

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>un</i> bounded 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.

mapping

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 $g(\mu) = \eta$.

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.

Pong

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.

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

$$\mathbb{V}[Y] = b''(\theta) \cdot a(\phi) \quad (11.5)$$

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

$$\text{logit}(\mu) := \log \left[\frac{\mu}{1 - \mu} \right] \quad (11.6)$$

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 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 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	$\ll 0.0001$
Banned Chickens	-0.0373	0.0200	-1.868	0.0723
Percent Religious	0.0095	0.0011	8.801	$\ll 0.0001$

Table 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

$$\mathbb{E}[\text{propWin}] = 0.1512 - 0.0201(\text{yearPassed}) - 0.0373(\text{chickens}) + 0.0095(\text{religPct}) \quad (11.30)$$

According to this model, what is the expected vote in Děčín? To answer this, we need this information about the Děčín ballot measure: `yearPassed = 9`, `chickens = 0`, `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).

impossible

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

	Estimate	Std. Error	t-value	p-value
Constant Term	-1.8909	0.2898	-6.53	$\ll 0.0001$
Year Passed (post 2000)	-0.0886	0.0157	-5.63	$\ll 0.0001$
Banned Chickens	-0.2318	0.0878	-2.64	0.0134
Percent Religious	0.0475	0.0047	10.06	$\ll 0.0001$

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 .

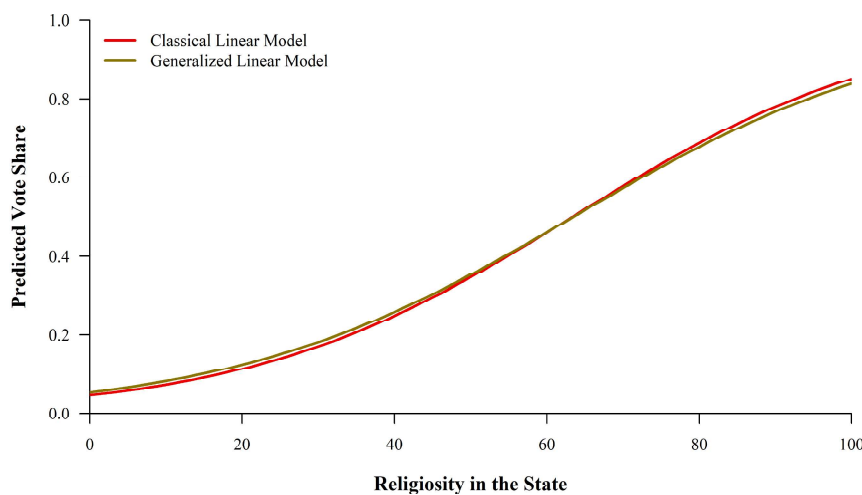


Figure 11.1: A plot of the predictions across various values of religiosity comparing the two models: CLM and GLM. Note that while the two results tables provided different results, the prediction plots are quite close together. The curves would have been equal only if we were to use the canonical link 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	$\ll 0.0001$
Level of Democracy	-0.0452	0.0061	-7.44	$\ll 0.0001$
Honesty in Government	0.3335	0.0219	15.20	$\ll 0.0001$

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).

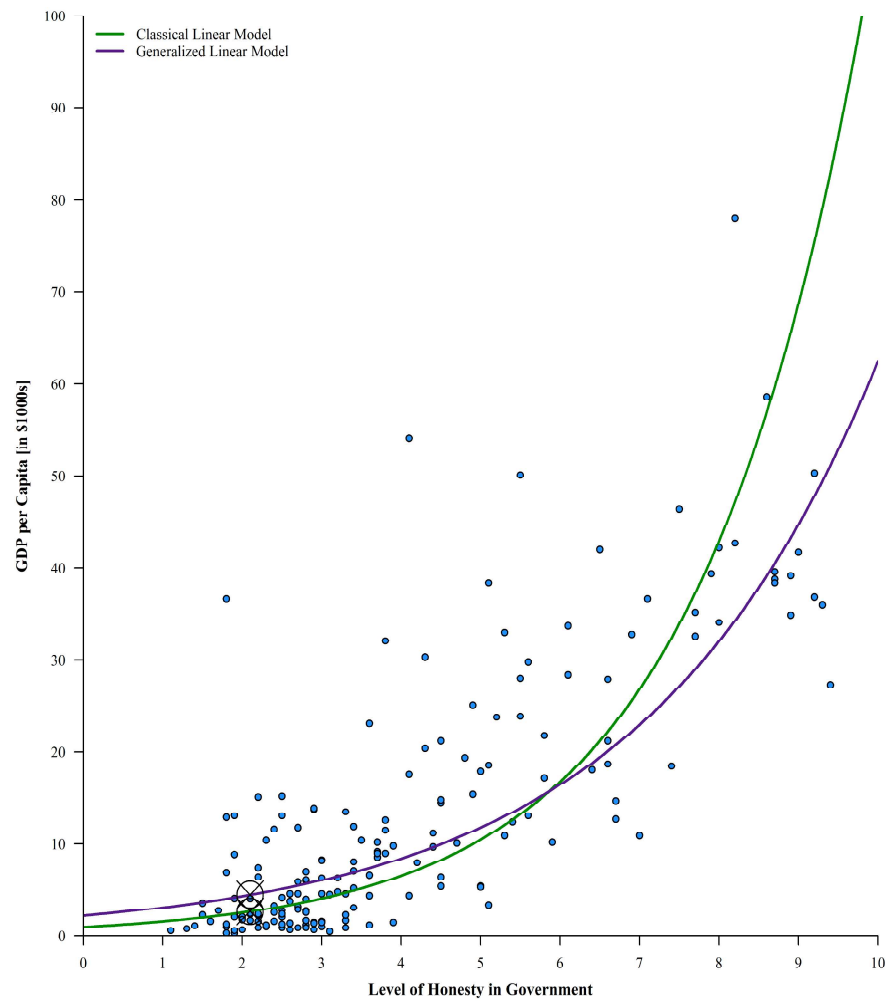


Figure 11.2: A plot of the two prediction curves, corresponding to the model fit using the Classical Linear Model and the Generalized Linear Model. Note that the two prediction curves are similar, but not really that close for large values of honesty in government. Estimates for Papua New Guinea are 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.

