for you. ♦

Example 8

What is the covariance between two constants?

Solution: I leave this as an exercise for you. ♦

S.4: Probability Distributions

There are several distributions you will have experienced in the past. They include the Binomial, Poisson, Normal, Chi-square, Student's t, and F distributions. Each of these distributions either arises in nature or as a consequence of needing to test a hypothesis. The next several sections covers these distributions — and more.

S.4.1 BERNOULLI Arguably, the Bernoulli is the grandfather of all discrete distributions.⁵ It models a random variable that has two possible outcomes, which we call “failure” and “success,” or 0 and 1. This is its “sample space,” the set of all outcomes with non-zero probability,

$$\mathcal{S} = \{0, 1\} \quad (\text{S.73})$$

sample space

The Bernoulli has a single parameter, π , which is the probability of the random variable being 1.

parameter

The probability mass function for the Bernoulli is

pmf

$$f(y) = \begin{cases} 1 - \pi & y = 0 \\ \pi & y = 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{S.74})$$

Technically, the probability mass function must return a value for every real y . Thus, the third line in the definition of the Bernoulli pmf. With that being said, it is frequently left off and assumed. Thus, we will often see it as

$$f(y) = \begin{cases} 1 - \pi & y = 0 \\ \pi & y = 1 \end{cases} \quad (\text{S.75})$$

without any loss of understanding.

Its cumulative distribution function of the Bernoulli distribution is

CDF

$$F(y) = \begin{cases} 0 & y < 0 \\ 1 - \pi & 0 \leq y < 1 \\ 1 & 1 \leq y \end{cases} \quad (\text{S.76})$$

Similarly, this will frequently be written as

$$F(y) = 1 - \pi \quad \text{for } 0 \leq y < 1 \quad (\text{S.77})$$

without confusion.

The mean (expected value) of a Bernoulli random variable is

mean

⁵The standard uniform distribution, which can be used to generate Bernoulli random variables, is the source of *all* distributions.

$$\mathbb{E}[Y] = \sum_{i \in \{0,1\}} y_i f(y_i) \quad (\text{S.78})$$

$$= 0 \times (1 - \pi) + 1 \times (\pi) \quad (\text{S.79})$$

$$= \pi \quad (\text{S.80})$$

Its variance is

variance

$$\mathbb{V}[Y] = \sum_{i \in \{0,1\}} (y_i - \mu)^2 f(y_i) \quad (\text{S.81})$$

$$= (0 - \pi)^2 \times (1 - \pi) + (1 - \pi)^2 \times (\pi) \quad (\text{S.82})$$

$$= \pi^2(1 - \pi) + \pi(1 - \pi)^2 \quad (\text{S.83})$$

$$= \pi^2 - \pi^3 + \pi - 2\pi^2 + \pi^3 \quad (\text{S.84})$$

$$= \pi(1 - \pi) \quad (\text{S.85})$$

The skew of a Bernoulli is

skew

$$\gamma_3(Y) = \mathbb{E}\left[\left(\frac{Y - \mu}{\sigma}\right)^3\right] \quad (\text{S.86})$$

$$= \sum_{i \in \{0,1\}} \left(\frac{y_i - \mu}{\sigma}\right)^3 f(y_i) \quad (\text{S.87})$$

$$= \left(\frac{0 - \pi}{\pi(1 - \pi)}\right)^3 \times (1 - \pi) + \left(\frac{1 - \pi}{\pi(1 - \pi)}\right)^3 \times (\pi) \quad (\text{S.88})$$

... some algebra ...

$$= \frac{1 - 2\pi}{\sqrt{\pi(1 - \pi)}} \quad (\text{S.89})$$

Here, we are defining skew as the “third standardized moment.” Other definitions are available, including the Hildebrand ratio, which you may have learned in your introductory statistics course:

Hildebrand

$$H = \frac{\overline{Y} - \tilde{Y}}{S} \quad (\text{S.90})$$

Example 9

What is the Hildebrand ratio for a Bernoulli distribution?

Solution: Because the median of the Bernoulli is not a smooth function, let us break this proof into three cases. Note, we will assume $\pi \in (0, 1)$. If $\pi = 0$ or $\pi = 1$, the statistics are not interesting.

Case 1 ($\pi > 0.500$). Let $Y \sim \text{Bern}(\pi)$. Let $\pi > 0.500$. From the probability mass function and the definition of the Hildebrand ratio, we have the following:

$$H = \frac{\bar{Y} - \tilde{Y}}{S} \tag{S.91}$$

$$= \frac{\pi - 1}{\pi(1 - \pi)} \tag{S.92}$$

$$= -\frac{1}{\pi} \tag{S.93}$$

Since this is always negative, we know that the Bernoulli is negatively skewed if $\pi > 0.500$.

Case 2 ($\pi > 0.500$). As previously, we have the following:

$$H = \frac{\bar{Y} - \tilde{Y}}{S} \quad (\text{S.94})$$

$$= \frac{\pi - 0}{\pi(1 - \pi)} \quad (\text{S.95})$$

$$= \frac{1}{1 - \pi} \quad (\text{S.96})$$

Since this is always positive, we know that the Bernoulli is positively skewed if $\pi < 0.500$.

Case 3 ($\pi = 0.500$). As previously, we have the following:

$$H = \frac{\bar{Y} - \tilde{Y}}{S} \quad (\text{S.97})$$

$$= \frac{\pi - \pi}{\pi(1 - \pi)} \quad (\text{S.98})$$

$$= 0 \quad (\text{S.99})$$

Thus, the Bernoulli is symmetric only when $\pi = 0.500$. ♦

In the previous example, I required $\pi \notin \{0, 1\}$. This is not a restriction for applied statistics. Why? What does it mean if $\pi = 0$ or $\pi = 1$? When would such a thing happen? Why would we need to study it?



kurtosis

The kurtosis of a Bernoulli is

$$\gamma_4(Y) = \mathbb{E}\left[\left(\frac{Y - \mu}{\sigma}\right)^4\right] \quad (\text{S.100})$$

$$= \sum_{i \in \{0,1\}} \left(\frac{y_i - \mu}{\sigma}\right)^4 f(y_i) \quad (\text{S.101})$$

$$= \left(\frac{0 - \pi}{\pi(1 - \pi)}\right)^4 \times (1 - \pi) + \left(\frac{1 - \pi}{\pi(1 - \pi)}\right)^4 \times (\pi) \quad (\text{S.102})$$

... some algebra ...

$$= \frac{1 - 3\pi(1 - \pi)}{\pi(1 - \pi)} \quad (\text{S.103})$$

Here, we are defining kurtosis as the “fourth standardized moment.”

The usual measure is called the “excess kurtosis.” This is just the kurtosis minus 3. Why 3? The kurtosis of the Normal distribution is 3. Thus, the **excess kurtosis** measures how its kurtosis differs from the Normal. Thus, the excess kurtosis of the Bernoulli is

$$\frac{1 - 6\pi(1 - \pi)}{\pi(1 - \pi)} \quad (\text{S.104})$$

Note: While the Bernoulli distribution is heavily used only in Chapter 12, it is a simple distribution that serves as a basis for better understanding other probability distributions.

S.4.2 BINOMIAL The Binomial distribution arises from modeling independent repeated trials where the variable is the number of successes out of a known number of trials.

Definition S.19. *Binomial Random Variable*

Let $Y_i \stackrel{iid}{\sim} \text{Bern}(\pi)$. If we define $X = \sum_{i=1}^n Y_i$, then $X \sim \text{Bin}(n, \pi)$.

Here, the symbol $\stackrel{iid}{\sim}$ indicates the random variables are independent and identically distributed; they are iid. In other words, they constitute a **random sample**.

iid

Interpreting the above definition gives us the following five requirements for a random variable to follow a Binomial distribution:

1. The number of trials, n , is known.
2. Each trial has two possible outcomes: failure (0) and success (1).
3. The probability of a success, π , does not change from trial to trial.
4. The trials are independent.
5. The random variable is the number of successes in those n trials.

If your random variable follows all five of these conditions, then it follows a Binomial distribution with parameters n and π . The sample space of a Binomial random variable is $\mathcal{S} = \{0, 1, 2, \dots, n\}$

The probability mass function (pmf) of a Binomial random variable is

$$\mathbb{P}[Y = y; n, \pi] = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad (\text{S.105})$$

The Binomial distribution is symmetric only when $\pi = 0.500$. If $\pi < 0.500$, then it is skewed right. Otherwise, it is skewed left. The sample space for the Binomial distribution is $\mathcal{S} = \{0, 1, \dots, n\}$. The expected value is $\mathbb{E}[Y] = n\pi$ and the variance is $\mathbb{V}[Y] = n\pi(1 - \pi)$.

Under certain circumstances, the Binomial distribution can be approximated with the Normal distribution. This arises from the Central Limit Theorem (Section S.6.4). This is especially important if we are examining the distribution of a proportion instead of a count.

CLT

Lemma S.20. Let $Y \sim \text{Bin}(n, \pi)$. The distribution of the number of successes is

$$Y \dot{\sim} \mathcal{N}(n\pi, n\pi(1 - \pi)) \quad (\text{S.106})$$

Here, the symbol $\dot{\sim}$ indicates the distribution is approximate. As one would expect, the approximation improves as the sample size, n , increases. This is due entirely to the Central Limit Theorem (Section S.6.4) and the fact that a Binomial random variable is the sum of independent and identically distributed (iid) Bernoulli random variables.

Also, the approximation improves if any of a variety of “continuity corrections” are used.

Lemma S.21. Let $Y \sim \text{Bin}(n, \pi)$. The distribution of the sample proportion, $P := \frac{Y}{n}$ is

$$P \dot{\sim} \mathcal{N}\left(\pi, \frac{\pi(1 - \pi)}{n}\right) \quad (\text{S.107})$$

As one would expect (Section S.6.4), the approximation improves as the sample size, n , increases.

exercise

Proof. This proof proceeds from approximating the distribution of X with a Normal distribution (Lemma S.20), then using the characteristics of the Normal distribution to obtain the answer. I leave this as an exercise. \square

S.4.3 POISSON The Poisson distribution arises from counting the number of successes over a time period or an area. Contrast this with the Binomial, where the successes were counted over the number of trials.

count

The probability mass function of the Poisson distribution is

$$\mathbb{P}[Y = y; \lambda] = \frac{e^{-\lambda} \lambda^y}{y!} \quad (\text{S.108})$$

The expected value is $\mathbb{E}[Y] = \lambda$, and the variance is $\mathbb{V}[Y] = \lambda$. The sample space is $\mathcal{S} = \{0, 1, 2, \dots\}$. It is skewed right, regardless of the value of λ , however that skew goes to zero as $\lambda \rightarrow \infty$. If the skew is defined as the standardized third moment, we can “easily” show that

skew

$$\gamma_3 = \mathbb{E}\left[\left(\frac{Y - \mu}{\sigma}\right)^3\right] = \lambda^{-1/2} \quad (\text{S.109})$$

Similarly, if we define the kurtosis as the standardized fourth moment, we know the *excess kurtosis* converges to zero as $\lambda \rightarrow \infty$:

kurtosis

$$\text{excess kurtosis} = \mathbb{E}\left[\frac{(X - \mu)^4}{\sigma^4}\right] - 3 = \lambda^{-1} \quad (\text{S.110})$$

These previous results show that the Poisson distribution becomes more and more Normal as $\lambda \rightarrow \infty$.

