Chapter 16:

Poisson Regression Modeling[1]

Dominique Lord
Zachry Dept. of
Civil Engineering
Texas A & M University
College Station, TX

Byung-Jung Park
Korea Transport Institute
Goyang, South Korea

Ned Levine
Ned Levine & Associates
Houston, TX

[1] The code for the Poisson and Negative Binomial models was developed by Ian Cahill of Cahill Software, Edmonton, Alberta, based on his MLE++ software package (http://cahillsoftware.com/2122/index.html). The integration and stepwise procedures were developed by us with the programming by Ms. Haiyan Teng of Houston.

Table of Contents

Chapter 16:
Poisson Regression Modeling

In this chapter, we discuss Poisson models for estimating count variables.

Count Data Models

In chapter 15, we examined Ordinary Least Squares (OLS) regression models. We showed that these models were bound by some strong assumptions of a normally-distributed dependent variable and errors that were normal and constant. We then demonstrated that OLS models are inadequate for describing skewed distributions, particularly counts. Given that crime analysis usually involves the analysis of counts, this is a serious deficiency.

Poisson Regression

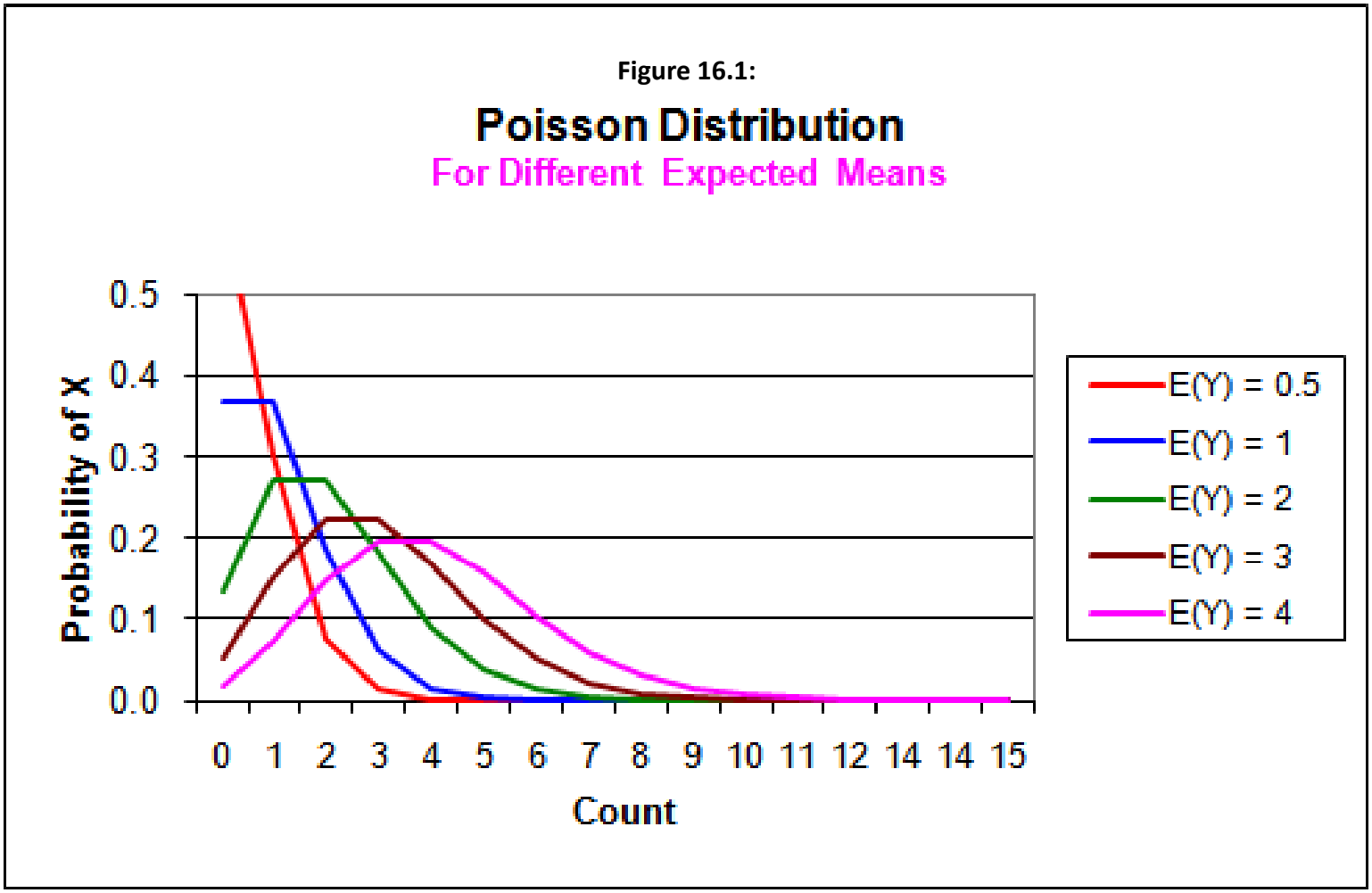
Consequently, we turn to count data models, in particular the Poisson family of models. This family is part of the generalized linear models (GLMs), in which the OLS normal model described above is a special case (McCullagh & Nelder, 1989). Poisson regression is a modeling method that overcomes some of the problems of traditional regression in which the errors are assumed to be normally distributed (Cameron & Trivedi, 1998). In the model, the number of events is modeled as a Poisson random variable with a probability of occurrence being:

$$\text{Prob}(y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \tag{16.1}$$

where y_i is the count for one group or class, i, λ is the mean count over all groups, and e is the base of the natural logarithm. The distribution has a single parameter, λ, which is both the mean and the variance of the function.

The "law of rare events" assumes that the total number of events will approximate a Poisson distribution *if* an event occurs in any of a large number of trials but the probability of occurrence in any given trial is small and assumed to be constant (Cameron & Trivedi, 1998). Thus, the Poisson distribution is very appropriate for the analysis of rare events such as crime incidents (or motor vehicle crashes or uncommon diseases or any other rare event). The Poisson model is not particularly good if the probability of an event is more balanced; for that, the normal distribution is a better model as the sampling distribution will approximate normality with increasing sample size. Figure 16.1 illustrates the Poisson distribution for different expected means.

Figure 16.1:


Poisson Distribution
For Different Expected Means
Probability of X
0.5
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 14 14 15
Count
E(Y) = 0.5
E(Y) = 1
E(Y) = 2
E(Y) = 3
E(Y) = 4

The Poisson distribution is part of a large family known as the exponential family of distributions (McCullagh & Nelder, 1989). The probability distribution for this family is expressed as (Hilbe, 2008):

$$f(y_i: \mu, \Phi) = e^{\left\{\frac{y_i\theta_i - b(\theta_i)}{\alpha(\Phi)} + C(y_i:\Phi)\right\}} \tag{16.2}$$

where θ_i is the canonical parameter or *link* function for observation i, b(θ_i) is the cumulant for observation i, α(Φ) is the scale parameter which is set to one in discrete and count models, and C(y_i: Φ) is a normalization (scaling) term that guarantees that the probability function sums to 1. This family of functions is unique in that the first and second derivatives of the cumulant, with respect to θ, produce the mean and variance function (Hilbe, 2008). All members of the class of generalized linear models can be converted to the exponential form.

