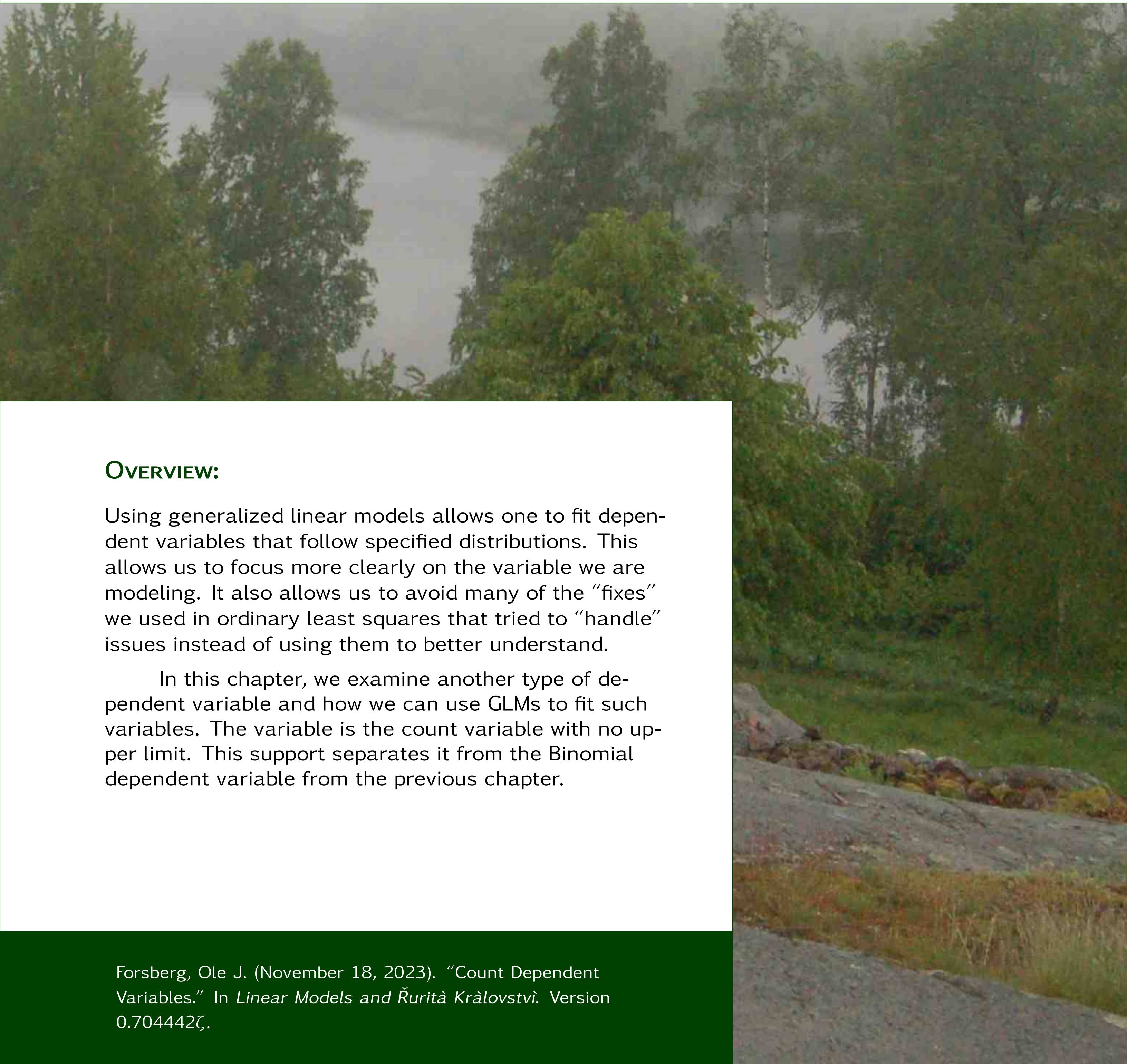
Chapter 13:

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.

Forsberg, Ole J. (November 18, 2023). "Count Dependent Variables." In *Linear Models and Řurità Kràlovstvì*. Version 0.704442ζ.