Note: The Poisson distribution is an example of an “**infinitely divisible**” distribution. This means that any Poisson distribution is the sum of two other Poisson distributions. This characteristic is rare, but it also holds for the Normal distribution. It is important because the Central Limit Theorem applies to sums of random variables. Since the Poisson distribution is a sum of other Poisson distributions, the CLT tells us that the Poisson distribution converges to the Normal distribution (as $\lambda \rightarrow \infty$).

infinitely
divisible

The Poisson distribution is the number of successes over a time period or an area. This suggests that the Poisson distribution arises as a limiting case of the Binomial distribution when $n \rightarrow \infty$, and $n\pi$ is a constant value. We prove this in the next theorem.

Theorem S.4.1

If $Y_n \sim \text{Bin}(n, \pi)$, then

$$\lim_{n \rightarrow \infty} Y_n \sim \mathcal{P}(\lambda) \quad (\text{S.111})$$

as long as $n\pi = \lambda$ remains constant.

Proof. This proof heavily relies on Sterling's approximation, $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$, which you could have seen in your calculus course.

$$\mathbb{P}[Y = y; n, \pi] = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad (\text{S.112})$$

$$= \frac{n!}{(n-y)! y!} \pi^y (1 - \pi)^{n-y} \quad (\text{S.113})$$

$$\approx \frac{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n}{\left(\sqrt{2\pi(n-y)} \left(\frac{n-y}{e}\right)^{n-y}\right) y!} \pi^y (1 - \pi)^{n-y} \quad (\text{S.114})$$

$$= \sqrt{\frac{n}{n-y}} \frac{n^n e^{-y}}{(n-y)^{n-y} y!} \pi^y (1 - \pi)^{n-y} \quad (\text{S.115})$$

Now, let $n \rightarrow \infty$ to give us

$$= 1 \frac{n^n e^{-y}}{(n-y)^{n-y} y!} \pi^y (1 - \pi)^{n-y} \quad (\text{S.116})$$

and holding $n\pi = \lambda$ (i.e., $\pi = \lambda/n$), we have

$$\mathbb{P}[Y = y; n, \pi] = \frac{n^n e^{-y}}{n^{n-y} \left(1 - \frac{y}{n}\right)^{n-y} y!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \quad (\text{S.117})$$

$$= \frac{\lambda^y \left(1 - \frac{\lambda}{n}\right)^{n-y} e^{-y}}{\left(1 - \frac{y}{n}\right)^{n-y} y!} \quad (\text{S.118})$$

As $n \rightarrow \infty$, we have $n - y \approx n$. This gives

$$\mathbb{P}[Y = y; n, \pi] \approx \frac{\lambda^y \left(1 - \frac{\lambda}{n}\right)^n e^{-y}}{\left(1 - \frac{y}{n}\right)^n y!} \quad (\text{S.119})$$

Remember from calculus that $n \rightarrow \infty$ means $\left(1 - \frac{y}{n}\right)^n \rightarrow e^{-y}$, by definition. Now, applying this limit, we have

$$\mathbb{P}[Y = y; n, \pi] = \frac{\lambda^y e^{-\lambda} e^{-y}}{e^{-y} y!} \quad (\text{S.120})$$

This simplifies to

$$\mathbb{P}[Y = y; n, \pi] = \frac{\lambda^y e^{-\lambda}}{y!} \quad (\text{S.121})$$

This is the probability mass function of the Poisson distribution. \square

Thus, we have shown that the Poisson distribution is the limiting distribution of a Binomial distribution when the number of trials goes to infinity and the expected value remains constant (the probability of success goes to zero).

Stop and think about why this fact suggests that the Poisson can model successes over time or space.

Note: The reason for understanding the Poisson distribution is that it is used to model counts (number of successes over time or space). It is the focus of count regression of Chapter 14.

Gaussian

i.i.d.

S.4.4 NORMAL The Normal distribution, also known as the Gaussian distribution and the Gauss-Laplace distribution, is ubiquitous in statistics. This is due to the Central Limit Theorem (see § S.6.4), which states that the distribution of the sample sum (or sample mean) approaches a Normal distribution, regardless of how the original variable is distributed (as long as the variance is finite and the sample is random).

The probability density function of the Normal is

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}\right] \quad (\text{S.122})$$

The Normal distribution is symmetric, has expected value $\mathbb{E}[Y] = \mu$, variance $\mathbb{V}[Y] = \sigma^2$, and sample space $\mathcal{S} = \mathbb{R}$, all real values.

Example 10

Calculate the skew of the Normal distribution.

Solution: There are several ways to show that the Normal is symmetric (skew = 0). Here are three.

Method 1: Math. The mathematical definition of symmetry is that $f(y)$ is symmetric about the vertical line μ if $f(\mu - y) = f(\mu + y)$. Here is the proof:

$$f(\mu - y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(\mu - y - \mu)^2}{\sigma^2}\right] \quad (\text{S.123})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(-y)^2}{\sigma^2}\right] \quad (\text{S.124})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y)^2}{\sigma^2}\right] \quad (\text{S.125})$$

and

$$f(\mu + y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(\mu + y - \mu)^2}{\sigma^2}\right] \quad (\text{S.126})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y)^2}{\sigma^2}\right] \quad (\text{S.127})$$

Thus since $f(\mu - y) = f(\mu + y)$, we have shown $f(y)$ is symmetric about μ .

Method 2: Hildebrand. We can also use the Hildebrand rule to show that the Normal distribution is symmetric:

$$H = \frac{\bar{Y} - \tilde{Y}}{S} \tag{S.128}$$

$$= \frac{\mu - \mu}{\sigma} \tag{S.129}$$

$$= 0 \tag{S.130}$$

Thus, since $H = 0$, the Normal distribution is symmetric about μ .

Note that I skipped over the part where I prove $\tilde{Y} = \mu$. I leave that for later (Lemma S.22).

Method 3: Third Standardize Moment. We can also use the third standardized moment to show that the Normal distribution is symmetric:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y - \mu)^2}{\sigma^2}\right] \tag{S.131}$$

$$\gamma_3 = \mathbb{E}\left[\left(\frac{Y - \mu}{\sigma}\right)^3\right] \tag{S.132}$$

$$= \int_{\mathbb{R}} f(y) \left(\frac{y - \mu}{\sigma}\right)^3 dy \tag{S.133}$$

I leave it as an exercise to expand the cube, separately calculate $\mathbb{E}[Y^2]$ and $\mathbb{E}[Y^3]$, and do the integration. There is a lot of algebra, but not much more than that if you are careful. \blacklozenge

There is a wealth of information on the Normal distribution. Arguably, it is the most studied distribution in statistics. It is also the most important distribution in statistics because of the Central Limit Theorem. For these reasons, I offer little to say about it beyond beyond the CLT (§ S.6.4).

Lemma S.22. *The median of a Normal distribution is μ .*

Solution: By definition for a continuous distribution, the median is the value \tilde{y} such that $F(\tilde{y}) = 0.500$. Thus, this proof reduces to a “mere calculation:”

$$F(\mu) = \int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}\right] dy \quad (\text{S.134})$$

Since this equals 0.500, we have shown that μ is the median of a Normal distribution.

◆

S.4.5 CHI-SQUARE The Chi-square distribution arose from multiple areas. One was the need to model the variation of the sample. A second was in examining categorical variables. It is used in the Chi-square goodness-of-fit test and the Chi-square test of independence. In both of those cases, the test statistic only *approximately* follows a Chi-square distribution.

If you want it (and I'm not entirely sure why you would), here is the probability density function of the Chi-square distribution based on its one parameter, ν , the "number of degrees of freedom:"

$$f(y; \nu) = \frac{1}{2^{\nu/2} \Gamma(\frac{\nu}{2})} y^{\nu/2-1} e^{-y/2} \quad (\text{S.135})$$

where $0 < \nu$ and the gamma function is defined as

$$\Gamma(y) := \int_0^{\infty} t^{y-1} e^{-t} dt \quad (\text{S.136})$$

gamma function

Note: The Chi-square distribution is actually the distribution of the second fraction in the exponent in the Normal probability density function. That is, if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{(Y - \mu)^2}{\sigma^2} \sim \chi_{\nu=1}^2 \quad (\text{S.137})$$

Mahalanobis

The square root of this fraction is called the "Mahalanobis" distance. The Mahalanobis distance is unitless, scale-invariant, and takes into account the correlations of the data set. One sees it in data science applications, especially in clustering and outlier detection.

When you read probability papers from around the turn of the 20th century, you may see χ as a variable of interest. Originally, it was just a curvey x used to illustrate that the Normal distribution could approximate the Binomial distribution:

$$\chi = \frac{x - n\pi}{\sqrt{n\pi(1 - \pi)}} \quad (\text{S.138})$$

With this, $\chi \sim \mathcal{N}(0, 1)$.

At the start of the 20th century, statisticians were able to turn this equation into a **definition** for the Chi-square distribution:

Definition S.23. *The Chi-square Distribution*

Let Z_i be a random sample of size ν from a standard Normal distribution, that is $Z_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$, then

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi_{\nu}^2 \quad (\text{S.139})$$

In this definition, ν (pronounced “nu”) is the number of degrees of freedom, the number of those Normal distributions that are independent.

It is *this* definition that is most helpful in determining what is (and what is *not*) a Chi-square random variable.

The expected value of a Chi-square distribution is ν . The variance is 2ν . The sample space is $\mathcal{S} = (0, \infty)$. It is always positively skewed ($\sqrt{8/\nu}$), although that skew goes to zero as $\nu \rightarrow \infty$. Its excess kurtosis also goes to 0 as $\nu \rightarrow \infty$ (ex.kurt. = $12/\nu$). If the number of degrees of freedom are 2 or less, then the probability density function is strictly decreasing (its first derivative is always less than zero). Among other things, this means that the mode is zero if $\nu \leq 2$.

S.4.6 STUDENT'S T Arguably, most of the previous distributions arise from Nature. The Student's t distribution arises from a need to test certain hypotheses. This distribution is named after its creator, William Sealy Gosset (as pictured in the figure to the right).

Gosset worked for Guinness Brewery as a master brewer (and statistician) near the end of the 19th century. He applied the well-known statistical techniques in his job, but found that these techniques had major flaws. Apparently, he felt that he was rejecting far too many bushels of grain based on current statistical techniques.

Specifically, Gosset determined that the test statistic

$$Z = \frac{Y - \mu}{s/\sqrt{n}} \quad (\text{S.140})$$

did not follow a standard Normal distribution as expected, when the sample size was small. In fact, it was not even “sufficiently” close.

The problem was that Gosset was working with small samples of barley, while most statistics at the time were concerned with large samples. In my opinion, this is Gosset's greatest contribution: paying attention to small-sample properties of estimators.