Since the Poisson family is a member of the exponential family, the mean can be modeled as a function of some other variables (the independent variables). Given a set of observations on one or more independent variables, $\mathbf{x}_i^T = (1, x_{1i}, \cdots, x_{Ki})$, the *conditional mean* of y_i can be specified as an exponential function of the *x*'s:

$$E(y_i \mid \mathbf{x}_i) = \lambda_i = e^{\mathbf{x}_i^T\boldsymbol{\beta}} \tag{16.3}$$

where *i* is an observation, $\mathbf{x}_i^T$ is a set of independent variables including an intercept, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_K)^T$ are a set of coefficients, and *e* is the base of the natural logarithm. Equation 16.3 can be also written as:

$$\ln(\lambda_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki} \tag{16.4}$$

where each independent variable, k, is multiplied by a coefficient, β_k, and is added to a constant, β_0. In expressing the equation in this form, we have transformed it using a ***link*** function, the link being the log-linear relationship. As discussed above, the Poisson model is part of the GLM framework in which the functional relationship is expressed as a linear combination of predictive variables. This type of model is sometimes known as a ***loglinear*** model as the natural log of the mean is a linear function of *K* independent variables and an intercept.

However, we will refer to it as a *Poisson model*. In more familiar notation, this is

$$\ln(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} \tag{16.5}$$

For the Poisson model, the log-likelihood is:

$$\ln L = \sum_{i=1}^{N} \left[-\lambda_i + y_i \ln(\lambda_i) - \ln y_i! \right] \tag{16.6}$$

where $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ is the conditional mean for zone i, and y_i is the observed number of events for observation i. Anselin provides a more detailed discussion of these functions in Appendix B. The data are assumed to reflect the Poisson model and the variance equals the mean. Therefore, it is expected that the residual errors should increase with the conditional mean. That is, there is inherent heteroscedasticity in a Poisson model (Cameron & Trivedi, 1998). This is different than a normal model where the residual errors are expected to be constant.

The model is estimated using a maximum likelihood (MLE) procedure, typically the Newton-Raphson method or, occasionally, using Fisher scores (Wikipedia, 2010; Cameron & Trivedi, 1998). In Appendix B, Anselin presents a more formal treatment of both the normal and Poisson regression models including the methods by which they are estimated.

Advantages of the Poisson Regression Model

The Poisson model overcomes some of the problems of the normal model. First, the Poisson model has a minimum value of 0. It will not predict negative values. This makes it ideal for a distribution in which the mean or the most typical value is close to 0. Second, the Poisson is a fundamentally skewed model; that is, it is data characterized with a long 'right tail'. Again, this model is appropriate for counts of rare events, such as crime incidents.

Third, because the Poisson model is estimated by the maximum likelihood method, the estimates are adapted to the actual data. In practice, this means that the sum of the predicted values is virtually identical to the sum of the input values, with the exception of a very slight rounding off error.

Fourth, compared to the normal model, the Poisson model generally gives a better estimate of the counts for each record. The problem of over- or underestimating the number of incidents for most records with the normal model is usually lessened with the Poisson. When the residual errors are calculated, generally the Poisson has a lower total error than the normal model, as was illustrated in chapter 15.

In short, the Poisson model has some desirable statistical properties that make it very useful for predicting crime incidents.

Example of Poisson Regression

Using the same Houston burglary database as in chapter 15, we estimate a Poisson model of the two independent predictors of burglaries (Table 16.1).

Likelihood Statistics

Log-likelihood

The summary statistics are quite different from the normal model. In the *CrimeStat* implementation, there are five separate statistics about the likelihood, representing a joint probability function that is maximized. First, there is the log-likelihood (L). The likelihood function is the joint (product) density of all the observations given values for the coefficients and the error variance. The log-likelihood is the log of this product or the sum of the individual densities. Because the function maximizes a probability, which is always between 0 and 1, the log-likelihood is ***always*** negative with a Poisson model.

Note that in comparing two models, the model with the **smallest** log-likehood will fit the data better assuming that the data set and the dependent variable are the same. For example, if one model has a log-likelihood of -4,000 and a second model on the same data set and dependent variable has a log-likelihood of -5,000, the first model is better because it has a *smaller* log-likelihood than the second model. While this is unintuitive, it makes sense in terms of probability theory. If the probability of the first model is 0.6 and that of the second 0.4, then the log-likelhood of the first model will be -0.51 and that of the second -.91. Since a likelihood is the product of the densities of each individual case (and, therefore, the log-likelihood is the sum of the individual logarithms), in practice the log-likelihood is proportional to the probability.

Aikaike Information Criterion (AIC)

Second, the Aikaike Information Criterion (AIC) adjusts the log-likelihood for degrees of freedom since adding more variables will always increase the log-likelihood. It is defined as:

$$AIC = -2L + 2(K+1) \quad (16.7)$$

where L is the log-likelihood and K is the number of independent variables. The model with the lowest AIC is 'best'.

Table 16.1:

Predicting Burglaries in the City of Houston: 2006

Poisson Model

(N= 1,179 Traffic Analysis Zones)

DepVar:	**2006 BURGLARIES**		
N:	1,179		
Df:	1,175		
Type of regression model:	Poisson		
Method of estimation:	Maximum likelihood		
Likelihood statistics			
Log-likelihood:	-13,639.5		
AIC:	27,287.1		
BIC/SC:	27,307.4		
Deviance:	23,021.4	p:	0.0001
Pearson Chi-square:	24,804.4	p:	0.0001
Model error estimates			
Mean absolute deviation:	16.0		
1st (highest) quartile:	33.9		
2nd quartile:	7.3		
3rd quartile:	8.8		
4th (lowest) quartile:	13.9		
Mean squared predicted error:	714.2		
1st (highest) quartile:	2,351.8		
2nd quartile:	203.7		
3rd quartile:	99.8		
4th (lowest) quartile:	206.7		
Dispersion tests			
Adjusted deviance:	19.6	p: 0.0001	
Adjusted Pearson Chi-Square:	21.1	p: 0.0001	
Dispersion multiplier:	21.1	p: 0.0001	Inverse dispersion multiplier: 0.05

Predictor	DF	Coefficient	Stand Error	Tolerance	VIF	Z-value	p
INTERCEPT	1	2.8745	0.014	-	-	212.47	0.001
HOUSEHOLDS	1	0.0006	0.000004	0.994	1.006	146.24	0.001
MEDIAN HOUSEHOLD INCOME	1	-0.000009	0.00000	0.994	1.006	-28.68	0.001

Bayes Information Criterion (BIC/SC)

Third, another measure which is very similar is the *Bayes Information Criterion* (BIC/SC, sometimes called *Schwartz Criterion*), which is defined as:

$$\text{BIC/SC} = -2L+[(K+1)\ln(N)] \tag{16.8}$$

These two measures penalize the number of parameters added in the model, and reverse the sign of the log-likelihood (L) so that the statistics are more intuitive. The model with the lowest BIC/SC value is 'best'.

Deviance

Fourth, a decision about whether the Poisson model is appropriate can be based on the statistic called the *deviance* which is defined as:

$$Dev = 2(L_F - L_M) = 2\sum_{i=1}^{N}\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - y_i - \hat{\lambda}_i \right] \tag{16.9}$$

where L_F is the log-likelihood that would be achieved if the model gave a perfect fit and L_M is the log-likelihood of the model under consideration. If the latter model is correct, the deviance (*Dev*) is approximately χ^2 distributed with degrees of freedom equal to $N-(K+1)$. A value of the deviance greatly in excess of $N-(K+1)$ suggests that the model is over-dispersed due to missing variables or non-Poisson form. This statistic is sometimes called the G^2 statistic (Bishop, Feinberg, & Holland, 1975). The deviance has N-*K*-1 degrees of freedom where *K* is the number of parameters estimated (including the constant).

