



Count Outcomes: Regression Models for Counts

Variables that count the number of times that something has happened are common in the social sciences. Hausman et al. (1984) examined the effect of R&D expenditures on the number of patents received by U.S. companies; Cameron and Trivedi (1986) analyzed factors affecting how frequently a person visited the doctor; Grogger (1990) studied the deterrent effects of capital punishment on daily homicides; and King (1989b) examined the effects of alliances on the number of nations at war. Other count outcomes include derogatory reports in an individual's credit history (Greene, 1994); consumption of beverages (Mullahy, 1986); illnesses caused by pollution (Portney & Mullahy, 1986); party switching by members of the House of Representatives (King, 1988); industrial injuries (Ruser, 1991); the emergence of new companies (Hannan & Freeman, 1989, p. 230); and police arrests (Land, 1992).

e.g.

Count variables are often treated as though they are continuous and the linear regression model is applied. The use of the LRM for count outcomes can result in inefficient, inconsistent, and biased estimates. Fortunately, there are a variety of models that deal explicitly with characteristics of count outcomes. The Poisson regression model is the most basic model. With this model the probability of a count is determined by a Poisson distribution, where the mean of the distribution is a function

Conseq.

of the independent variables. This model has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance. In practice, the conditional variance often exceeds the conditional mean. Dealing with this problem leads to the negative binomial regression model, which allows the variance to exceed the mean. A second problem is that the number of 0's in a sample often exceeds the number predicted by either the Poisson or the negative binomial regression model. Zero modified count models explicitly model the number of predicted 0's, and also allow the variance to differ from the mean. A third problem is that many count variables are only observed after the first count occurs. This requires a truncated count model, corresponding to the truncated regression model of Chapter 7. Each of these models for counts is based on the Poisson distribution, which is now considered.

8.1. The Poisson Distribution

Let y be a random variable indicating the number of times that an event has occurred during an interval of time. y has a Poisson distribution with parameter $\mu > 0$ if

$$\Pr(y | \mu) = \frac{\exp(-\mu)\mu^y}{y!} \quad \text{for } y = 0, 1, 2, \dots \quad [8.1]$$

For example, if $y = 0$, then $\Pr(y = 0 | \mu) = \exp(-\mu)\mu^0/0! = \exp(-\mu)$; for $y = 1$, $\Pr(y = 1 | \mu) = \exp(-\mu)\mu^1/1! = \exp(-\mu)\mu$; and for $y = 3$, $\Pr(y = 3 | \mu) = \exp(-\mu)\mu^3/3! = \exp(-\mu)\mu^3/6$. Figure 8.1 plots the Poisson distribution for μ equal to .8, 1.5, 2.9, and 10.5, and illustrates several important properties of the Poisson distribution (see Taylor & Karlin, 1994, pp. 241–242, for proofs):

1. As μ increases, the mass of the distribution shifts to the right. Specifically,

$$E(y) = \mu$$

The parameter μ is known as the rate since it is the expected number of times that an event has occurred per unit of time. μ can also be thought of as the mean or expected count.

2. The variance equals the mean:

$$\text{Var}(y) = E(y) = \mu$$

The equality of the mean and the variance is known as equidispersion. In practice, count variables often have a variance greater than the mean, which

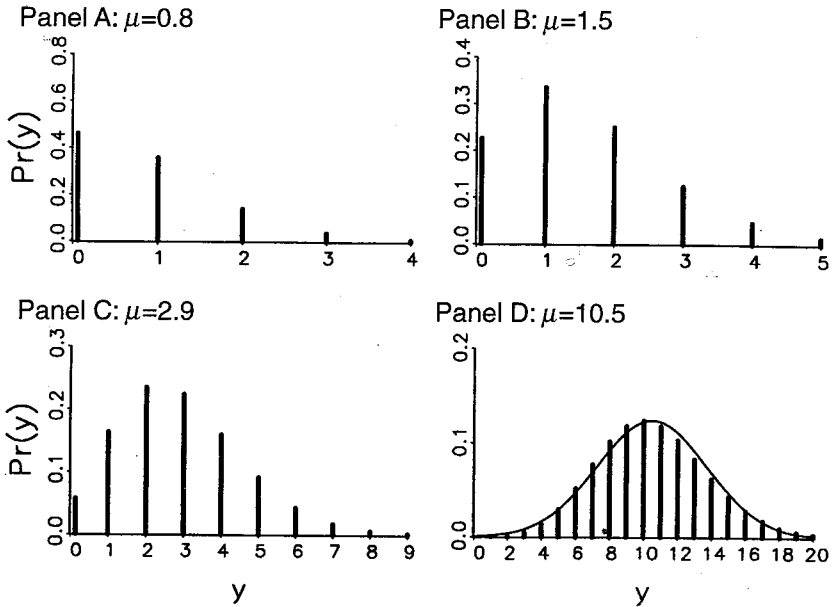


Figure 8.1. Poisson Probability Distribution

is called *overdispersion*. The development of many models for count data is an attempt to account for overdispersion.