While Guinness supported him, they did not want a repeat of a previous employee who published trade secrets in a scientific journal. Thus, to get his discoveries out there, Gosset had to publish under a pseudonym. He chose “Student.”



Figure S.1:
William Sealy Gosset (1908).

z-score



Figure S.2:
Ronald Aylmer Fisher (1913).

finite samples

The following is Gosset's **definition** of his Student's t distribution.

Definition S.24. *Student's t Distribution*

Let $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_\nu$ with Z and V independent. Define the following ratio:

$$T = \frac{Z}{\sqrt{V/\nu}} \quad (\text{S.141})$$

The random variable T follows a Student's t distribution with ν degrees of freedom.

If you care to see it (and some do like the mathematics), this is the probability density function calculated by Fisher:

$$f(y; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (\text{S.142})$$

Again, $\Gamma(\cdot)$ is the gamma function.

The mean of the t distribution is $\mathbb{E}[Y] = 0$ if $\nu > 1$. The variance is $\mathbb{V}[Y] = \frac{\nu}{\nu-2}$ if $\nu > 2$, ∞ if $\nu \in (1, 2]$, and undefined elsewhere. It has zero skew. Its sample space is $\mathcal{S} = \mathbb{R}$. As $\nu \rightarrow \infty$, the Student's t distribution converges to the Normal distribution.⁶

Note: While the t-distribution was formulated by Gosset to deal with the distribution of a test statistic, it was actually first seen as a posterior distribution a couple decades earlier. However, as Bayesian inference and frequentist statistics were rarely aware of each other, the results by Helmert and Lüroth remained unknown to Gosset and Fisher.

Bayes

⁶The proof of this is an excellent exercise in exponentials and Stirling's approximation. I leave it as an exercise for you.

S.4.7 CAUCHY The Cauchy distribution is named after Augustin-Louis Cauchy, a French mathematician who specialized in complex analysis and abstract algebra. It is originally (and most helpfully) **defined** as

Definition S.25. *Cauchy Distribution*

Let $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 \sim \mathcal{N}(0, 1)$, and Z_1 and Z_2 be independent. Then, the ratio

$$Y := \frac{Z_1}{Z_2}$$

follows a standard Cauchy distribution.

Note that this is just a t distribution with one degree of freedom. As such, the probability density function for the standard Cauchy is

$$f(y) = \frac{1}{\pi(1+y^2)} \tag{S.143}$$

Regardless of the fact that the standard Cauchy is symmetric with median 0, neither its mean nor its variance exist. I leave this proof as an exercise for you.

The standard Cauchy can be generalized to different medians (locations) and spreads (scales):

$$f(y; \eta, \gamma) := \frac{1}{\pi\gamma\left(1 + \left(\frac{y-\eta}{\gamma}\right)^2\right)} \tag{S.144}$$

Again, neither the mean nor the variance exist, however the median and mode are now η (“eta”) and the interquartile range is 2γ (“gamma”). [I have you prove this later.]

Finally, its cumulative distribution function (CDF) is

$$F(y; \eta, \gamma) = \frac{1}{\pi} \arctan\left(\frac{y-\eta}{\gamma}\right) + \frac{1}{2} \tag{S.145}$$

As befits the magical world of mathematics, the Cauchy distribution pops up in many places. For me, the most interesting place is as the Witch of Agnesi.

I find it very interesting to see how frequently mathematical equations find their way into many different areas of mathematics and science. The key, as I see



Figure S.3: Baron Augustin-Louis Cauchy (1901).

exercise

exercise

CDF

it, is to understand how the mathematical expressions came into being, then see what applies to the current probability/statistics problem. A trip through history may suggest that there is “no progress in the history of [mathematical] knowledge — merely a continuous and sublime recapitulation” (Umberto Eco, *The Name of the Rose*, 1980).

Note: The Cauchy distribution is one of the most helpful distributions available. Its first (and above) moments are not finite. This means the Central Limit Theorem does *not* apply to it. As such, it is frequently used to illustrate the importance of finite variances.

S.4.8 SNEDECOR'S F The F distribution, developed by George W. Snedecor (*Calculation and Interpretation of Analysis of Variance and Covariance*, 1934) and named for Fisher, is frequently used for testing compound hypotheses, notably in the analysis of variance (ANOVA) procedure (see page 64).

Definition S.26. *Snedecor's F Distribution*

Let $X_1 \sim \chi^2(\nu_1)$, $X_2 \sim \chi^2(\nu_2)$, and X_1 and X_2 independent (i.e., $X_1 \perp X_2$). Define the following ratio:

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \quad (\text{S.146})$$

The random variable F follows Snedecor's F distribution with ν_1 and ν_2 degrees of freedom.

If $\nu_2 > 2$, then the mean of the F distribution is

$$\mathbb{E}[F] = \frac{\nu_2}{\nu_2 - 2} \quad (\text{S.147})$$

When $\nu_2 > 4$, its variance is

$$\mathbb{V}[F] = \frac{2\nu_2^2(\nu_1 + \nu_2 + 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad (\text{S.148})$$

If $\nu_2 \leq 2$, then the expected value is infinite. If $\nu_2 \leq 4$, then the variance is infinite. The distribution (function) is always right-skewed. The support set is $\mathcal{S} = (0, \infty)$.

If you want it, here is its probability density function. Note, however, that the definition of the F distribution is **much more helpful**.

$$f(y; \nu_1, \nu_2) = \frac{1}{y B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \sqrt{\frac{y^{\nu_1} \nu_1^{\nu_1} \nu_2^{\nu_2}}{(y \nu_1 + \nu_2)^{\nu_1 + \nu_2}}} \quad (\text{S.149})$$

In this formula, $B(\cdot, \cdot)$ is the beta function, defined as

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt \quad (\text{S.150})$$

Interesting... There seems to be some relationship between the beta function and the Binomial distribution.

By the way, if you would prefer writing the beta function in terms of the gamma function,

$$B(x, y) = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x+y)} \quad (\text{S.151})$$

Most likely, you will have seen the gamma function in your calculus class. If not, there is nothing to worry about, since the computer will be performing the calculations for you. Focus on **how these distributions arise** (i.e., their actual *definition*) and how they are used.

S.5: Distributions of Sample Statistics

Now, let us combine the previous two sections. It is knowing the distributions of the sample statistics that the power of statistics first shows itself. It allows us to draw conclusions about the population parameter based solely on the sample statistic.

S.5.1 SAMPLE MEAN The distribution of the sample mean is one of the first sampling distributions you would have come across in your introductory statistics course. It usually serves as the link between the Normal distribution and confidence intervals.

Theorem S.5.1

If $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, and we collect an independent and identically distributed (iid) sample of size n , then

$$\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (\text{S.152})$$

Proof. This proof proceeds in three parts. The first part notes that a linear combination of independent Normal random variables is also a Normal random variable. (The proof of this can be found at Corollary S.37.)

Since the Normal distribution has two parameters, mean and variance, the second and third parts determine the expected value and variance of that random variable.

$$\mathbb{E}[\bar{Y}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] \quad (\text{S.153})$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i] \quad (\text{S.154})$$

$$= \frac{1}{n} \sum_{i=1}^n \mu \quad (\text{S.155})$$

$$= \frac{1}{n} n\mu \quad (\text{S.156})$$

Thus, the expected value of \bar{Y} is μ . Now for the variance:

$$\mathbb{V}[\bar{Y}] = \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] \quad (\text{S.157})$$

$$= \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n Y_i\right] \quad (\text{S.158})$$

Since the Y_i values are independent, we can pass the variance operator through the summation:

$$= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}[Y_i] \quad (\text{S.159})$$

$$= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 \quad (\text{S.160})$$

$$= \frac{1}{n^2} n\sigma^2 \quad (\text{S.161})$$

$$= \frac{\sigma^2}{n} \quad (\text{S.162})$$

Putting these parts together gives us our conclusion. \square

Note: Notice the procedure in the previous proof. The first step is to determine the distribution. The remaining steps determine the values of the distribution's parameters. We needed to determine the values of μ and σ^2 because the Normal distribution uses both as parameters.

S.5.2 SAMPLE VARIANCE The distribution of the sample variance is rarely covered in introductory statistics. However, it is needed when determining some important test statistics.

Theorem S.5.2

If $Y_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and we collect a random sample of size n , then

$$\frac{(n-1)S_y^2}{\sigma^2} \sim \chi_{v=n-1}^2 \quad (\text{S.163})$$

Proof. Since we need to show that it follows a Chi-square distribution, we simply use the definition of that distribution, performing some algebra to ensure it is in the right form.

First, since we will need it in the future of this proof, please recall

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (\text{S.164})$$

This is equivalent to

$$(n-1)S_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (\text{S.165})$$

Now, remember the definition of a Chi-square random variable:

$$\sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma} \right)^2 \sim \chi_n^2 \quad (\text{S.166})$$

Note that this is very similar to our definition of S^2 . The difference is the presence of μ . So, for reasons that will become obvious later, let us subtract and add \bar{Y} to the numerator of S.166, then expand the square:

$$\chi_n^2 \sim \sum_{i=1}^n \left(\frac{Y_i - \bar{Y} + \bar{Y} - \mu}{\sigma} \right)^2 \quad (\text{S.167})$$

$$= \sum_{i=1}^n \left(\frac{(Y_i - \bar{Y}) + (\bar{Y} - \mu)}{\sigma} \right)^2 \quad (\text{S.168})$$

$$= \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{Y} - \mu}{\sigma} \right)^2 + 2 \sum_{i=1}^n \left(\frac{(Y_i - \bar{Y})(\bar{Y} - \mu)}{\sigma} \right) \quad (\text{S.169})$$

Note that the third term is zero:

$$2 \sum_{i=1}^n \left(\frac{(Y_i - \bar{Y})(\bar{Y} - \mu)}{\sigma} \right) = 2 \left(\frac{(\bar{Y} - \mu)}{\sigma} \right) \sum_{i=1}^n (Y_i - \bar{Y}) \quad (\text{S.170})$$

$$= 2 \left(\frac{(\bar{Y} - \mu)}{\sigma} \right) 0 \quad (\text{S.171})$$

$$= 0 \quad (\text{S.172})$$

With that, we have:

$$\chi_n^2 \sim \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{Y} - \mu}{\sigma} \right)^2 \quad (\text{S.173})$$

$$= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} + n \left(\frac{\bar{Y} - \mu}{\sigma} \right)^2 \quad (\text{S.174})$$

$$= \frac{(n-1)S^2}{\sigma^2} + n \left(\frac{\bar{Y} - \mu}{\sigma} \right)^2 \quad (\text{S.175})$$

$$\text{And: } \chi_n^2 \sim \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \right)^2 \quad (\text{S.176})$$

The second term is the square of a standard Normal distribution; that is, it follows a χ_1^2 distribution. Since the sum of two Chi-square distributions is another Chi-square distribution (with ν being the sum of the degrees of freedom), we have our result:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (\text{S.177})$$

□

Note: Do you remember from Section S.4.5 that the expected value of a Chi-square random variable is ν ? Use that to see why we needed to divide by $n-1$ in the formula for the sample variance to ensure S^2 is unbiased for σ^2 . That is, you should be able to prove $\mathbb{E}[S^2] = \sigma^2$.

exercise

By the way, there is another proof using algebra, but it gives little insight into probability. Because of this, I forgo it. However, it is a really nice exercise in high school algebra and statistics definitions.