Chapter Contents

Remember that we are examining all of these different types of regressions for one primary reason: the assumptions of Ordinary Least Squares are not met by discrete dependent variables. Rather than seeing this as a problem, we can use it as an indicator that we can better model the data and extract more information from the data.

This marks the next chapter of discrete dependent variables. In Chapter 11, 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. 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 in Northern Ireland 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 of the United Kingdom — both Conservative and Labour — 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 Provisional Irish Republican Army (PIRA) was best able to organize its actions to affect the elections.

The question is whether they did — Did the PIRA adjust its tactics in reaction to the political ideology of the current prime minister? Unfortunately, the extant literature is divided on the direction of the effect. Some research suggests that the PIRA became more violent and killed more people when the Conservatives held power. Other research suggests that the PIRA became more violent under the Labour party. Which is it?

For the unbounded count variables in this chapter, there are three identifying characteristics: the variable can never be negative, has no theoretic upper bound, and is discrete. If Y is this type of count variable, then $Y \in \left\{0, 1, 2, 3, \ldots\right\}$.

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.

13.1: 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(577)

n=75
x1 = sort( runif(n, min=0, max=2) )
beta0 = 0
beta1 = 2
lambda = exp( beta0 + beta1*x1 )

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 `lambda` is defined and keep this in mind as you read forward.

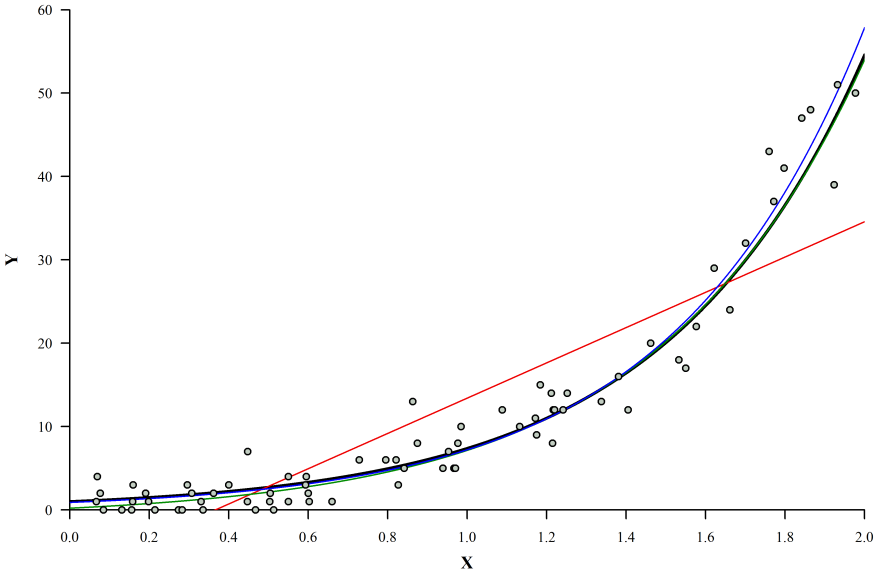


Figure 13.1: *Plot of the pseudo data with three regression equations overlaying. The linear regression is in red, the linear regression on the log-transformed data is in green, and the Poisson regression is in blue. The black curve is the "correct" curve.*

For this example, the true parameters are $\tilde{\beta}_0 = 0$ and $\tilde{\beta}_1 = 2$. Both of these are in log units (hence the tildes to serve as reminders of this). Except for those provided for the linear model, it is difficult to compare the estimates the the true value. It is much easier to compare the prediction curves.

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. 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. Check that the linear model fails the fit test (runs test provides a p-value of 3×10^{-9}). The assumption of homoskedasticity fails for Model 2 (where it is required; the Breusch-Pagan test returns a p-value of 2.8×10^{-5}). The Poisson model passes the runs test and does not require homoskedasticity. Thus, on the basis of meeting assumptions, the third model is the best of these three.

[1] Remember 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.

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 13.1). Numerically, we could also check how much the uncertainty in y has changed. The uncertainty using the null model (predicting $y = \overline{y}$) 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.

13.2: 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. First, we need to know the distribution of the dependent variable, conditioned on the independent variables. Second, we need to know the linear predictor, η. 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 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 Section 13.3.3.