Pearson Chi-square

Fifth, there is the Pearson Chi-square statistic which is defined by

$$Pearson-\chi^2 = \sum_{i=1}^{N}\frac{(y_i-\hat{\mu}_i)^2}{VAR(y_i)} \tag{16.10}$$

If the mean and the variance are properly specified, then $E\left[\sum_{i=1}^{N}(y_i - \frac{\mu_i)^2}{VAR(y_i)})\right] = N$ (Cameron and Trivedi, 1998). Values closer to N (the sample size) show a better fit. The Pearson Chi-square has N-*K*-1 degrees of freedom where *K* is the number of parameters

estimated (including the constant). Note, that the expected value depends on the variance function, which we will discuss below.

Model Error Estimates

Next, there are two statistics that measure how well the model fits the data (goodness-of-fit). Mean Absolute Deviation (MAD) and Mean Squared Predicted Error (MSPE) were defined in Chapter 15. Comparing these with the results of the normal model (Table 15.1), it can be seen that the overall MAD and MSPE are slightly worse than for the normal model, though much better than with the log transformed linear model (Table 15.4). Comparing the four quartiles, it can be seen that for three of the four quartiles the normal model had slightly better MAD and MSPE scores than for the Poisson but the differences were not great.

Dispersion Tests

The remaining four summary statistics measure *dispersion*. A more extensive discussion of dispersion is given later in the chapter. But, very simply, in the Poisson framework, the variance should equal the mean. These statistics indicate the extent to which the variance exceeds the mean.

First, the *adjusted deviance* is defined as the deviance divided by the degrees of freedom (N-K-1); a value closer to 1 indicates a satisfactory goodness-of-fit. Usually, values greater than 1 indicate signs of over-dispersion.

Second, the *adjusted Pearson Chi-square* is defined as the Pearson Chi-square divided by the degrees of freedom; again, a value closer to 1 indicates a satisfactory goodness-of-fit.

Third, the *dispersion* multiplier, γ, measures the extent to which the conditional variance exceeds the conditional mean (conditional on the independent variables and the intercept term) and is defined by $Var(y_i) = \lambda_i + \gamma\lambda_i^2$. The Z-test of the dispersion multiplier indicates whether the amount of dispersion is significantly greater than that assumed by the Poisson model (Hilbe, 2008). The test is:

$$Z = \frac{(\sum(y_i - \mu_i)^2 - y_i)}{\sum \mu_i \sqrt{2}} \qquad (16.11)$$

where y_i is the observed value of Y and μ_i is the predicted value of Y. The statistic is a test of *over-dispersion*, that the conditional variance is *greater* than the conditional mean. A significant value for Z indicates that the assumption of equi-dispersion of the conditional variance is

rejected and the model should be estimated as a negative binomial or lognormal for over-dispersion.

In some cases, there may be *under-dispersion*, that is where the conditional variance is less than the conditional mean. In this case, a Poisson with linear correction should be used. Unfortunately, the Z-test will identify that as being not significant. We are not aware of a good test for under-dispersion and the user will have to use judgment.

Fourth, the *inverse dispersion multiplier* (ψ) is simply the reciprocal of the dispersion multiplier ($\psi = 1/\gamma$); some users are more familiar with it in this form.

As seen in Table 16.1, the four dispersion statistics are much greater than 1 and indicate *over-dispersion*. In other words, the conditional variance is greater – in this case, much greater, than the conditional mean. The 'pure' Poisson model (in which the variance is supposed to equal the mean) is not an appropriate model for these data.

Individual Coefficient Statistics

Finally, the signs of the coefficients are the same as for the normal and transformed normal models, as would be expected. The relative strengths of the variables, as seen through the Z-values, are also approximately the same.

In short, the Poisson model has produced results that are an alternative to the normal model. While the likelihood statistics indicate that, in this instance, the normal model is slightly better, the Poisson model has the advantage of being theoretically sounder. In particular, it is not possible to get a minimum predicted value less than zero (which is possible with the normal model) and the sum of the predicted values will always equal the sum of the input values (which is rarely true with the normal model). With a more skewed dependent variable, the Poisson model will usually fit the data better than the normal as well.

Problems with the Poisson Regression Model

On the other hand, the Poisson model is not perfect. The primary problem is that count data are usually *over-dispersed*.

Over-dispersion in the Residual Errors

In the Poisson distribution, the mean equals the variance. In a Poisson regression model, the mathematical function, therefore, equates the conditional mean (the mean controlling for all the predictor variables) with the conditional variance. However, most actual distributions have a

high degree of skewness, much more than are assumed by the Poisson distribution (Cameron & Trivedi, 1998; Mitra & Washington, 2007).

As an example, figure 16.2 shows the distribution of Baltimore County and Baltimore City crime origins and Baltimore County crime destinations by TAZ. For the origin distribution, the ratio of the variance to the mean is 14.7; that is, the variance is 14.7 times that of the mean! For the destination distribution, the ratio is 401.5!

In other words, the simple variance is many times greater than the mean. We have not yet estimated some predictor variables for these variables, but it is probable that even when this is done the conditional variance will far exceed the conditional mean. Many real-world count data are similar to this; the variance will usually be much greater than the mean (Lord, 2006) although, occasionally, the variance can be smaller than the conditional mean (Lord, 2010). What this means in practice is that the residual errors - the difference between the observed and predicted values for each zone, will be greater than what is expected. The Poisson model calculates a standard error as if the variance equals the mean. Thus, the standard error will be underestimated using a Poisson model and, therefore, the significance tests (the coefficient divided by the standard error) will be greater than they really should be. In a Poisson multiple regression model, we might end up selecting variables that really should not be selected because we think they are statistically significant when, in fact, they are not (Park & Lord, 2007).

Under-dispersion in the Residual Errors

There are also cases where the conditional variance is less than the conditional mean (under-dispersion). This happens sometimes with crime data. For example, in an analysis of drunk driving crashes in Baltimore County, we found that the modeled variance was substantially less than the modeled mean (Levine & Canter, 2011). In both cases, one needs to correct the estimated standard error from the Poisson model.

To visualize over- and under-dispersion, Figure 16.3 shows three different skewed distributions, over-dispersed, equi-dispersed (Poisson), and under-dispersed. These are based on the variance-to-mean ratios of the raw data. Note that the over-dispersed distribution is extremely skewed while the under-dispersed distribution is mildly skewed. Still, with under-distribution, one cannot assume a normal distribution because it will still underestimate the high values of the dependent variable.

Also, the actual dispersion is conditional on the independent variables (i.e., after the model has been run). However, Cameron and Trivedi (1998) suggest that if the raw variance-to-mean ratio is less than 2.0, most likely the conditional variance will be less than the conditional mean.