S.6: Other Topics

This section holds some topics that I do not know where to put otherwise. You will have discussed the effects of multiple comparisons on the legitimacy of statistical conclusions (Section S.6.2). That is why you had to use analysis of variance instead of multiple t-tests for comparing more than two population means.

Next, I am introducing the runs test here (Section S.6.3). The runs test is important in testing the claim that the model systematically fits the observed data. The logic behind it is very straight-forward, being based on a Binomial random variable.

Finally, I provide a proof of the Central Limit Theorem. You will most certainly have heard of the Central Limit Theorem in your earlier statistics course but not seen its proof. However, the work below with moment generating functions is provided solely for those who are interested in how to prove the Central Limit Theorem (Section S.6.4).

S.6.1 CHECK FOR SKEWNESS There are several tests for skewness, and for estimating the level of skew in a distribution. However, most are beyond the scope of this book. In lieu of these, let us use the Hildebrand Rule as a good Rule of Thumb for determining if the data are “sufficiently” skewed (Hildebrand 1986).

For the sample, the ratio is defined as

$$H = \frac{\bar{x} - \tilde{x}}{s} \quad (\text{S.178})$$

According to Hildebrand, if this ratio is greater than 0.20, then the data (or population) are positively skewed. If this ratio is less than -0.20 , then the data (or population) are negatively skewed. Otherwise, there is no significant skew.

Of course, this is just a Rule of Thumb that completely ignores the sample size and precision of the estimate. However, it is not terrible. If, however, the skew of the data is important, one should use a genuine statistical test.

consistently

CLT

S.6.2 MULTIPLE COMPARISONS Remember from your previous statistics class the problem with multiple comparisons. When performing multiple tests, one needs to adjust the value of the p-value to ensure that you do not reject true null hypotheses at too high a rate (α).

The first part illustrates why. The second part introduces the Bonferroni adjustment. Note that the Bonferroni *always* “works,” but it tends to be rather conservative. That is, it controls the Type I error rate, but at the expense of reducing the power of the test (higher Type II error rate, β).

power

MULTIPLE TESTING: To start this section, recall the meaning of α . It is the proportion of the time we wrongly reject a true null hypothesis (i.e., commit a Type I error). Since we are basing our statistical inference on our *selected* value of α , we want to ensure that we are rejecting at the right frequency (or with the right probability). At the very least, we want to ensure that we do not reject too frequently. If we reject too frequently, we are biasing our conclusions towards rejecting null hypotheses when we should not.⁷

To see the effect of multiple tests on the probability of rejecting a true null hypothesis (committing a Type I error), let us look at the following two examples.

Example 11

Let the null hypothesis be true. Let us also collect a sample and test the null hypothesis using that sample. What is the probability we reject the true null hypothesis if we state our accepted Type I error rate is α ?

Solution: By definition, as long as the test is appropriate for the null hypothesis, the probability of wrongly rejecting the null hypothesis is α . ♦

⁷If we reject too *infrequently*, then we will fail to reject the null hypothesis too often. This could also be a bad thing. It really depends on the penalty for being each type of wrong. In the framework of the traditional hypothesis testing, we wish to avoid rejecting the null hypothesis too frequently. This leads to the focus on ensuring we reject the null hypothesis at the α rate or less. Chapter 12 explores consequences of only controlling α .

Example 12

Let the null hypothesis be true. Let us collect a sample and test the null hypothesis using that sample. Let us now collect a second, independent, sample and test the null hypothesis. What is the probability we reject the true null hypothesis in *either* sample if we state our Type I error rate is α ?

each

Solution: By definition, for an appropriate test of the null hypothesis, the probability of wrongly rejecting the null hypothesis in *each* test is α .

Let us define Y as the number of samples for which the test rejects the null hypothesis. Clearly, if the tests are independent (if the samples are independent), then $Y \sim \text{Bin}(2, \alpha)$, and we need to calculate $\mathbb{P}[Y \geq 1]$.

As such, the probability is

$$\mathbb{P}[Y \geq 1] = 1 - \mathbb{P}[Y = 0] \tag{S.179}$$

$$= 1 - \binom{2}{0} \alpha^0 (1 - \alpha)^2 \tag{S.180}$$

If $\alpha = 0.05$, then $\mathbb{P}[Y \geq 1] = 0.0975$. In other words, we claim the Type I error rate is 0.05, but it is really 0.0975... almost twice as much!! \blacklozenge

Allow me to repeat that result: If we test the null hypothesis twice, then the *actual* Type I error rate is almost twice what we claim. This is very problematic, especially for hypotheses that require several sub-tests to fully test, or for times we test multiple hypotheses using the same data.

BONFERRONI ADJUSTMENT: One of the first ways for ensuring that the real Type I error rate is not larger than the claimed is the Bonferroni adjustment (Bonferroni 1936; Dunn 1958; Dunn 1961). While the Bonferroni adjustment always controls the Type I error rate, it is only the best option when nothing else is available; it reduces the power of the test (increases the probability of a Type II error).

Theorem S.6.1

Let the target Type I error rate be α , with k tests being performed. To ensure that the experimentwise Type I error rate is no greater than α , each test should be rejected at the α/k level.

Proof. If we reject each test at the α/k level, the experiment-wise error rate for the k tests is

$$\text{EWER} = \mathbb{P} \left[\bigcup_{i=1}^k \mathbb{I}_{\{P_i \leq \frac{\alpha}{k}\}} \right] \quad (\text{S.181})$$

$$\leq \sum_{i=1}^k \mathbb{P} \left[P_i \leq \frac{\alpha}{k} \right] \quad (\text{S.182})$$

$$= \sum_{i=1}^k \frac{\alpha}{k} \quad (\text{S.183})$$

$$= k \frac{\alpha}{k} = \alpha \quad (\text{S.184})$$

Note that $\mathbb{I}_{\{P_i \leq \frac{\alpha}{k}\}}$ equals 1 when $P_i \leq \frac{\alpha}{k}$ and 0 otherwise. The function $\mathbb{I}_{\{\cdot\}}$ is called the “indicator function” for this very reason. It ‘indicates’ if the condition in the braces is true. (What is the distribution of the indicator function?)

indicator

Also note that P_i is the p-value for each test. If the test is appropriate, then the p-value follows a standard uniform distribution.

With those reminders, the rest is just algebra. \square

Note: Note the inequality. Under what circumstances is it an equality?

The Bonferroni adjustment always “works.” In other words, applying the Bonferroni adjustment always ensures that the Type I error rate for the entire battery of tests is at most α . However, it is usually a poor adjustment in that it over-corrects for multiple comparisons.

To see this, let us make the assumption that the tests are independent. This additional information allows us to create a tighter adjustment.

Theorem S.6.2

Let the target Type I error rate be α , with k independent tests being performed. To ensure that the experiment-wise Type I error rate is no greater than α , each test should be rejected at the $1 - (1 - \alpha)^{1/k}$ level.

Proof. If we reject each test at the $1 - (1 - \alpha)^{1/k}$ level, the experiment-wise error rate for the k tests is

$$\text{EWER} = \mathbb{P} \left[\bigcup_{i=1}^k \mathbb{I}_{\{P_i \leq 1 - (1 - \alpha)^{1/k}\}} \right] \quad (\text{S.185})$$

By De Morgan's Laws, this is

$$= 1 - \mathbb{P} \left[\bigcap_{i=1}^k \mathbb{I}_{\{P_i > 1 - (1 - \alpha)^{1/k}\}} \right] \quad (\text{S.186})$$

By independence, this is

$$= 1 - \prod_{i=1}^k \mathbb{P} \left[P_i > 1 - (1 - \alpha)^{1/k} \right] \quad (\text{S.187})$$

$$= 1 - \prod_{i=1}^k \left(1 - \mathbb{P} \left[P_i \leq 1 - (1 - \alpha)^{1/k} \right] \right) \quad (\text{S.188})$$

$$= 1 - \prod_{i=1}^k \left(1 - (1 - (1 - \alpha)^{1/k}) \right) \quad (\text{S.189})$$

Finishing with algebra

$$\text{EWER} = 1 - \prod_{i=1}^k (1 - \alpha)^{1/k} \quad (\text{S.190})$$

$$= 1 - \left((1 - \alpha)^{1/k} \right)^k \quad (\text{S.191})$$

$$= 1 - (1 - \alpha) \quad (\text{S.192})$$

$$= \alpha \quad (\text{S.193})$$

Thus, if we know that the tests are independent, we can create a smaller multiplier. This will still protect the Type I error rate, while not affecting the Type II error rate as much. \square

Lemma S.27. *The original Bonferroni adjustment is a larger divisor (smaller multiplier; lower power) than the adjustment that assumes independence among the tests. In other words,*

$$\frac{\alpha}{k} \leq 1 - (1 - \alpha)^{1/k} \quad (\text{S.194})$$

I leave this proof as an exercise.

exercise

Theorem S.6.1 shows that we can adjust any case of multiple testing. However, Theorem S.6.2 and Corollary S.27 show that additional information (or assumptions) can help make adjustments that affect the test's power less.

In other words, additional information can help. We should not throw anything out.

Note: This section explored the need for adjusting the α to ensure that the claimed Type I error rate is no larger than the true Type I error rate. This is extremely important in inferential statistics. It is not, however, the only thing that needs to be examined. Power (the probability of rejecting a false null hypothesis) is also quite important. Because of this, while the Bonferroni adjustment always works, it may reduce the power of the test to unacceptable levels. As such, understanding the relationships between and among the various tests is important in creating a superior adjustment.

S.6.3 RUNS TEST This section introduces the runs test (Bradley 1968; Mood 1940; Wald and Wolfowitz 1940). While you will not have seen it in your introductory statistics course, it is based on elementary concepts.

To begin, let us define a couple of terms.

Definition S.28. *Run*

A run is a sequential set of values all either above or below zero.

Definition S.29. *Run Length*

The length of a run is the number of elements in that run.

Example 13

Let the following be residuals from a model.

1.2, 3.5, -0.5, 8.9, -1.2, -1.6, -4.5, 0.6, 1.5, -7.9

Let us calculate the number of runs.

Solution: The first run is of length 2: $\{1.2, 3.5\}$. The second run is of length 1: $\{-0.5\}$. The third run is also of length 1: $\{8.9\}$. The fourth through sixth runs are $\{-1.2, -1.6, -4.5\}$, $\{0.6, 1.5\}$, and $\{-7.9\}$.

Thus, there are six runs in those 10 residuals. ♦

Before continuing, remind yourself of Figures 5.4 and 5.5. In both graphics, the residual is colored blue if it is positive and pink otherwise. In Figure 5.4, there are long unbroken streaks (runs) of blue and pink. There are just three runs.

run

In Figure 5.5, the length of those runs is much reduced *and* the number is increased. There are now 12 runs in the 20 residuals.