13.2.1 The Poisson Distribution

The Poisson distribution has the following probability mass function (pmf):

$$f_Y(y;\ \lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad y \in \left\{0, 1, 2, 3, \dots\right\} \tag{13.1}$$

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*, $\mathbb{E}[Y \mid X]$, 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:

$$\mathbb{E}[Y] := \sum_{y=0}^{\infty} y\ f_Y(y) \tag{13.2}$$

$$= \sum_{y=0}^{\infty} y \frac{e^{-\lambda}\lambda^y}{y!} \tag{13.3}$$

$$= \sum_{y=1}^{\infty} y \frac{e^{-\lambda}\lambda^y}{y!} \tag{13.4}$$

$$= \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^y}{(y-1)!} \tag{13.5}$$

$$= \lambda \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^{(y-1)}}{(y-1)!} \tag{13.6}$$

Let us define $z := y - 1$:

$$= \lambda \sum_{z=0}^{\infty} \frac{e^{-\lambda}\lambda^z}{z!} \tag{13.7}$$

and so, we have

$$\mathbb{E}[Y] = \lambda \tag{13.8}$$

This last step is correct as $\frac{e^{-\lambda}\lambda^z}{z!}$ is the probability mass function for the Poisson, therefore $\sum_{z=0}^{\infty} \frac{e^{-\lambda}\lambda^z}{z!} = 1$. 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 an assumption 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 13.3), we will discuss what it means and what we should do.

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 6).

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,

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{13.9}$$

- the distribution of the dependent variable,

$$Y \mid x \sim \mathcal{P}(\lambda) \tag{13.10}$$

  with the formula for the expected value,

$$\mu = \lambda \tag{13.11}$$

- and the link function,

$$\log(\mu) = \eta \tag{13.12}$$

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.

13.2.2 Deriving the *Canonical* Link

In Chapter 10, 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 10), Bernoulli (Chapter 11), and binomial (Chapter 12):

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:

$$\begin{aligned} \text{pmf}: \quad f(y \mid \lambda) &= \frac{e^{-\lambda}\ \lambda^y}{y!} && (13.13) \\ &= \exp\left[\log\left(e^{-\lambda}\right) + \log\left(\lambda^y\right) - \log\left(y!\right)\right] && (13.14) \\ &= \exp\left[-\lambda + y\log(\lambda) - \log(y!)\right] && (13.15) \\ & && (13.16) \end{aligned}$$

$$f(y \mid \lambda) = \exp\left[\frac{y\ \log(\lambda) - \lambda}{1} + -\log(y!)\right] \tag{13.17}$$

This is in the required form:

(13.18)

$$\exp\left[\frac{y\ \theta - b(\theta)}{a(\phi)} + c(y, \theta)\right] \tag{13.19}$$

(13.20)

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 10.2.4.

> ***Note***: As mentioned in previous chapters (10, 11, and 12), we are not *required* to use the canonical link. *Any* monotonic, increasing function that maps the restricted domain to the unrestricted domain works. With that said, however, few links work better than the logarithm link in this situation.

Example 13.1: 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:

$$\texttt{inituse} \mid \lambda \sim \mathcal{P}(\lambda) \tag{13.21}$$

The linear predictor will use our explanatory variable:

$$\eta = \beta_0 + \beta_1 \texttt{pop90} \tag{13.22}$$

The link function will be the logarithm function:

$$\log(\lambda) = \eta \tag{13.23}$$

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 *in*consistent 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)