Figure 16.2:

Distribution of Crime Origins and Destinations: Baltimore County, MD: 1993-1997

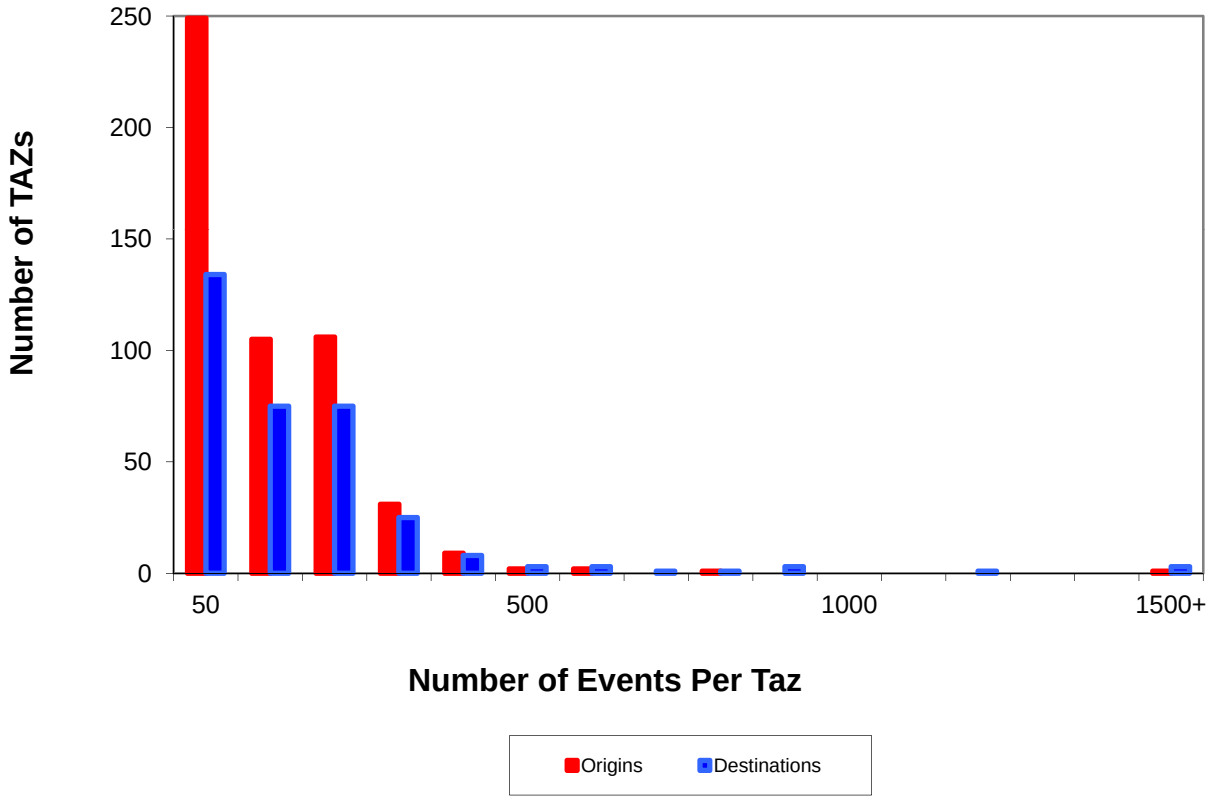
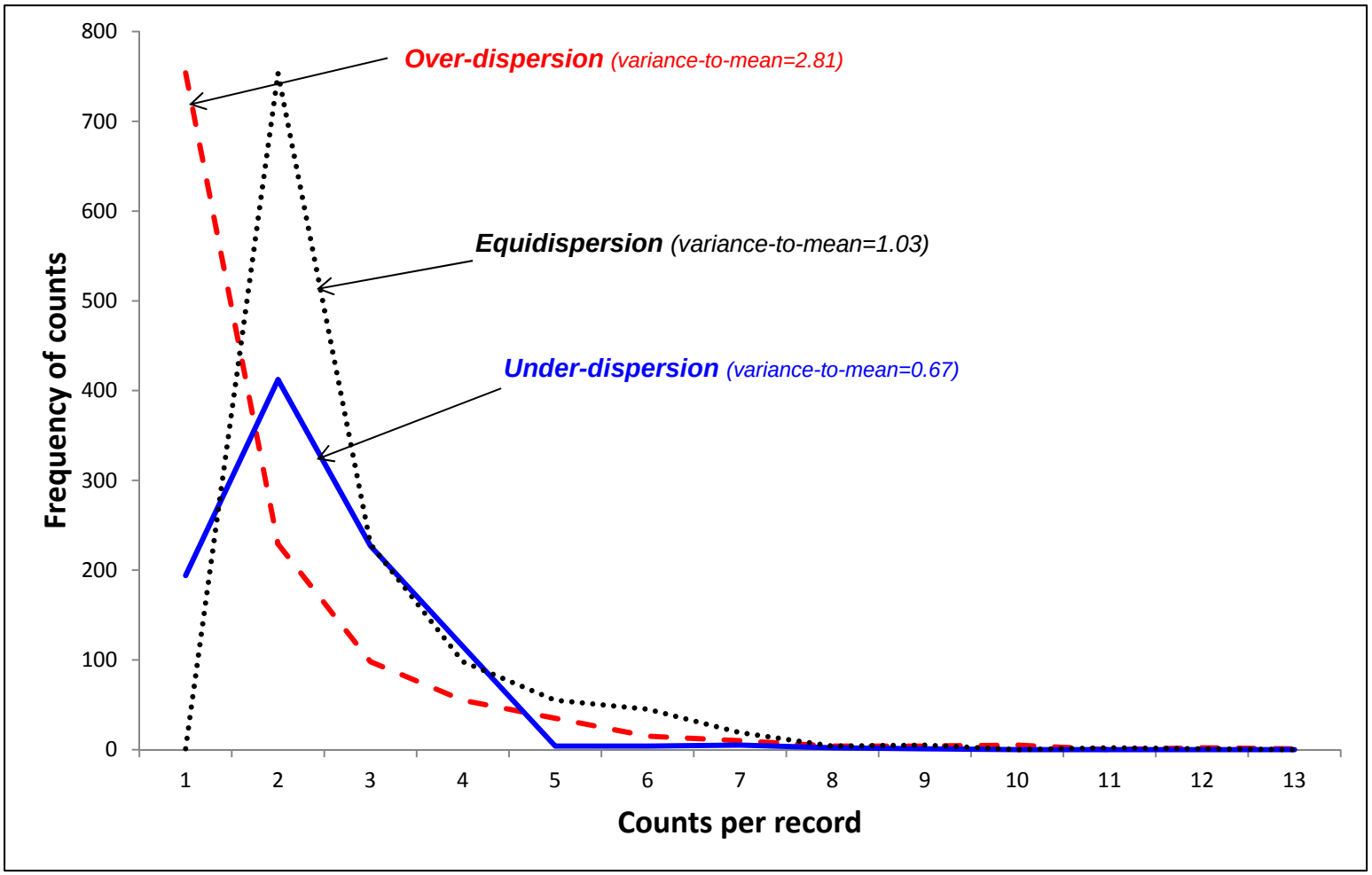


Figure 16.3:

Skewed Distributions and Type of Dispersion



Poisson Regression with Linear Dispersion Correction

There are a number of methods for correcting the over-dispersion in a count model. Most of them involve modifying the assumption of the conditional variance equal to the conditional mean. The first is a simple linear correction known as the *linear negative binomial* (or NB1 model; Cameron & Trivedi, 1998, 63-65). The variance of the function is assumed to be a linear multiplier of the mean. The conditional variance is defined as:

$$\omega_i = V[y_i \mid \mathbf{x}_i] \tag{16.12}$$

where $V[y_i \mid \mathbf{x}_i]$ is the variance of y_i given the independent variables.