Since the distribution of the residuals is assumed/required to be Normal, the probability of each residual being above the line $e = 0$ is $\pi = 0.500\dots$ a Binomial random variable! Because we know the distribution of the number of residuals above the $e = 0$ line, we know *everything* about the distribution of residuals, including the distribution of the number of runs. Too few runs indicates that one positive residual will tend to be followed by another. This usually suggests the model is misspecified. In Figure 5.4, the number of runs was just three. In the properly specified model of Figure 5.5, the number of runs is 12.



Eureka!

A statistician will pay attention to the number of runs as well as the *distribution of the number of runs under the null hypothesis*. In other words, what is the distribution of the number of runs?

frequentist

The following estimates that distribution. Working through the code may help you better understand the runs test, the test statistic, and its distribution.

```
resids = 20

calculateRuns = function(r) {
  n = length(r)
  x = sign(r)
  y = x[-n]-x[-1]
  runs = 1+sum(y!=0)
}

numRuns = numeric()
for(i in 1:1e5) {
  x = rnorm(resids)
  numRuns[i] = calculateRuns(x)
}
```

```

    }

    par(mar=c(2,1,1,1))
    par(yaxt="n")
    par(family="serif")
    barplot(table(numRuns), col=rgb(0.85, 0.37, 0.01))
    abline(h=0)

```

The first line specifies the number of residuals. Since I will be using this distribution to analyze the residuals from Section 5.2, I set the number to 20. The `calculateRuns` function calculates the number of runs in a sequence of values. The `for`-loop repeatedly generates values from the null hypothesis and calculates the number of runs.

The last chunk of code plots the histogram, which is shown in Figure S.5. Note that it is bell-shaped. This should give you an insight into how we can create an approximate test statistic based on the Normal distribution, which is what the functions do.

From this, the probability of having 7 or more runs is 95%. So, we would claim too few runs if we observed 6 or fewer runs in our 20 residuals. The p-value for observing 3 runs is 0.0003. Since this p-value is less than $\alpha = 0.05$, we reject the null hypothesis that the residuals are randomly distributed about the zero line. Thus, we would conclude that the relationship is not linear.

The following lines calculated the critical value and the p-value for the runs distribution.

```

quantile(numRuns, 0.05)      ## critical value
mean(numRuns<=3)            ## p-value

```

NORMAL APPROXIMATION*: In addition to the simulation experiment above, Wald and Wolfowitz (1940) provided a Normal approximation that did not require simulation. They concluded that the number of runs, R , approximately follows the following distribution:

$$R \sim \mathcal{N}\left(\frac{2N^+N^-}{N} + 1; \frac{(\mu-1)(\mu-2)}{N-1}\right) \quad (\text{S.195})$$

In these, N^+ is the number of positive values, N^- is the number of negative values, and N is the number of values. As with any “Normal approximation,” the approximation improves as the sample size increases.

THE EXACT DISTRIBUTION*: While Wald and Wolfowitz (1940) did create an approximate distribution for the number of runs, we can also provide the exact distribution that preceded them using the hypergeometric distribution. All that we need is to remember combinatorics and the “choose” (or nCr or binomial coefficient) function.

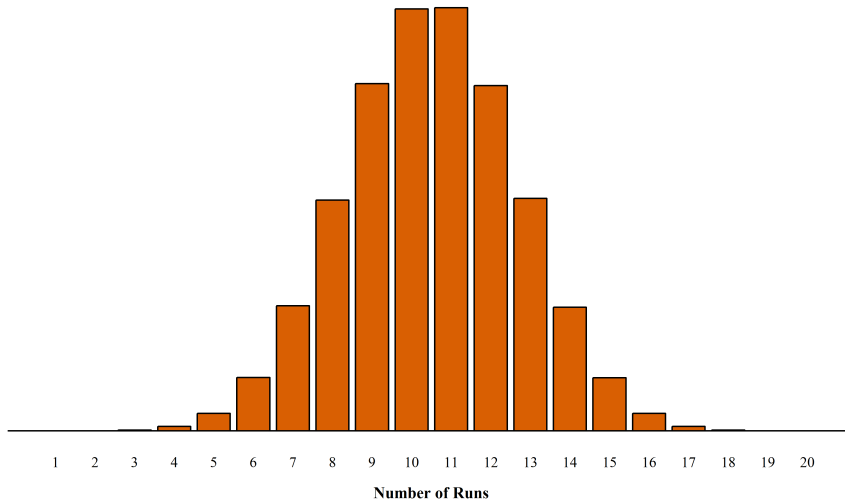


Figure S.5: *Distribution of the number of runs with 20 residuals.*

Under the null hypothesis, the probability of having exactly r runs, given a sample size of N , of which N^+ are positive values and N^- are negative, is

$$\mathbb{P}[R = r] = 2 \frac{\binom{N^+ - 1}{m - 1} \binom{N^- - 1}{m - 1}}{\binom{N}{N^+}} \quad (\text{S.196})$$

if r is even and $r = 2m$, and

$$\mathbb{P}[R = r] = \frac{\binom{N^+ - 1}{m} \binom{N^- - 1}{m - 1} + \binom{N^- - 1}{m} \binom{N^+ - 1}{m - 1}}{\binom{N}{N^+}} \quad (\text{S.197})$$

if r is odd and $r = 2m + 1$.

Those look rather “simple” to calculate. Remember, however, that p-values are cumulative probabilities. Thus, to calculate p-values, one would need to calculate $\mathbb{P}[R \leq r]$, which means a lot more calculation.

Computers are awesome!

THE R FUNCTIONS: Instead of going through the above steps every time to obtain p-values and critical values, the function is implemented in a couple packages in R. First, it is implemented in the `lawstat` package as the function `runs.test`. It takes just one piece of information: the residuals in the order of the independent variable. It is also implemented in the `randtests` and `snpar` packages, as well as the `RFS` add-on.

Note that order matters. Thus, you will need to order the residuals based on each independent variable separately. That is, if you have three independent variables, you will need to run the runs test three times, once for each independent variable.

S.6.4 THE CENTRAL LIMIT THEOREM The Central Limit Theorem (CLT) is one of the most important theorems in statistics. It explains why we can rely so heavily on the Normal distribution for inferential statistics, even if the data do not arise from a Normal process.

Theorem S.6.3: The Central Limit Theorem

Let X be a random variable with finite variance σ^2 . Let $\{X_1, X_2, \dots, X_n\}$ be a random sample of size n from this distribution. Finally, define T as

$$T_n := \sum_{i=1}^n X_i \quad (\text{S.198})$$

As $n \rightarrow \infty$, the distribution of T_n converges to the Normal distribution.

Note that this theorem requires that the random variable has a finite variance *and* that the sample is a random sample (independent and identically distributed). If either of these two conditions is not met, then this version of the Central Limit Theorem does not hold.

Other versions exist for some different types of dependent sampling. However, even those require the variance be finite.

In your previous statistics course, you may remember the Central Limit Theorem as saying something about the distribution of the sample mean. Well, that is actually a corollary of the CLT.

Corollary S.30. *Given the conditions of the Central Limit Theorem, define \bar{X} as*

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (\text{S.199})$$

The distribution of the sample mean, \bar{X}_n , converges to the Normal distribution.

To use the Central Limit Theorem, we need to ensure the data are randomly generated from a distribution with a finite variance. We also need to ensure that we are trying to describe the sample total or sample mean. If those conditions are met, then we can use the CLT.

How would we prove the Central Limit Theorem? To do that, we need to introduce something called “moment generating functions.” This we do in the next section. Continue, if you wish. Proving the Central Limit Theorem is beyond the scope of this type of course.

S.6.5 MOMENT GENERATING FUNCTIONS* This section introduces moment generating functions. You most certainly did not experience these in your previous statistics course. For us, they are used solely to prove the Central Limit Theorem (Theorem S.6.6) and to prove that a linear combination of Normally distributed random variables is also a Normally distributed random variable (Theorem S.37).

A moment generating function is a function that can be used to generate all of the moments of a distribution. We have already experienced moments.

moments

Definition S.31. *Definition*

The k^{th} (raw) moment is $\mathbb{E}[X^k]$.

Definition S.32. *Definition*

The k^{th} central moment is $\mathbb{E}[(X - \mu)^k]$, where $\mu = \mathbb{E}[X]$.

Definition S.33. *Definition*

The k^{th} standardized central moment is $\mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^k\right]$, where $\mu = \mathbb{E}[X]$ and $\sigma = \sqrt{\mathbb{V}[X]}$.

Thus, by definition, the first raw moment is $\mathbb{E}[X]$. This is just the mean. The second central moment is $\mathbb{E}[(X - \mu)^2]$. This is just the variance. Similarly, the third central moment is the skew (when standardized by the cube of σ), and the fourth central moment is the kurtosis (when standardized by the fourth power of σ).

exercise

Prove to yourself that the first central moment is always 0, as long as $\mathbb{E}[X]$ exists. (Note that this is a problem from your first Math-Stat course.)



Moments are interesting in their own right. However, the following theorem makes them important in the realm of probability theory.

Theorem S.6.4

If two distributions have all moments the same, then the two distributions are equal, with probability 1.

The “with probability 1” clause is added to take into consideration the inherent issues with continuous distributions. [It also clearly specifies the difference between probability (and statistics) and mathematics that some mathematicians may not appreciate.]

with probability 1

Here is an example of that inherent issue. Let us define two distributions X and Y . Let $f(x) = 1$ for $x \in (0, 1)$. Let $f(y) = 1$ for $y \in [0, 1]$. The random variables X and Y are not the same. The support for Y includes the entire support of X as well as the values 0 and 1. Thus, they are mathematically not the same distribution.

However, they are the same *with probability 1*. In other words, you would spend the rest of your life (and the life of the universe and the life of the multiverse) drawing samples from Y and never obtaining 0 or 1. Thus, for all intents and purposes, X and Y are the same. In probability, we would say $X = Y$ with probability 1, symbolized as $X \xrightarrow{wp1} Y$. In symbols, we would write $P[X \leq x] = P[Y \leq y]$.

With this theorem, it becomes clear that if two distributions have the same moment generating function, then those distributions are the same (with probability 1).

And so, at this point, we now what a moment is and what we can use them for. We know that if two distributions have all the same moments, then the two distributions are identical. This led to us knowing that if two distributions have the same moment generating functions, then the two distributions are the same. All that remains is to define a moment generating function (MGF).

Definition S.34. Moment Generating Function

Let X be a random variable. Define the function $M_x(t)$ as

$$M_x(t) := \mathbb{E}[e^{Xt}] \tag{S.200}$$

Then $M_x(t)$ is called the moment generating function for X .

Why is it called a “moment generating function”? It generates the raw moments. The k^{th} raw moment of X is calculated from

$$\mathbb{E}[X^k] = \left. \frac{d^k}{dt^k} M_x(t) \right|_{t=0} \quad (\text{S.201})$$

Note that the moment generating function must be defined in a small neighborhood of 0 for us to apply this formula.

ε -neighborhood

Example 14

Let X be a Bernoulli random variable with success probability π . What is its moment generating function?