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

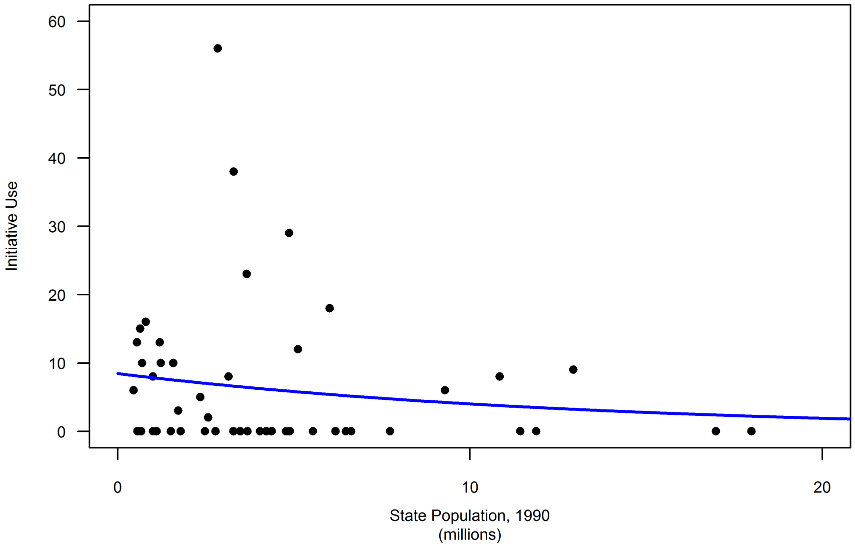


Figure 13.2: *A plot of initiative use against the population of the state in 1990, with the Poisson regression curve superimposed.*

> specify which data to include in the analysis. Here, I removed the state with ccode 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 13.2 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 is *overdispersed*.

13.3: 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 $a(\phi) = 1$. Overdispersion means that $a(\phi) > 1$.

In essence, this means $\mathbb{V}[Y \mid X] > \mathbb{E}[Y \mid X]$. 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 13.1), 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 12.3.2 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 to an acceptable level.

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 over-fitting the data; that is, you are fitting the data and not the data-generating process in which we are actually interested (see Section 13.5).

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

[2] The code I used was `qchisq( c(0.025,0.975), df=48)`.

	Parameter	Original	Adjusted		
	Estimate	Std. Error	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 13.1: *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.*

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. Quasi-maximum likelihood estimation (QMLE) 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.

13.3.1 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* by the square root of the dispersion factor.

Table 13.1 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.

13.3.2 Quasi Maximum Likelihood Estimation

The maximum likelihood estimation method makes assumptions about the relationship between the

	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 13.2: *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.*

mean and variance of the underlying distribution. For the Poisson distribution, that relationship is the identity function; that is, $\mathbb{E}[Y] = \mathbb{V}[Y]$. 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 Quasi Maximum Likelihood Estimation (Section 12.3.2, page 338). 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 QMLE 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. R can. To model using QMLE in R, we prefix the distribution with the world quasi. Thus, we would use

```
glm(y ~  x, family=quasipoisson(link=log))
```

to fit this model. This command produces the results in Table 13.2. 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 QMLE option in R are the Poisson (quasipoisson) and the binomial (quasibinomial). These are the only two main distributions that have a specific numeric value for $a(\phi) = 1$. The rest have a value for $a(\phi)$ that is estimated from the data; for instance, the Gaussian has $a(\phi) = \sigma^2$.

13.3.3 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

an

$$Y \mid \mu,\theta \ \sim NegBin(\mu,\theta) \tag{13.24}$$

where

$$Y \mid W \sim \mathcal{P}(\mu W) \tag{13.25}$$

with

$$W \sim \frac{1}{\theta}\text{GAM}(\theta) \tag{13.26}$$

where $\mathbb{E}[W] = 1$ and $\mathbb{V}[W] = 1/\theta$.

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

exercise

$$f(y;\theta) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)\ y!}\ \frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}} \tag{13.27}$$

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` — the programmers designed it that way. Thus, the command

	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 13.3: *The results table for modeling the initiative use using the Negative Binomial distribution. Note that the population is no longer statistically significant.*

```
m2n = glm.nb( inituse ~  pop90, data=vcr, subset=(ccode!=93) )
```

will perform negative binomial regression similar to the regression performed in Section 13.3.2. 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 13.3). The second thing to notice is that the effect of population is still no longer statistically significant. This agrees with our observation in Sections 13.3.1 and 13.3.2. 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]

exercise

This model estimates that Utah will have had approximately 7.9 initiatives during the 1990s. I leave it as an exercise to determine this.

[3] That 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.

13.4: 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 13.2: The Troubles in Northern Ireland 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 of the United Kingdom — both Conservative and Labour — 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 Provisional Irish Republican Army (PIRA) was best able to organize its actions to affect the elections.

The question is whether they did —

> Did the PIRA react to the political ideology of the current prime minister?