The conditional variance is then a function of the mean:

$$\omega_i = \lambda_i + \tau\lambda_i^p \tag{16.13}$$

where τ is the *dispersion parameter* and p is a constant (usually 1 or 2). In the case where p is 1, the equation simplifies to:

$$\omega_i = \lambda_i + \tau\lambda_i \tag{16.14}$$

This is the NB1 correction. In the special case where $\tau = 0$, the variance becomes equal to the mean (the Poisson model). The model is estimated in two steps. First, the Poisson model is fitted to the data and the degree of over- (or under) dispersion is estimated. The dispersion parameter is defined as:

$$\hat{\tau} = 1/\hat{\psi} = \frac{1}{N-K-1}\sum_{i=1}^{N}\frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} - 1 \tag{16.15}$$

where *N* is the sample size, *K* is the number of independent variables, Y_i is the observed number of events that occur in observation (or zone) *i*, and $\hat{\lambda}_i$ is the predicted number of events for observation (or zone) *i*. The test is similar to an average chi-square in that it takes the square of the residuals $(y_i - \hat{\lambda}_i)^2$ and divides it by the predicted values, and then averages it by the degrees of freedom. The dispersion parameter is a standardized number. A value greater than 0 indicates over-dispersion while a value less than 0 indicates under-dispersion. A value of 0 indicates *equidispersion* (or the variance equals the mean). The dispersion parameter can also be estimated based on the deviance.

In the second step, the Poisson standard error is multiplied by the square root of the dispersion parameter to produce an *adjusted standard error*:

$$SE_{adj} = SE \times \sqrt{\hat{\tau}} \tag{16.16}$$

The new standard error is then used with the t-test to produce an adjusted t-value. This adjustment is found in most Poisson regression packages using a Generalized Linear Model (GLM) approaches (McCullagh and Nelder, 1989, 200). Cameron & Trivedi (1998) have shown that this adjustment produces results that are virtually identical to that of the negative binomial, but involving fewer assumptions. *CrimeStat* includes an NB1 correction and is called *Poisson with linear correction*.

Example of Poisson Model with Linear Dispersion Correction (NB1)

Table 16.2 shows the results of running the Poisson model with the linear dispersion correction. The likelihood statistics are the same as for the simple Poisson model (Table 16.1) and the coefficients are identical. The dispersion parameter, however, has now been adjusted to be 1.0. This affects the standard errors, which are now greater. In the example, the two independent variables are still statistically significant, but the Z-values are smaller.

Poisson-Gamma (Negative Binomial) Regression

A second type of dispersion correction involves a mixed function model. Instead of simply adjusting the standard error by a dispersion correction, different assumptions are made for the mean and the variance (dispersion) of the dependent variable. In the *negative binomial* model, the number of observations (Y_i) is assumed to follow a Poisson distribution but the mean (λ_i) follows a Gamma distribution (Lord, 2006; Cameron & Trivedi, 1998, 62-63; Venables & Ripley, 1997, 242-245). This is frequently called an NB2 model.

Mathematically, the negative binomial distribution is one derivation of the binomial distribution in which the sign of the function is negative, hence the term *negative binomial* (for more information on the derivation, see Wikipedia, 2010). For our purposes, it is defined as a mixed distribution with a Poisson mean and a one parameter Gamma dispersion function having the form:

$$f(y_i / \theta_i) = \frac{e^{-\theta_i} \theta_i^{y_i}}{y_i!} \tag{16.17}$$

Table 16.2:

Predicting Burglaries in the City of Houston: 2006
Poisson with Linear Dispersion Correction Model (NB1)

(N= 1,179 Traffic Analysis Zones)

DepVar:	**2006 BURGLARIES**			
N:	1,179			
Df:	1,175			
Type of regression model:	Poisson with linear dispersion correction			
Method of estimation:	Maximum likelihood			
Likelihood statistics				
Log-likelihood:	-13,639.5			
AIC:	27,287.1			
BIC/SC :	27,307.4			
Deviance:	12,382.5	p: 0.0001		
Pearson Chi-square:	12,402.2	p: 0.0001		
Model error estimates				
Mean absolute deviation:	16.0			
1st (highest) quartile:	33.9			
2nd quartile:	7.3			
3rd quartile:	8.8			
4th (lowest) quartile:	13.9			
Mean squared predicted error:	714.2			
1st (highest) quartile:	2,351.8			
2nd quartile:	203.7			
3rd quartile:	99.8			
4th (lowest) quartile:	206.7			
Dispersion tests				
Adjusted deviance:	10.5	P: 0.001		
Adjusted Pearson Chi-Square:	10.6	p: 0.001		
Dispersion multiplier:	1.0	p: n.s.	Inverse dispersion multiplier:	1.0

Predictor	DF	Coefficient	Stand Error	Tolerance	VIF	Z-value	p
INTERCEPT	1	2.87452	0.062	-	-	46.26	0.001
HOUSEHOLDS	1	0.00059	0.00002	0.994	1.006	31.84	0.001
MEDIAN HOUSEHOLD INCOME	1	-0.000009	0.000001	0.994	1.006	-6.24	0.001

where

$$\theta_i = e^{\beta_0 + (\Sigma \beta_i x_i) + \varepsilon_i} \tag{16.18}$$

$$\theta_i = e^{\beta_0 + (\Sigma \beta_i x_i)} e^{\varepsilon_i} \tag{16.19}$$

$$\theta_i = \mu_i \nu_i \tag{16.20}$$

and where θ_i is a function of a one-parameter gamma distribution where the parameter, τ, is greater than 0 (ignoring the subscripts):

$$h(y/\mu,\tau) = \frac{\Gamma(\tau^{-1}+y)}{\Gamma(\tau^{-1})\Gamma(y+1)}\left(\frac{(\tau^{-1})}{\tau^{-1}+\mu}\right)^{\tau^{-1}}\left(\frac{\mu}{\tau^{-1}+\mu}\right)^{y} \tag{16.21}$$

The model is used traditionally with integer (count) data though it can also be applied to continuous (real) data. Sometimes the integer model is called a *Pascal* model while the real model is called a *Polya* model (Wikipedia, 2010; Springer, 2010). Boswell and Patil (1970) argued that there are at least 12 distinct probabilistic processes that can give rise to the negative binomial function including heterogeneity in the Poisson intensity parameter, cluster sampling from a population which is itself clustered, and the probabilities that change as a function of the process history (i.e., the occurrence of an event breeds more events). The interpretation we adopt here is that of a heterogeneous population with different observations coming from different sub-populations, and the Gamma distribution is the mixing variable.

Because both the Poisson and Gamma functions belong to the single-parameter exponential family of functions and are convex in shape (increasing smoothly up to a peak and then decreasing smoothly), they can be solved by the maximum likelihood method. The mean is always estimated as a Poisson function. However, there are slightly different parameterizations of the variance function (Hilbe, 2008). In the original derivation by Greenwood and Yule (1920), the conditional variance was defined as:

$$\omega_i = \mu_i + \mu_i^2/\psi \tag{16.22}$$

whereupon ψ (Psi) became known as the *inverse dispersion parameter* (McCullagh & Nelder, 1989).