Solution: The moment generating function is defined as $M_x(t) := \mathbb{E}[e^{Xt}]$. Since X is a discrete random variable, the MGF is a sum weighted by the probability mass function.

$$\mathbb{E}[e^{Xt}] = \sum_i e^{x_i t} \mathbb{P}[X = x_i] \quad (\text{S.202})$$

$$= \sum_{x=0}^1 e^{xt} \pi^x (1-\pi)^{1-x} \quad (\text{S.203})$$

$$= e^{0t} \pi^0 (1-\pi)^{1-0} + e^{1t} \pi^1 (1-\pi)^{1-1} \quad (\text{S.204})$$

$$= 1 \cdot 1 \cdot (1-\pi) + e^t \cdot \pi^1 \cdot 1 \quad (\text{S.205})$$

And, simplification gives us the MGF for a Bernoulli random variable:

$$\mathbb{E}[e^{Xt}] = (1-\pi) + \pi e^t \quad (\text{S.206})$$

Now, with that, we can calculate the mean of X :

$$\mathbb{E}[X^1] = \left. \frac{d^1}{dt^1} M_x(t) \right|_{t=0} \quad (\text{S.207})$$

$$= \left. \frac{d}{dt} \left((1-\pi) + \pi e^t \right) \right|_{t=0} \quad (\text{S.208})$$

$$= \left. \pi e^t \right|_{t=0} \quad (\text{S.209})$$

$$= \pi e^0 \quad (\text{S.210})$$

$$= \pi \quad (\text{S.211})$$

It should come as no surprise that the expected value of a Bernoulli random variable is the success probability. The second moment is

$$\mathbb{E}[X^2] = \left. \frac{d^2}{dt^2} M_x(t) \right|_{t=0} \quad (\text{S.212})$$

$$= \left. \frac{d^2}{dt^2} ((1 - \pi) + \pi e^t) \right|_{t=0} \quad (\text{S.213})$$

$$= \left. \frac{d}{dt} (\pi e^t) \right|_{t=0} \quad (\text{S.214})$$

$$= \pi e^t \Big|_{t=0} \quad (\text{S.215})$$

$$= \pi \quad (\text{S.216})$$

This is the second raw moment. Recall that $\mathbb{E}[X^2] = \sigma^2 + \mu^2$. Thus, we can calculate the variance of X as

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mu^2 \quad (\text{S.217})$$

$$= \pi - \pi^2 \quad (\text{S.218})$$

$$= \pi(1 - \pi) \quad (\text{S.219})$$

Thus, we have obtained our usual variance formula for a Bernoulli random variable.



Example 15

Next, let us determine the moment generating function for a Binomial random variable with parameters n and π .

Solution: By our definition of the Binomial random variable, we know that it is just the sum of n independent Bernoulli random variables. Let us use that to make the calculations easier.

Thus, let X_i be a series of Bernoulli random variables and $Y := \sum_{i=1}^n X_i$ be the resulting Binomial random variable. The moment generating function of Y is

$$M_y(t) = \mathbb{E}[\exp[Yt]] \quad (\text{S.220})$$

$$= \mathbb{E}\left[\exp\left[\sum_{i=1}^n X_i t\right]\right] \quad (\text{S.221})$$

$$= \mathbb{E}\left[\prod_{i=1}^n \exp[X_i t]\right] \quad (\text{S.222})$$

Since the X_i are independent and identically distributed, this is

$$= \prod_{i=1}^n \mathbb{E}[\exp[X_i t]] \quad (\text{S.223})$$

Note that $\mathbb{E}[\exp[X_i t]]$ is just the MGF for a Bernoulli random variable. Thus, this becomes

$$= \prod_{i=1}^n ((1 - \pi) + \pi e^t) \quad (\text{S.224})$$

and the MGF for a Binomial random variable is

$$M_y(t) = ((1 - \pi) + \pi e^t)^n \quad (\text{S.225})$$

With that, we can calculate the various moments, should we desire. ♦

Lemma S.35. Let X and Y be two independent random variables. Define $Z := X + Y$. The moment generating function of Z is the product of the moment generating functions of X and Y :

$$M_Z(t) = M_X(t) M_Y(t) \quad (\text{S.226})$$

Proof. The proof is simply an exercise in definitions and algebra:

$$M_Z(t) := \mathbb{E} \left[e^{tZ} \right] \quad (\text{S.227})$$

$$= \mathbb{E} \left[e^{(X+Y)t} \right] \quad (\text{S.228})$$

$$= \mathbb{E} \left[e^X e^Y \right] \quad (\text{S.229})$$

$$= \mathbb{E} \left[e^X \right] \mathbb{E} \left[e^Y \right] \quad (\text{S.230})$$

$$= M_X(t) M_Y(t) \quad (\text{S.231})$$

Note that we needed the assumption of independence in the fourth line (S.230), where we split the expected value. \square

Lemma S.36. The moment generating function for the Normal distribution is

$$M_X(t) = \exp \left[\mu t + \frac{1}{2} \sigma^2 t^2 \right] \quad (\text{S.232})$$

I leave this proof as an exercise. It is a simple application of the definition of moment generating function (Definition S.34) and the probability density function of the Normal (Equation S.122).

With that last theorem, we are able to prove that a linear combination of independent Normal random variables also is a Normal random variable.

Lemma S.37. Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. Let X and Y be independent; that is, $X \perp Y$. If a , b , and c be scalars, then

$$aX + bY + c \sim \mathcal{N}(a\mu_x + b\mu_y + c, a^2\sigma_x^2 + b^2\sigma_y^2) \quad (\text{S.233})$$

Proof. As expected, the proof relies on moment generating functions. First, the moment generating function of $aX + bY + c$ is the product of the MGFs of aX , bY , and c because they are independent. So, let us examine the MGFs of each.

$$M_{ax}(t) = M_x(at) \quad (\text{S.234})$$

$$= \exp\left[\mu_x(at) + \frac{1}{2}\sigma_x^2(at)^2\right] \quad (\text{S.235})$$

$$M_{by}(t) = M_y(bt) \quad (\text{S.236})$$

$$= \exp\left[\mu_y(bt) + \frac{1}{2}\sigma_y^2(bt)^2\right] \quad (\text{S.237})$$

$$M_c(t) = M(ct) \quad (\text{S.238})$$

$$= \exp[ct] \quad (\text{S.239})$$

Putting these together gives

$$M_{ax+by+c}(t) = \exp\left[\mu_x(at) + \frac{1}{2}\sigma_x^2(at)^2\right] \cdot \exp\left[\mu_y(bt) + \frac{1}{2}\sigma_y^2(bt)^2\right] \cdot \exp[ct] \quad (\text{S.240})$$

$$= \exp\left[\mu_x(at) + \frac{1}{2}\sigma_x^2(at)^2 + \mu_y(bt) + \frac{1}{2}\sigma_y^2(bt)^2 + ct\right] \quad (\text{S.241})$$

$$= \exp\left[(a\mu_x + b\mu_y + c)t + \frac{1}{2}(a^2\sigma_x^2 + b^2\sigma_y^2)t^2\right] \quad (\text{S.242})$$

Note that this is the moment generating function of a Normal distribution with mean $a\mu_x + b\mu_y + c$ and variance $a^2\sigma_x^2 + b^2\sigma_y^2$. Thus, we have proven this theorem. \square

degenerate

$$\mathbb{P}[X = c] = 1$$

Note: We are not quite done; there is one missing piece. How can we have a moment generating function for the scalar value c ? Aren't MGFs only for random variables? Well, there is a distribution called the "degenerate distribution" that takes on a single value with probability 1. Think of a two-headed coin. The probability mass function for a degenerate distribution is $\delta(X - c)$, where $\delta(\cdot)$ is the Dirac delta function (Calculus I), X is the random variable, and c is the only possible value of x .

If you have taken a course in differential equations, you will certainly have come across the Dirac delta function. If not, then you may not have. In my experience, I saw $\delta(\cdot)$ in Calculus I.

The take-away is that there is a lot of statistics that you have not even seen. You have just caught little glimpses. Sometimes, you have seen probability without actually noticing it. For instance, the probability density functions you have seen are just solutions to certain differential equations.

Astounding!

Note: Alright, I'm not done with that. If you do not like thinking about the degenerate distribution as an actual distribution, think of it as a limiting distribution. For instance, define

$$f_s(x; \mu) := \frac{1}{\sqrt{2\pi s^2}} \exp\left[-\frac{1}{2} \frac{(x-\mu)^2}{s^2}\right] \quad (\text{S.243})$$

The limit of this (Normal) distribution as $s \rightarrow 0$ is the degenerate distribution at the point $x = 0$.

Neat-o skeet-o!

Note that the previous distribution showed that the sum of two Normally-distributed random variables also has a Normal distribution. The converse, that a Normally-distributed random variable can only be decomposed into the sum of two other Normally-distributed random variables is much more difficult to prove. It is the purpose of Cramér's Theorem.

In other words, Cramér's Theorem states that if $Z = X + Y$ and if Z follows a Normal distribution, then both X and Y must also follow Normal distributions. The proof can be found in Cramér (1936).

S.6.6 PROOF OF THE CENTRAL LIMIT THEOREM* And so, with all of this preparation, we are now ready to prove the Central Limit Theorem.

Theorem S.6.5: The Central Limit Theorem

Let X be a random variable with expected value μ and finite variance σ^2 . Let $\{X_1, X_2, \dots, X_n\}$ be a random sample of size n from this distribution. Finally, define T_n as

$$T_n := \sum_{i=1}^n X_i \quad (\text{S.244})$$

Then,

$$T_n \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2) \quad (\text{S.245})$$

as $n \rightarrow \infty$.

Here is an **overview of this proof**: This proof of the Central Limit Theorem relies on moment generating functions and Taylor's theorem. It starts with the distribution of T , which we need to determine. It then creates two subsequent distributions, Y_n and Z_i , to help with the calculations. The moment generating function for Z_i is approximated using Taylor's theorem. From that, the moment generating function for Y_n is determined. That provides us the approximate distribution of Y_n , which is $\mathcal{N}(0, 1)$. With that, we determine the approximate distribution of T , as required.

With that overview, let us start the actual proof.

Proof. Let X be a random variable with mean μ and finite variance σ^2 . Let $\{X_1, X_2, \dots, X_n\}$ be a random sample from X . The sum $T_n := \sum X_i$ has mean $n\mu$ and variance $n\sigma^2$.

Define the random variable Y_n as

$$Y_n := \frac{T_n - n\mu}{\sqrt{n\sigma^2}} \quad (\text{S.246})$$

In other words, the Y_n random variable is just the T_n random variable standardized.

Note that this definition is equivalent to

$$Y_n = \sum_{i=1}^n \frac{X_i - \mu}{\sqrt{n\sigma^2}} \quad (\text{S.247})$$

Eventually, we will show that the distribution of Y_n converges to standard Normal as $n \rightarrow \infty$. For now, we use it as an intermediate between T and Z_i . Define Z_i as

$$Z_i := \frac{X_i - \mu}{\sigma} \quad (\text{S.248})$$

which is the z-score for the X_i random variable. It is useful, because we can now see

$$Y_n = \sum_{i=1}^n \frac{1}{\sqrt{n}} Z_i \quad (\text{S.249})$$

This link between Z_i and Y_n will be exploited later, when we finally approximate the distribution of Z_i and use it to approximate the distribution of Y_n .