Unfortunately, the extant literature is divided on the direction of the effect. Some research suggests that the PIRA became more violent and killed more people when the Conservatives held power. Other research suggests that the PIRA became more violent under the Labour 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

$$\texttt{deaths} \sim \texttt{riteleft} \tag{13.28}$$

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. However, this may not really make sense. If the model tells us that the coefficient estimate for days is 2.35, what does that really mean?

It is usually 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?

mathemagic

Fear not! Through the magic of mathematics, we can handle it.

Recall in Section 13.2 that the link function we used was the logarithm: $\log[\lambda] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. If, instead of the expected count, λ, we wanted to model the expected ratio, $\frac{\lambda}{days}$, we would have:

$$\log\left[\frac{\lambda}{\texttt{days}}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{13.29}$$

Using one of the properties of logarithms, this is equal to

$$\log[\lambda] - \log[\texttt{days}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{13.30}$$

This, in turn, is mathematically equivalent to

$$\log[\lambda] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \log[\texttt{days}] \tag{13.31}$$

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 ($\log[days]$ on the right as a non-random variable). Note that *there is no parameter* to estimate for $\log[days]$. 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 was constant). Most statistical programs have an offset option available when you specify the model to be fit. In R, the offset is specified in the model call by the keyword ‘offset’.

offset

For the glm function,

```
glm(pira ~  riteleft, offset=log(days), data=terror)
```

For glm.nb,

```
glm.nb(pira ~  riteleft, offset(log(days)), data=terror)
```

	Poisson	Quasi-Poisson	Negative Binomial
Constant Term	2.2622	2.2622	2.0363
	(0.1190)	(0.7071)	(0.6004)
Conservatism	-0.0115	-0.0115	-0.0142
	(0.0011)	(0.0065)	(0.0093)
Days in year	0.0050	0.0050	0.0058
	(0.0004)	(0.0021)	(0.0019)
AIC	1482.8	——	369.5

Table 13.4: *Results of three different "families:" Poisson, quasiPoisson, and Negative Binomial. The numbers in parentheses, below the coefficient estimates, are the standard errors.*

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 13.4.

specific

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 quasiPoisson families. The only difference is the size of the standard errors. The quasiPoisson 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 quasiPoisson 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 at the more-appropriate results of fitting using Quasi Maximum Likelihood Estimation (or using the negative binomial family), we see

	Poisson	Quasi-Poisson	Negative Binomial
Constant Term	-1.8280	-1.8280	3.8744
	(0.0254)	(0.1495)	(0.0969)
Conservatism	-0.0106	-0.0106	-0.0069
	(0.0011)	(0.0063)	(0.0041)
AIC	1479.6	——	2080.2

Table 13.5: *Results of three different families: Poisson, quasiPoisson, and Negative Binomial. The numbers represent the estimated coefficients. The numbers in parentheses are the estimated standard errors.*

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.

exposure

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 13.5.

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 QMLE 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 PIRA terrorism in the United Kingdom during the Troubles in Northern Ireland ($p = 0.0905$).

13.4.1 Bettering the Fit*

Using the results from both the quasiPoisson 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]

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.

extreme

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

$$\begin{aligned}
\texttt{pira} = \beta_0 + \beta_1\ \texttt{riteleft} + \beta_2\ \texttt{riteleft}^2 + \beta_3\ \texttt{riteleft}^3 \qquad (13.32)\\
+ \beta_4\ \texttt{riteleft}^4 + \beta_5\ \texttt{labour} \qquad (13.33)\\
+ \beta_6\ \texttt{labour} \times \texttt{riteleft} \qquad (13.34)
\end{aligned}$$

Of course, we would need to have good theory to provide this model, but let's just have fun with this.

theory!!

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

$$\begin{aligned}
\texttt{pira} \sim \texttt{riteleft + labour + I(riteleft}\wedge\texttt{2)} \qquad (13.35)\\
\texttt{+ I(riteleft}\wedge\texttt{3) + I(riteleft}\wedge\texttt{4) + I(labour}\star\texttt{riteleft)}\\
(13.36)
\end{aligned}$$

The use of `I()` indicates that `R` should evaluate what is in the parentheses as a new variable. Fitting this model using Quasi Maximum Likelihood Estimation indicates that none of the terms have a statistically significant effect.