However, in more recent years, the conditional variance was defined within the Generalized Linear Models tradition as a direct adjustment of the squared Poisson mean, namely:

$$\omega_i = \mu_i + \tau \mu_i^2 \quad (16.23)$$

where the variance is now a quadratic function of the Poisson mean (i.e., p is 2 in formula 16.13) and τ is called the *dispersion multiplier*. This is the formulation proposed by Cameron & Trivedi (1998; pp. 62-63). That is, it is assumed that there is an unobserved variable that affects the distribution of the count so that some observations come from a population with higher expected counts whereas others come from a population with lower expected counts. The model then has a Poisson mean but with a 'longer tail' variance function. The dispersion parameter, τ, is directly related to the amount of dispersion. This is the interpretation that we will use in the chapter and in *CrimeStat*.

Formally, we can write the negative binomial model as a Poisson-gamma mixture form:

$$y_i \mid \lambda_i \sim Poisson(\lambda_i) \quad (16.24)$$

The Poisson mean λ_i is organized as:

$$\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i) \quad (16.25)$$

where exp() is an exponential function, $\boldsymbol{\beta}$ is a vector of unknown coefficients for the *k* covariates plus an intercept, and ε_i is the model error independent of all covariates. The $\exp(\varepsilon_i)$ is assumed to follow the gamma distribution with a mean equal to 1 and a variance equal to $\tau = 1/\psi$ where ψ is a parameter that is greater than 0 (Lord, 2006; Cameron & Trivedi, 1998).

For a negative binomial generalized linear model, the deviance can be computed the following way:

$$D = \sum_{i=1}^{N} \left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i + \hat{\psi}) \ln\left(\frac{y_i + \hat{\psi}}{\hat{\lambda}_i + \hat{\psi}}\right) \right] \quad (16.26)$$

For a well-fitted model the deviance should be approximately χ^2 distributed with $N - K - 1$ degrees of freedom (McCullagh and Nelder, 1987). If $D/(N - K - 1)$ is close to 1, we generally conclude that the model's fit is satisfactory.

Example 1 of Negative Binomial Regression

To illustrate, Table 16.3 presents the results of the negative binomial model for Houston burglaries. Even though the individual coefficients are similar, the likelihood statistics indicate

that the model fit the data better than the Poisson with linear correction for over-dispersion. The log-likelihood is higher, the AIC and BIC/SC statistics are lower as are the deviance and the Pearson Chi-square statistics.

On the other hand, the model error is higher than for the Poisson and Poisson NB1 models, both for the mean absolute deviation (MAD) and the mean squared predicted error (MSPE). Accuracy and precision need to be seen as two different dimensions for any method, including a regression model (Jessen, 1979, 13-16). Accuracy is 'hitting the target', in this case maximizing the likelihood function. Precision is the consistency in the estimates, again in this case the ability to replicate individual data values. A normal model will often produce lower overall error because it minimizes the sum of squared residual errors though it rarely will replicate the values of the records with high values and often does poorly at the low end.

For this reason, we say that the negative binomial is a more accurate model though not necessarily a more precise one. To improve the precision of the negative binomial, we would have to introduce additional variables to reduce the conditional variance further. Clearly, residential burglaries are associated with more variables than just the number of households and the median household income (e.g., ease of access into buildings, lack of surveillance on the street, having easy contact with individuals willing to distribute stolen goods).

Nevertheless, the negative binomial is a better model than the Poisson and certainly the normal, Ordinary Least Squares. It is theoretically sounder and does better with highly skewed (over-dispersed) data. In Appendix C, Lord and Park present a more formal presentation of the model.

Example 2 of Negative Binomial Regression with Highly Skewed Data

To illustrate further, the negative binomial is very useful when the dependent variable is extremely skewed. Figure 16.4 show the number of crimes committed (and charged for) by individual offenders in Manchester, England in 2006. The X-axis plots the number of crimes committed while the Y-axis plots the number of offenders. Of the 56,367 offenders, 40,755 committed one offence during that year, 7,500 committed two offences, and 3,283 committed three offences. At the high end, 26 individuals committed 30 or more offences in 2006 with one individual committing 79 offences. The distribution is very skewed.

A negative binomial regression model was set up to model the number of offences committed by these individuals as a function of conviction for previous offence (prior to 2006), age, and distance that the individual lived from the city center. Table 16.4 shows the results.

Table 16.3:
Predicting Burglaries in the City of Houston: 2006
MLE Negative Binomial Model
(N= 1,179 Traffic Analysis Zones)

DepVar:	**2006 BURGLARIES**			
N:	1,179			
Df:	1,175			
Type of regression model:	Poisson with Gamma dispersion			
Method of estimation:	Maximum likelihood			
Likelihood statistics				
Log-likelihood:	-4,430.8			
AIC:	8,869.6			
BIC/SC :	8,889.9			
Deviance:	1,390.1	p: 0.0001		
Pearson Chi-square:	1,112.7	p: n.s.		
Model error estimates				
Mean absolute deviation:	39.6			
1st (highest) quartile:	124.1			
2nd quartile:	19.4			
3rd quartile:	6.2			
4th (lowest) quartile:	8.9			
Mean squared predicted error:	62,031.2			
1st (highest) quartile:	242,037.1			
2nd quartile:	6,445.8			
3rd quartile:	118.3			
4th (lowest) quartile:	154.9			
Dispersion tests				
Adjusted deviance:	1.2	p: n.s		
Adjusted Pearson Chi-Square:	0.9	p: n.s.		
Dispersion multiplier:	1.5	p: n.s.	Inverse dispersion multiplier: 0.7	

Predictor	DF	Coefficient	Stand Error	Tolerance	VIF	Z-value	p
INTERCEPT	1	2.3210	0.083	-	-	27.94	0.001
HOUSEHOLDS	1	0.0012	0.00007	0.994	1.006	17.66	0.001
MEDIAN HOUSEHOLD INCOME	1	-0.00001	0.000002	0.994	1.006	-5.13	0.001

Figure 16.4:

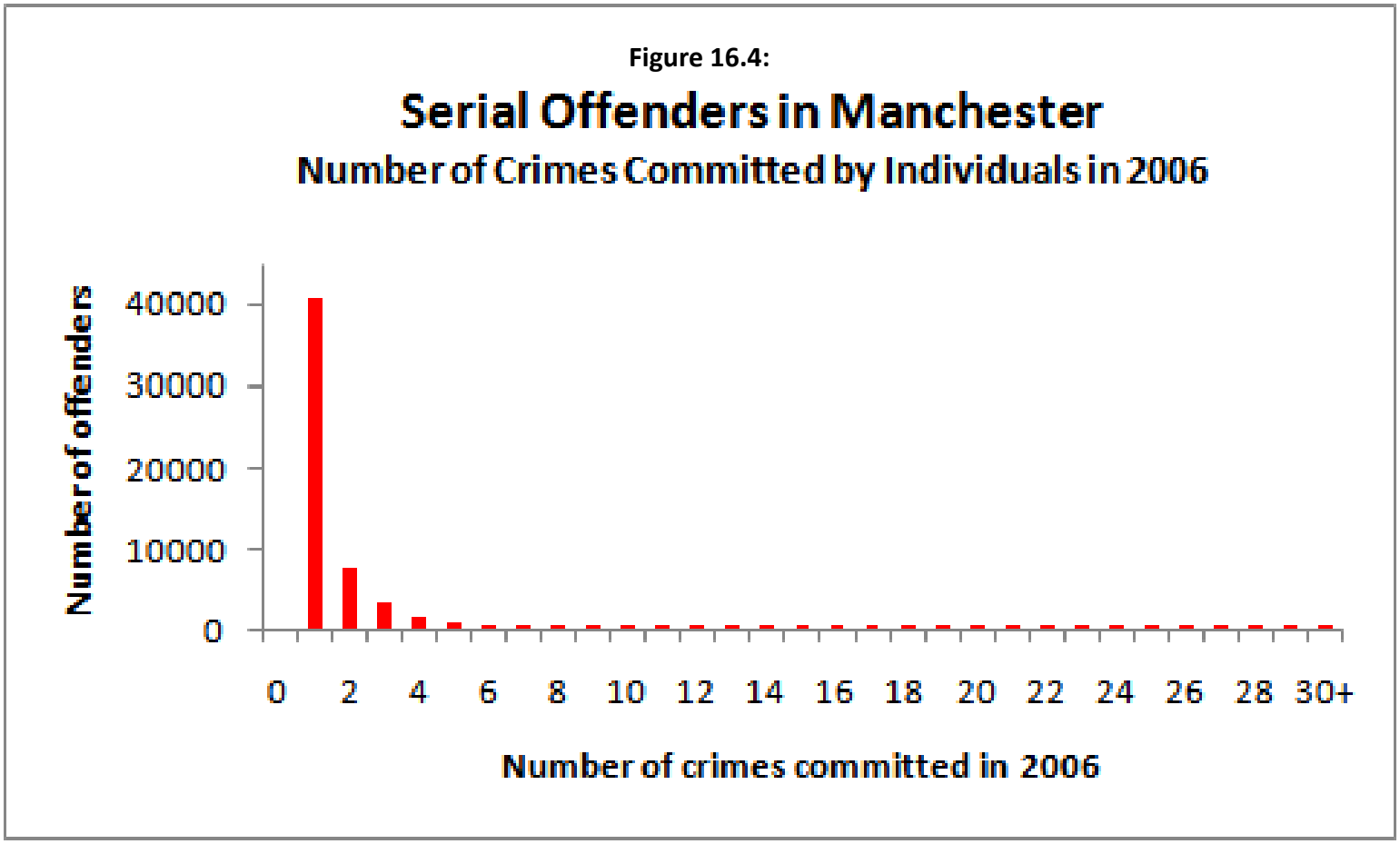

Serial Offenders in Manchester
Number of Crimes Committed by Individuals in 2006
Number of offenders
40000
30000
20000
10000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30+
Number of crimes committed in 2006

Table 16.4:

Number of Crimes Committed in Manchester in 2006
Negative Binomial Model

(N= 56,367 Offenders)

DepVar:	**NUMBER OF CRIMES COMMITTED IN 2006**				
N:	56,367				
Df:	56,362				
Type of regression model:	Poisson with Gamma dispersion				
Method of estimation:	Maximum likelihood				
Likelihood statistics					
Log-likelihood:	-89,103.7				
AIC:	178,217.4				
BIC/SC:	178,262.1				
Deviance:	36,616.6	p: n.s.			
Pearson Chi-square:	80,950.2	p: 0.0001			
Model error estimates					
Mean absolute deviation:	0.93				
1st (highest) quartile:	1.9				
2nd quartile:	0.7				
3rd quartile:	0.6				
4th (lowest) quartile:	0.6				
Mean squared predicted error:	3.90				
1st (highest) quartile:	13.8				
2nd quartile:	0.7				
3rd quartile:	0.6				
4th (lowest) quartile:	0.6				
Dispersion tests					
Adjusted deviance:	0.6	p: n.s.			
Adjusted Pearson Chi-Square:	1.4	p: n.s.			
Dispersion multiplier:	0.2	p : n.s.	Inverse dispersion multiplier:	6.2	

Predictor	DF	Coefficient	Stand Error	Tolerance	Z-value	p
INTERCEPT	1	0.509	0.012	-	41.90	0.001
DISTANCE FROM CITY CENTER	1	-0.022	0.003	0.999	-6.74	0.001
PRIOR OFFENCE	1	0.629	0.008	0.982	80.24	0.001
AGE OF OFFENDER	1	-0.012	0.0003	0.981	-35.09	0.001

The model was discussed in a recent article (Levine & Lee, 2013). The closer an offender lives to the city center, the greater than number of crimes committed. Also, younger offenders committed more offences than older offenders. However, the strongest variable is whether the individual had an earlier conviction for another crime. Offenders who have committed previous offences are more likely to commit more of them again. Crime is a very repetitive behavior!

The likelihood statistics indicates that the model was reasonably closely. The likelihood statistics were better than that of a normal OLS and a Poisson NB1 models (not shown). The model error was also slightly better for the negative binomial. For example, the MAD for this model was 0.93 compared to 0.95 for the normal and 0.93 for the Poisson NB1. The MSPE for this model was 3.90 compared to 3.93 for the normal and also 3.90 for the Poisson NB1. The negative binomial and Poisson models produce very similar results because, in both cases, the means are modeled as Poisson variables. The differences are in the dispersion statistics. For example, the standard error of the four parameters (intercept plus three independent variables was 0.012, 0.003, 0.008, and 0.0003 respectively for the negative binomial compared to 0.015, 0.004, 0.010, and 0.0004 for the Poisson NB1 model. In general, the negative binomial will fit the data better when the dependent variable is highly skewed and will usually produce lower model error.

Advantages of the Negative Binomial Model

The main advantage of the negative binomial model over the Poisson and Poisson with linear dispersion correction (NB 1) is that it incorporates the theory of Poisson but allows more flexibility in that multiple underlying distributions may be operating. Further, mathematically it separates out the assumptions of the mean (Poisson) from that of the dispersion (Gamma) whereas the Poisson with linear dispersion correction only adjust the dispersion after the fact (i.e., it determines that there is over- or under-dispersion and then adjusts it). This is neater from a mathematical perspective. Separating the mean from the dispersion can also allow alternative dispersion estimates to be modeled, such as the lognormal (Lord, 2006). This is very useful for modeling highly skewed data.

Disadvantages of the Negative Binomial Model

The biggest disadvantage is that the constancy of sums is not maintained. Whereas the Poisson model (both "pure" and with the linear dispersion correction) maintains the constancy of the sums (i.e., the sum of the predicted values equals the sum of the input values), the negative binomial does not. Usually, the degree of error in the sum of the predicted values is not far from the sum of the input values. But, occasionally substantial distortions are seen.

A second disadvantage is that the negative binomial model cannot handle under-dispersion. There are crime data sets that we have seen which show under-dispersion. For those, one needs another type of model. In Levine and Canter (2011), a Poisson with linear correction was used to adjust the standard errors (essentially, making them smaller). But, better methods need to be developed.

A final disadvantage of the negative binomial is related to the small sample size and low sample mean bias. It has been shown that the dispersion parameter of NB2 models can be significantly biased or misestimated when not enough data are available for estimating the model (Lord, 2006). For that, a Poisson-lognormal model is a better solution.

Alternative Poisson Regression Models

There are a number of variations of these involving different assumptions about the dispersion term, such as a lognormal function. There are also a number of different Poisson-type models including the zero-inflated Poisson (or ZIP; Hall, 2000), the Generalized Extreme Value family (Weibul, Gumbel and Fréchet), the lognormal function (see NIST 2004 for a list of common non-linear functions), and the Negative binomial-Lindley (Lord and Greedipally, 2011).