With that literary foreshadowing, note that the moment generating function of Y_n , in terms of the Z_i , is

$$M_y(t) = \left(M_z \left(\frac{t}{\sqrt{n}} \right) \right)^n \quad (\text{S.250})$$

This result relies on the Z_i being independent and identically distributed. The moment generating function for Z_1 (or any of the Z_i) can be approximated by using Taylor's theorem (expanded around $t = 0$):

Taylor

$$M_z \left(\frac{t}{\sqrt{n}} \right) = \sum_{i=0}^{\infty} \left(\frac{t}{\sqrt{n}} \right)^i \frac{M_z^{(i)}(0)}{i!} \quad (\text{S.251})$$

$$= \left(\frac{t}{\sqrt{n}} \right)^0 \frac{M_z^{(0)}(0)}{0!} + \left(\frac{t}{\sqrt{n}} \right)^1 \frac{M_z^{(1)}(0)}{1!} + \left(\frac{t}{\sqrt{n}} \right)^2 \frac{M_z^{(2)}(0)}{2!} + \dots \quad (\text{S.252})$$

$$= 1 \frac{1}{0!} + \frac{t}{\sqrt{n}} \frac{\mathbb{E}[Z_i]}{1!} + \left(\frac{t}{\sqrt{n}} \right)^2 \frac{\mathbb{E}[Z^2]}{2!} + \dots \quad (\text{S.253})$$

$$= 1 + \frac{t}{\sqrt{n}} \frac{0}{1!} + \left(\frac{t}{\sqrt{n}} \right)^2 \frac{1}{2!} + \dots \quad (\text{S.254})$$

$$= 1 + 0 + \frac{t^2}{2n} + \dots \quad (\text{S.255})$$

If we include the remainder term from Taylor (the \dots above), then we have this conclusion

$$M_z \left(\frac{t}{\sqrt{n}} \right) \approx 1 + \frac{t^2}{2n} + \xi \frac{t^3}{6n^{3/2}} + o \left(\frac{t^3}{n^{3/2}} \right) \quad (\text{S.256})$$

The last two terms are the remainder. Here, ξ is a constant and $o \left(\frac{t^3}{n^{3/2}} \right)$ is "little-o" notation indicating that the remainder (whatever it is) goes to 0 faster than $\frac{t^3}{n^{3/2}}$ (as $t \rightarrow 0$).

Now, we return to Y_n . It was defined as $\sum_{i=1}^n \frac{1}{\sqrt{n}} Z_i$, where the Z_i were independent and identically distributed. This means the moment generating function of Y_n is

$$M_y \left(\frac{t}{\sqrt{n}} \right) \approx \left(1 - \frac{t^2}{2n} + \xi \frac{t^3}{6n^{3/2}} + o \left(\frac{t^3}{n^{3/2}} \right) \right)^n \quad (\text{S.257})$$

The third term goes to 0 as $n \rightarrow \infty$, as does the fourth term. The remaining two terms are

$$M_y\left(\frac{t}{\sqrt{n}}\right) \approx \left(1 - \frac{t^2/2}{n}\right)^n; \quad n \rightarrow \infty \quad (\text{S.258})$$

Note that this is just the definition of the exponential function $e^{\frac{1}{2}t^2}$. This means we have the MGF for Y_n :

$$M_y\left(\frac{t}{\sqrt{n}}\right) \approx e^{\frac{1}{2}t^2} \quad (\text{S.259})$$

... and this is just the moment generating function for a standard Normal distribution, $\mathcal{N}(0, 1)$. Thus, Y_n will approach $\mathcal{N}(0, 1)$ as $n \rightarrow \infty$.

Finally, using our definition of Y_n (Eqn S.247), we can conclude that the distribution of T_n converges in distribution to $\mathcal{N}(\mu, \sigma^2)$ as $n \rightarrow \infty$. \square



And so, in this final section of the appendix, we were finally able to prove the Central Limit Theorem. To do so, we needed to learn about moment generating functions.

Again, the only purpose of moment generating functions for this course of study is to be able to prove the Central Limit Theorem. I included the proof of the Central Limit Theorem simply because many have asked for it in the past.

Et voilà!

S.7: End-of-Appendix Materials

S.7.1 R FUNCTIONS In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

PACKAGES:

RFS This is a “book package,” that is not yet complete. In lieu of installing this package and loading it with `library(RFS)`, you will activate all of its important parts by running
`source("http://rfs.kvasaheim.com/rfs.R")`.

STATISTICS:

runs.test(r, order=x) The runs test tests if the residuals **r**, as ordered by **x**, are sufficiently distributed around the zero line. This is the version in the **RFS** package.

S.7.2 EXERCISES

1. Prove Lemma S.4.
2. Prove Lemma S.9.
3. Prove Lemma S.10.
4. Prove Lemma S.14.
5. Prove Lemma S.15.
6. Prove Lemma S.13.
7. Prove that the Cauchy distribution is equivalent to the Student's t distribution with 1 degree of freedom.
8. Prove that the Normal distribution is equivalent to the Student's t distribution as $\nu \rightarrow \infty$.
9. Prove that the interquartile range of the standard Cauchy is 2.
10. Prove $\mathbb{E}[S^2] = \sigma^2$ when $Y \sim \mathcal{N}(\mu, \sigma^2)$.
11. Prove that the first central moment is always 0, as long as $\mathbb{E}[X]$ exists.
12. Prove that the third central moment (skew) for the Normal distribution is zero.
13. Use moment generating functions to determine when the Binomial distribution has zero skew.
14. Calculate the moment generating function for the Poisson distribution. Check that it generates the first two moments. Use it to determine the variance of a Poisson random variable.
15. Prove Theorem S.36.
16. Use the moment generating function of the Bernoulli distribution to determine the moment generating function of the degenerate distribution.
17. Use moment generating functions to calculate the variance of a degenerate distribution (see page 598).

S.7.3 THEORY READINGS

- Carl E. Bonferroni (1936). “Teoria statistica delle classi e calcolo delle probabilità.” *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*. 8: 3–62.
- James V. Bradley (1968). *Distribution-Free Statistical Tests*, Chapter 12. Prentice-Hall.
- Harald Cramér (1936). “Über eine Eigenschaft der Normalen Verteilungsfunktion.” *Mathematische Zeitschrift* (in German). 41(1): 405–414.
doi:10.1007/BF01180430.
- Olive Jean Dunn (1958). “Estimation of the Means for Dependent Variables.” *Annals of Mathematical Statistics*. 29(4): 1095–1111.
doi:10.1214/aoms/1177706374.
- Olive Jean Dunn (1961). “Multiple Comparisons Among Means.” *Journal of the American Statistical Association*. 56(293): 52–64.
doi:10.1080/01621459.1961.10482090.
- Alexander M. Mood (1940). “The Distribution Theory of Runs.” *Annals of Mathematical Statistics*. 11(4): 367–392.
- Abraham Wald and Jacob Wolfowitz (1940). “On a test whether two samples are from the same population.” *Annals of Mathematical Statistics*. 11(2): 147–162.

Index

Applications

- ANCOVA, 158
- approval, 456
- binomial regression, 381, 392
- cattle feed, 458
- coins, 347
- crime, 153
- differential invalidation, 392
- geography, 243
- GLM, 319, 322
- heteroskedasticity, 246
- imputation, 463
- lifetime, 291
- logistic regression, 381, 392
- median regression, 267
- nominal regression, 447
- occupation, 447
- ordinal regression, 456, 458, 463
- physics, 246
- Poisson regression, 426
- property crime, 267
- South Sudan, 209
- terrorism, 426
- time series, 241
- voting, 166, 195, 209, 250, 319
- wealth, 200, 322

R data

- cattleData, 458
- coinflips, 348
- cows, 166, 167, 195, 319
- crime, 153, 266, 418
- fakepoisson, 410

- gdpcap, 322
- gdp, 200
- gssocc, 447, 456
- ocanada, 381
- rur2013parl, 251
- sri2010pres, 392
- summary.aov, 175
- suvr, 463
- terrorism, 426
- xsd2011referendum, 209

R functions

- (drop), 211
- AIC, 364
- BIC, 365
- I, 430
- ROC, 357
- &&, 133, 134
- accuracy, 352, 355
- anova, 388
- aov, 160, 388
- attach, 153
 - no, 210
- autocor.test, 241
- barplot, 587
- binom.test, 118
- bptest, 131, 132, 154
- cbind, 383
- col, 242
- confint, 133, 134, 157, 165, 247, 248
- cor.test, 168, 458
- c, 140, 247, 248

`data.frame`, **178**, 197, 201, 323, 418, 460, 464
`exp`, 410
`fligner.test`, 161
`for-loop`, 117, 119, 133, 134, 154, 587
`function`, 587
`glm.nb`, 424, 428
`glm.nb,offset`, 428
`glm`, 319, **348**, 383, 418, 428
`glm,family`, 320, 322
`glm,link`, 320, 322
`glm,offset`, 428
`glm,quasipoisson`, 422
`head`, 112
`hetero.test`, 160
`hist`, 110, 121
`ks.test`, 118
`legend`, 461
`length`, 122
`library`, 125, 266
`lines`, 213
`lm`, 116, 117, 121, 131, 136, 140, 153, 162, 201, 211, 231
`lm,weights`, 248, 252
`logistic`, 195, 197
`logit.inv`, 195
`logit`, 195
`make.link`, 320, 322, 323, 344
`matrix`, 355
`mean`, 136, 184
`multinom`, 450, 458
`ordered`, **463**
`overlay`, 110, 153, 159
`plot`, 122, 131, 154
`points`, 157
`polr`, 456, 463
`predict`, 136, 157, 165, 178, 197, 201, 212, 323, 349, 418, 450, 457, 460, 464
`predict,se.fit`, 212
`qchisq`, 420
`qqline`, 108
`qqnorm`, 108
`quantile`, 203
`rcauchy`, 120
`rchisq`, 120
`read.csv`, **153**, 167, 266, 458
`residuals`, 122, 153, 159, 241, 388
`rexp`, 109, 119
`rnorm`, 108, 116, 117, 121, 122, 125, 131, 133, 134, 587
`row`, 242
`rq`, 265, 266, **266**, 267, 269, 270
`rt`, 114
`runif`, 125
`runs.test`, 125, 154, 159, 388, 590
`sample`, 154, 355
`seq`, 122, 131, 133, 134, 140
`set.base`, 162, 452
`set.seed`, 112, 117, 122, 125, 136, 140, 410
`shapiroTest`, 111, 113, 115, 153, 159
`sort`, 410
`source`, **107**, 125, 153, 167
`subset`, 418, 424
`summary.aov`, 158, 161
`summaryHCE`, 217

