Linear Models and Řuritȁ Krȁlovstvì

Using the Kingdom for Greater Insight

Version: 0.704442ζ

Ole J. Forsberg

Department of Mathematics
Knox College
Galesburg, IL, USA

December 2023

This document was typeset using the LaTeX 2_ε 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
2	King Harald V of Norway	Government of Norway	CC0
3	Ruritanian Flag	Ole J. Forsberg	BY-NC-SA
??	Ambassador Georgette Mosbacher	US Embassy to Poland	CC0
11	King Olav V of Norway	Government of Norway	CC0
15	St. Canute's Cathedral, Odense, Denmark	Ole J. Forsberg	BY-NC-SA
57	Street View, Karlstad, Sweden	Ole J. Forsberg	BY-NC-SA
79	A Steadfast Tin Soldier statue, Odense, Denmark	Ole J. Forsberg	BY-NC-SA
159	A bridge in Karlstad, Sweden	Ole J. Forsberg	BY-NC-SA
195	A canal in The Hague, Netherlands	Ole J. Forsberg	BY-NC-SA
208	A map of San Marino	Wikipedia User Aotearoa	BY-SA
223	The Peace Palace in The Hague, Netherlands	Ole J. Forsberg	BY-NC-SA
241	Sunne Church in Sunne, Sweden	Ole J. Forsberg	BY-NC-SA
265	The Vigeland Park in Oslo, Norway	Ole J. Forsberg	BY-NC-SA
289	Bergen, Norway	Ole J. Forsberg	BY-NC-SA
329	Bergenhus Fortress in Bergen, Norway	Ole J. Forsberg	BY-NC-SA
357	Windmill scene in Malmö, Sweden	Ole J. Forsberg	BY-NC-SA
387	Street scene in Malmö, Sweden	Ole J. Forsberg	BY-NC-SA
421	Street scene in Karlstad, Sweden	Ole J. Forsberg	BY-NC-SA
443	Royal Guard in Oslo, Norway	Ole J. Forsberg	BY-NC-SA
465	William Sealy Gosset, 1908	Wikipedia	CC0
466	Ronald Fisher, 1913	Wikipedia	CC0
467	Augustin-Louis Cauchy, 1901	Wikipedia	CC0

Contents

List of Figures

List of Tables

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 five 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 statistical analysis. Scripting provides 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 benefit of being in academia 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 Denmark's second-city (Odense) 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

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

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

---

[1] Furthermore, 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.

[2] Be aware of the difference between Σ, $\sum$, and $\boldsymbol{\Sigma}$. 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: first, from context; second, because the sum will be performed over an index and symbolized as $\sum_i$ or as $\sum_{i=1}^{n}$.

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
o	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, its 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 **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.

Vyčkej Času
Adolph Heyduk

Nespěchej k štětí, tiše jen,
vše nemůž' najednou býti;
nejprv se travou zjeví len,
později 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`