There are also alternative methods than maximum likelihood for estimating the likely value of a count given a set of independent predictors. In Chapter 17, we will examine several other approaches to estimating the Poisson model and will develop several alternative Poisson models.

Likelihood Ratios

One test that we have not implemented in the regression I module is the *likelihood ratio* because it is so simple. A likelihood ratio is the ratio of the log-likelihood of one model to that of another. For example, a Poisson-Gamma model run with three independent variables can be compared with a Poisson-Gamma model with two independent variables to see if the third independent variable significantly adds to the prediction.

The test is very simple. Let L_C be the log-likelihood of the comparison model and let L_B be the log-likelihood of the baseline model (the model to which the comparison model is being compared). Then,

$$LR = 2(L_C - L_B) \tag{16.27}$$

LR is distributed as a χ^2 statistic with K degrees of freedom where K is the difference in the number of parameters estimated between the two models including the intercepts. In the

example above, *K* is 1 since a model with three independent variables plus an intercept (d.f. = 4) is being compared with a model with two independent variables plus an intercept (d.f.=3).

Limitations of the Maximum Likelihood Approach

The functions considered up to this point are part of the single-parameter exponential family of functions where the function is smooth and convex. Because of this, maximum likelihood estimation (MLE) can be used. However, there are more complex functions that are not part of this family. Also, some functions come from multiple families and are, therefore, too complex to solve for a single maximum. They may have multiple 'peaks' for which there is not a single optimal solution. For these functions, a different approach has to be used.

Also, one of the criticisms leveled against maximum likelihood estimation (MLE) in general is that the approach *overfits* data. That is, it finds the values of the parameters that maximize the joint probability function. This is similar to the old approach of fitting a curve to data points with higher-order polynomials. While one can find some combination of higher-order terms to fit the data almost perfectly, such an equation has no theoretical basis nor cannot easily be explained. Further, such an equation does not usually do very well as a predictive tool when applied to a new data set.

MLE has been seen as analogous to this approach. By finding parameters that maximize the joint probability density distribution, the approach may be fitting the data too tightly. The original logic behind the AIC and BIC/SC criteria were to penalize models that included too many variables (Findley, 1993). However, these corrections only partially adjust the model. It is still possible to overfit a model with MLE. Radford (2006) has suggested that, in addition to a penalty for too many variables, that the gradient assent in a maximum likelihood algorithm be stopped before reaching the peak. This would require modifying the MLE algorithm substantially.

Further, Nannen (2003) has argued that overfitting creates a paradox because as a model fits the data better and better, it will do worse on other datasets to which it is applied for prediction purposes. In other words, it is better to have a simpler, but more robust, model than one that closely models one data set. Probably the biggest criticism against the MLE approach is that it underestimates the sampling errors by, again, overfitting the parameters (Husmeier & McGuire, 2002).

Instead, we will now examine a method that overcomes some of these difficulties, the Markov Chain Monte Carlo (MCMC) approach. Because the algorithm samples from a larger space rather than maximizes a function *per se*, it has the ability to find solutions to very complex problems for which the MLE approach is not appropriate. Chapter 17 presents this approach.

References

Bishop, Y. M. M., Feinberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press: Cambridge, MA.

Boswell, M. T. & Patil, G. P. (1970). "Chance mechanisms generating negative binomial distributions". In *Random Counts in Scientific Work*, Vol. 1, G. P. Patil, ed., Pennsylvania State University Press:University Park, PA, 3-22.

Cameron, A. Colin & Trivedi, Pravin K. (1998). *Regression Analysis of Count Data*. Cambridge University Press: Cambridge, U.K.

Findley, D. F. (1993). *The Overfitting Principles Supporting AIC*. Statistical Research Division Report Series, SRD Research Report no. CENSUS/SRD/ RR-93/04, U.S. Bureau of the Census: Washington, DC. http://www.census.gov/srd/papers/pdf/rr93-04.pdf.

Geedipally, S.R., D. Lord, S.S. Dhavala (2012) The Negative Binomial-Lindley Generalized Linear Model: characteristics and Application using Crash Data. Accident Analysis & Prevention, in press.

Greenwood, M. & Yule, G. U. (1920). "An inquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents". *Journal of the Royal Statistical Society*, 83, 255-279.

Hall, D. B. (2000). "Zero-inflated Poisson and binomial regression with random effects: a case study". *Biometrics*, 56, 1030-1039.

Hilbe, J. M. (2008). *Negative Binomial Regression (with corrections)*. Cambridge University Press: Cambridge.

Husmeier, D. & McGuire, G. (2002). "Detecting recombination in DNA sequence alignments: A comparison between maximum likelihood and Markov Chain Monte Carlo". Biomathematics and Statistics Scotland, SCRI: Dundee.
http://www.bioss.ac.uk/~dirk/software/BARCEtdh/Manual/em/em.html

Jessen, R.J. (1979). *Statistical Survey Techniques*. John Wiley & Sons: New York.

Levine, N. & Lee, P. (2013). Crime travel of offenders by gender and age in Manchester, England. Leitner, M. (ed), *Crime Modeling and Mapping Using Geospatial Technologies,* Springer. 145-178.

References (continued)

Levine, N. & Canter, P. (2010). "Linking origins with destinations for DWI motor vehicle crashes: An application of Crime Travel Demand modeling". *Crime Mapping*, 3, 7-41.

Lord, D. (2006). "Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter". *Accident Analysis and Prevention*, 38, 751-766.

Lord, D. & Geedipally, S. R. (2011) The Negative Binomial-Lindley Distribution as a Tool for Analyzing Crash Data Characterized by a Large Amount of Zeros. Accident Analysis & Prevention, Vol. 43, No. 5, pp. 1738-1742.

Lord, D., Geedipally, S. R., & Guikema, S. (2010) Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion. Risk Analysis, Vol. 30, No. 8, pp. 1268-1276.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). Chapman & Hall/CRC: Boca Raton, FL.

Mitra, S. & Washington, S. (2007). "On the nature of over-dispersion in motor vehicle crash prediction models", *Accident Analysis and Prevention*, 39, 459-468.

Nannen, V. (2003). *The Paradox of Overfitting*. Artificial Intelligence, Rijksuniversitat: Groningen, Netherlands. http://volker.nannen.com/pdf/the_paradox_of_overfitting.pdf. Accessed March 11, 2010.

NIST (2004). "Gallery of distributions". *Engineering Statistics Handbook*. National Institute of Standards and Technology: Washington, DC. http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm.

Park, E.S., and Lord, D. (2007) Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. In Transportation Research Record 2019: Journal of the Transportation Research Board, TRB, National Research Council, Washington, D.C., pp. 1-6.

Radford, N. (2006). "The problem of overfitting with maximum liklihood" . CSC 411: Machine Learning and Data Mining, University of Toronto: Toronto, CA. http://www.cs.utoronto.ca/~radford/csc411.F06/10-nn-early-nup.pdf Accessed March 11, 2010.

References (continued)

Springer (2010). "Polya distribution", *Encyclopedia of Mathematics*, Springerlink: London, http://eom.springer.de/p/p073540.htm.

Venables, W.N. & Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus (second edition)*. Springer-Verlag: New York.

Wikipedia (2010). "Negative binomial distribution", *Wikipedia*, http://en.wikipedia.org/wiki/Negative_binomial_distribution Accessed February 24, 2010.