summary, 116, 117, 121, 153, 160, 319
 tail, 113
 vif, 141, 176
 which, 211
 R packages
 Epi, 357
 MASS, 424, 456, 463
 RFS, 111, 125, 195, 217, 352, 452, 590, **603**
 car, 141, 176
 lawstat, 125, 153
 lmtest, 131, 153
 nnet, 450, 458
 quantreg, 265
 accuracy, 351, 355, 452, 457
 AUC, 356
 maximum, 353
 rate, 351
 relative, 352
 additive model, 140, 159, 176
 AIC, 320, 321, **364**, 387, 412, 452, 464
 alpha-testing, 116, 119, 120
 ANCOVA, 158
 ANOVA, 64, 160, 574
 effects model, 69, 70
 means model, 68
 table, 158
 toy example, 28
 ANOVA table, 173
 appropriate test, **117**
 AR(1), 242
 AUC, 356
 autocorrelation, 241, 242
 back-transform, 197, 206, 349
 base category, 162, 452
 bias, 34, 51, 81, 125, 127, 299, 544
 bias-variance trade-off, 426, **433**
 BIC, **365**, 387, 412
 bijection, 295
 Binomial test, 118
 boolean, 133
 bounds, **192**
 one, 200
 two, 192, **193**
 others, **206**
 Breusch-Pagan test, **129**, 154, 161
 canonical link, 312, 314, 317, 342, 416
 CDF, 344, 554
 center of gravity, 31
 centering, 53
 Central Limit Theorem, *see* CLT
 classical linear model, *see* CLM
 CLM, 152, 190, 232, 240, 294, 308, 309
 CLT, 115, 120, 561, **591**, 600
 coefficient of variation, 167
 conditional distribution, **311**, 341, 379, 382, 416
 confidence band, 100, 212, 355
 confidence interval, 86, 94, 95, 99, 135, 157, 165, 266, 344
 confusion matrix, 355
 correlation, 168, 169, 543, 555
 population, 555
 sample, 543
 symmetric, 543
 covariance, 52, 82, 83, 503, 504, 541, 555
 population, 555

- sample, 541
 - symmetric, 542
- covariance matrix, 309
- coverage, **133**, 136
- cumulative distribution
 - Cauchy, 573
- cumulative distribution function, *see*
 - CDF
- data matrix, *see* design matrix
- degrees of freedom, 496
- design matrix, 54, 107
- deviance, 538
- differential invalidation, **209**, 250, 392
- directional hypotheses, 166
- electoral forensics, 209, 250, 392
- estimator, **48**
- exponential class, **313**, 379
- exposure, *see* offset
- extrapolation, 27
- factor analysis, 143
- Fligner-Killeen test, 161
- function
 - absolute value, 262
 - beta, 575
 - exponential, 202
 - gamma, 569, 572, 575
 - identity, 422
 - logarithm, 200, 413
 - logistic, **194**, 197, 337
 - logit, **193**, **314**, 337
- functional form, 122, 128, 154, 159, 315
- Gaussian process, *see* normality
- generalized least squares, *see* GLS
- generalized linear model, *see* GLM
- geographic regression, 243
- GLM, **309**, 310
 - vs* CLM, 309
 - assumptions, 315
 - components, 310, 332, 339, 341, 348, 413
 - distribution, **311**
 - exponential class, **313**
 - linear predictor, **310**
 - link function, **312**, 318
- GLS, **240**
- grammar of formulas, **171**
- graphics, 155, 163, **179**, 182, **460**, 465
 - R philosophy, 212
- hat matrix, **57**, **237**
- heteroskedasticity, 133, 217, 309, 335, 409
 - adjustment, 216
 - bulge, **134**
 - effects, 125, 133, 135
 - funnel, **134**
 - modeling, 230
 - none, *see* homoskedasticity
 - trumpet, **133**, 136, 414
- heteroskedasticity test, 160, 161
- Hildebrand, **557**
- Hildebrand Rule, 155, 447, 580
- homoskedasticity, 78, **129**, 154, 160, 229, 340
- Huber-White, 216
- iid, 78, 80, 229, 566
- independent, 80, 230, 295
- independent and identically distributed, *see* iid

influential point, 267, 419
 inner product, 483
 interaction model, 158, 171, 173
 interaction term, 431
 interpolation, 27, 155
 interpretation, 135, 182, 209, 254, 320, 452, 461
 of logarithm, 202
 of logit, 196

 Kingdom of Ruritania, *see* Ruritania
 Kolmogorov-Smirnov test, 118, 119–121
 kurtosis, 560, 563, 570, 592

 latent variable, 336
 likelihood, 285, 287, 291
 log-likelihood, 289, 291
 likelihood ratio test, 365, 387
 line of “best” fit, 18, 282
 linear, 78
 linear predictor, 310, 337, 341, 350, 379, 381
 link function, 312, 320, 341, 359, 380, 382, 416, 446
 canonical, 312, 342, 379, 416
 cauchit, 344
 complementary log-log, 344, 359
 identity, 317, 422
 log-log, 344, 361
 logarithm, 415
 logit, 314, 337, 341, 344, 347, 446, 450, 457
 probit, 344
 logistic regression, 332

 machine epsilon, 138
 matrix, 476

 addition, 479
 additive identity, 479
 additive inverse, 479
 adjacency, 244
 associative, 480, 481, 492
 commensurate, 479, 482
 commutative, 480, 481, 489
 confusion, 355
 covariance, 229, 240, 309, 501, 504
 determinant, 484
 diagonalization, 494
 dimension, 476
 distributive, 481
 eigenvalue, 495
 eigenvector, 495
 hat, 237
 idempotent, 57, 496
 inner product, 483
 inverse, 67, 73, 484
 invertible, 484
 multiplication, 482
 multiplicative identity, 484
 multiplicative inverse, 484, 499
 orthogonal, 57, 62, 497
 positive definite, 240, 497, 500
 projection, 58, 497
 rank, 67, 484, 496, 501
 representation, 477
 sample covariance, 504
 sample mean, 502
 sample variance, 503
 singular, 138, 484
 square, 476
 symmetric, 57, 244
 symmetrize, 493

trace, 493
 transpose, 73, 493, 499
 maximum likelihood estimation, *see* MLE
 mean, 313, 413, 502, 537
 distribution of, 576
 linear, 537
 population, 545
 sample, **537**
 mean squared error, *see* MSE
 median regression, 18, **266**, 267
 MGF, 592–597
 minimum-width interval, 95
 vs central, 95, 96, **97**
 MLE, 18, 285, 287, 288, 294, 312, 446
 bias, 292
 function of, 292
 of β_0 , 294
 of β_1 , 297
 of σ^2 , 299
 uniqueness, 292
 model selection, 171, 174, 195, **362**, 429
 AIC, **364**
 BIC, **365**
 likelihood ratio test, **365**
 model specification, 128
 model stability, 446
 moment, 592
 central, 592
 raw, 592
 standardized, 592
 moment generating function, *see* MGF
 Monte Carlo, 181, 197, 203, 355
 MQLE, 420, **422**, 429
 MSE, 36, 84, 95, 238
 multicollinearity, 49, 64, **138**
 approximate, 138
 CS, 138
 fix, **142**
 indications, 140
 logic, 139
 super, 138, 139
 test, 141
 multiple comparisons, 581
 Bonferroni adjustment, **582**
 nominal regression, 449
 norm, 265
 L_1 , 265
 L_2 , 265
 Normality, 108
 normality, 78, 81, 153, 159
 nuisance parameter, **314**
 null model, 38, 452, 464
 Occam's Razor, 158
 odds ratio, 196, 199
 offset variable, **427**
 OLS, 19, 190, 229
 assumptions, 34, **50**, 77, 152, 176, 190, 229
 b_0 , 22, 82, 127
 b_1 , 22, 24, 32, 33, 35, 82, 125
 derivation, matrix, 46–48
 derivation, scalar, 20–22
 matrix, 229
 matrix model, **46**
 scalar, 229
 scalar model, **20**
 toy example, 25, 53
 OLS assumptions, 122

ordinary least squares, *see* OLS
 overdispersion, 384, 415, **420**, 428
 adjustment, 421
 causes, 420
 effects, 420
 test, 420

 p-value, 117
 parameter, 536, **545**, 556
 point estimate, 183
 Poisson regression, 410
 PRE, 386
 definition, **38**, **63**, 352
 R^2 , 38, 141, 170
 \bar{R}^2 , 39
 pseudo- R^2 , 40, 63, 319, 351
 preamble, 167
 precision, 88
 prediction interval, 89, 99, **99**, 157, 165
 principal component analysis, 143
 probability distribution
 Bernoulli, 339, **556**, 594
 Binomial, **561**, 596
 binomial, 283, 284, 377
 bivariate normal, 301
 Cauchy, 120, 573
 Chi-square (χ^2), 536, 569, 578
 chi-square (χ^2), 84, 91, 95, 120
 degenerate, 598
 exponential, 119, 290
 F, 574
 gamma, 423
 Gaussian, 316, 566
 multinomial, 442
 multivariate, **300**
 multivariate normal, 301
 Negative Binomial, 413, 423
 Normal, 566, 576, 597
 Poisson, 286, 288, **413**, 423, **563**
 Snedecor's F, *see* F
 Student's t, *see* t
 t, 91, 94, 95, 114, 571
 uniform, 116, 583
 proportional reduction in error, *see* PRE

 Q-Q plot, 108
 QMLE, 312
 quantile function, 344
 quantile regression, 18, 269
 quasi maximum likelihood estimation,
 see MQLE
 quasi-likelihood, 385, 420

 random variable, 181, **534**
 regression table, 154, 162, 179
 rejection rate, 121, 135
 representative sample, 447
 residuals, 122, 240
 residuals plot, 122, 124, 160
 ROC curve, 355
 Rule of Thumb, 142, 170, 453, 580
 runs test, 123, 159, **586**
 Ruritania, 2
 currency, 3
 drugs, 5
 economics, 5
 geography, 3
 government, 3
 kraj, 4, 243
 Rudolph II, 3
 stát, 4

Státní Univerzita v Ruritánii, 463
 US embassy, 6
 US relations, 6
Valné Shromáždění, 287
Vlajka (1954), 3

sample space, **556**
 sensitivity, 355
 Shapiro-Wilk test, 113, 153, 159
 skew, 109, **557**, 563, 570, 580, 592
 specificity, 355
 standard form, 313, 316, 417
 statistics experiment, 108, 116, 119,
 120, 133, 135, 140, 203, 204,
 355, 410
 S_{xx} , 24, 82, 88
 systematic error, 122

test statistic, **91**, 94, 135
 testing systematic error, 122
 threshold, 337, 353
 time series, 242
 Type I error, **116**, 121, 135, 357, 581
 Type II error, 116, 132, 135, 357

variability, 167
 variable type
 binary, 332
 count, 409
 dichotomous, 332
 nominal, **442**
 ordinal, **454**
 variance, 314, 503, 504, **535**, 538, 552
 distribution of, 578
 population, 552
 quadratic, 539
 sample, 538
 variance inflation factor, *see* VIF
 VGLM, 398
 VIF, 141, 170, 176
 voting, 236

weighted least squares, *see* WLS
 WLS, **230**
 estimator, 235
 matrix, 230
 Working-Hotelling, 100
 bands, 213