[4] With 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	−6.980
	(4.478)	(2.396)
Labour	−4.742	−4.8430
	(1.553)	(0.2856)
Conservatism	1.847	1.8660
	(0.07101)	(0.3778)
Conservatism^2	−0.03830	−0.03833
	(0.01425)	(0.00750)
Conservatism^3	−0.002585	−0.0026070
	(0.0009421)	(0.0005005)
Conservatism^4	−0.00007314	−0.00007361
	(0.00002642)	(0.00001398)
AIC	——	1787.8

Table 13.6: *Results of two different models: fitting with QMLE and using the Negative Binomial family. The numbers are the parameter estimates; in parentheses, the estimated standard errors.*

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 from the model is the interaction term. Doing this gives us the research model:

$$\begin{aligned} \texttt{pira} = \beta_0 + \beta_1\ \texttt{riteleft} + \beta_2\ \texttt{riteleft}^2 + \beta_3\ \texttt{riteleft}^3 \\ + \beta_4\ \texttt{riteleft}^4 + \beta_5\ \texttt{labour} + \varepsilon \end{aligned} \tag{13.37}$$

Fitting this model using both the quasiPoisson family and the Negative Binomial family gives us the results in Table 13.6.

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.

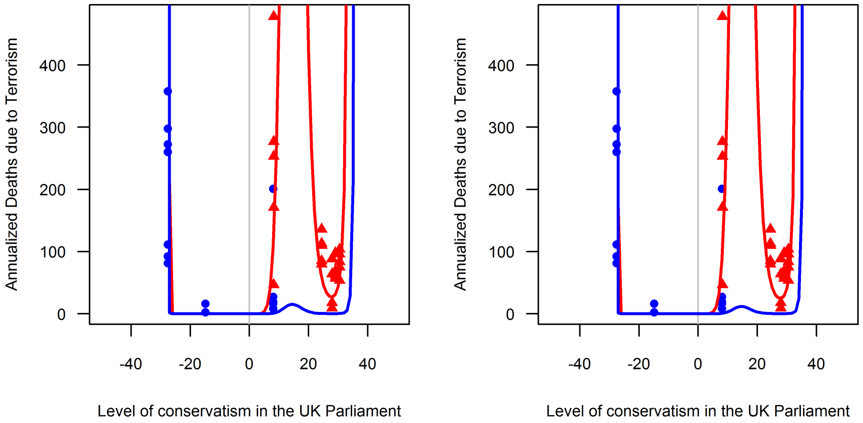


Figure 13.3: *Plot of the number of deaths due to terrorism, caused by the Provisional Irish Republican Army (PIRA), in the United Kingdom during the Troubles in Northern Ireland. The points are overlaid with the quasiPoisson model (Left Panel) and the Negative Binomial model (Right Panel). In both cases, the upper curve (red) corresponds to the prediction when the Conservative Party is in power.*

As such, we can show the effects on the same graph (one graph for each family), with two prediction curves per graph. Figure 13.3 shows the predictions from both the quasiPoisson 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.

13.5: 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

[5] Explanation 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 cam to being).

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 13.4), we can see that we used too many explanatory variables in our model. A glance at the graphs in Figure 13.3 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).

13.6: 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 6 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 quasiPoisson; 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.

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:

MASS This package is a "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.

13.7.2 Exercises

1. Show that $\mathbb{E}[Y] = \lambda$ and $\mathbb{V}[Y] = \lambda$ using the methods of Section 10.2.4.

2. Example 13.2.2 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 13.3.3, 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 13.37). Consider the comments about the model made in Section 13.5. Create a better model. Fit it with both the quasiPoisson and the Negative Binomial. Plot graphs like those in Figure 13.3. 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 13.27 (the formula for the probability mass function) on Page 371 is true.

7. Given the probability mass function in Equation 13.27, prove $\mathbb{E}[Y] = \mu$ and $\mathbb{V}[Y] = \mu + \mu^2/\theta$.

8. Given the definition of the Negative Binomial distribution (Equations 13.25 and 13.26), prove that an overdispersion of $\theta = \infty$ reduces the Negative Binomial to a Poisson.

13.7.3 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.

13.7.4 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.