3. As μ increases, the probability of 0's decreases. For $\mu = .8$, the probability of a 0 is .45; for $\mu = 1.5$, it is .22; for $\mu = 2.9$, it is .05; and for $\mu = 10.5$, the probability is .00002. For many count variables, there are more observed 0's than predicted by the Poisson distribution.
4. As μ increases, the Poisson distribution approximates a normal distribution. This is shown in panel D where a normal distribution with a mean and variance of 10.5 has been superimposed on the Poisson distribution.

The Poisson distribution can be derived from a simple stochastic process, known as a Poisson process, where the outcome is the number of times that something has happened (see Taylor & Karlin, 1994, pp. 252–258, for a formal derivation of the Poisson distribution). A critical assumption of a Poisson process is that events are independent. This means that when an event occurs it does not affect the probability of the event occurring in the future. For example, consider the publication of articles by scientists. The assumption of independence implies that when a sci-

entist publishes a paper, her rate of publication does not change. Past success in publishing does not affect future success.

Example of Fitting the Poisson Distribution: Article Counts

In a study of scientific productivity, Long (1990) considered factors affecting the number of papers published during graduate school by a sample of 915 biochemists. The average number of articles was 1.7 with a variance of 3.7, which indicates that there is overdispersion in the distribution of articles. The form of this overdispersion is shown in Figure 8.2. The observed proportions for each count are indicated by diamonds that are connected by a solid line. The circles show the predicted probabilities from a Poisson distribution with $\mu = 1.7$. Compared to the Poisson distribution, the observed distribution has substantially more 0's, fewer cases in the center of the distribution, and more observations in the upper tail. Overall, the sample variance is larger than would be expected if the publication of articles was governed by a Poisson process in which all scientists had the same rate of productivity. Of course, the idea that all scientists have the same rate of productivity is unrealistic, which leads us to the idea of heterogeneity.

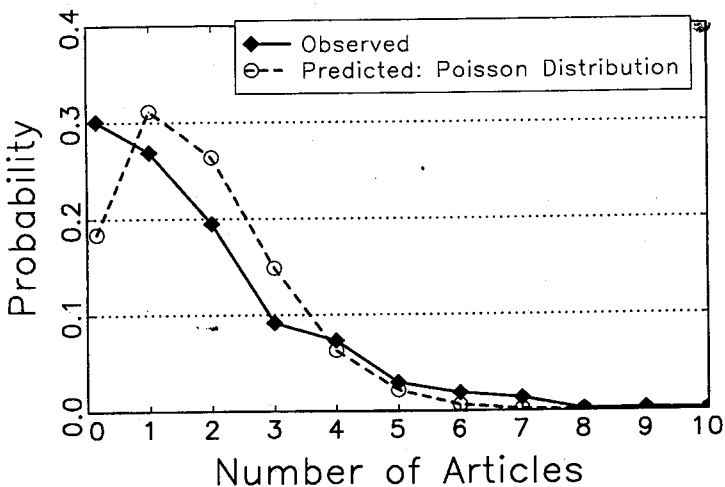


Figure 8.2. Distribution of Observed and Predicted Counts of Articles

8.1.1. The Idea of Heterogeneity

One explanation for the failure of the Poisson distribution to fit the observed data is that the rate of productivity μ differs across individuals. This is known as heterogeneity. Failure to account for heterogeneity in the rate results in overdispersion in the marginal distribution of the count. For example, suppose that the mean productivity for men is $\mu + \delta$, with a corresponding variance of $\mu + \delta$, while the mean and variance for women is $\mu - \delta$. Publications are assumed to be generated by a Poisson process in which the rate of publication differs for men and women. What will the marginal distribution look like? Assume there are equal numbers of men and women. Then the mean rate of productivity for the combined sample is the average of the rates for men and women, $\mu = [(\mu + \delta) + (\mu - \delta)]/2$, but the variance will exceed μ . (Draw the two conditional distributions and show that the marginal distribution would have a larger variance.) In general, failure to account for heterogeneity among individuals in the rate of a count variable leads to overdispersion in the marginal distribution. This result leads to the Poisson regression model which introduces heterogeneity based on observed characteristics.

8.2. The Poisson Regression Model

In the Poisson regression model, hereafter the PRM, the number of events y has a Poisson distribution with a conditional mean that depends on an individual's characteristics according to the structural model:

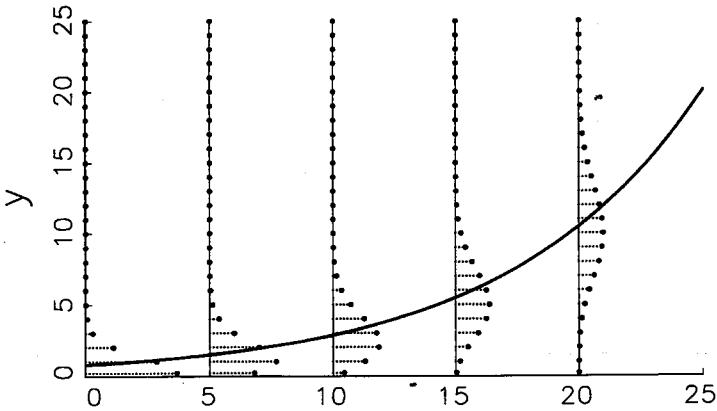
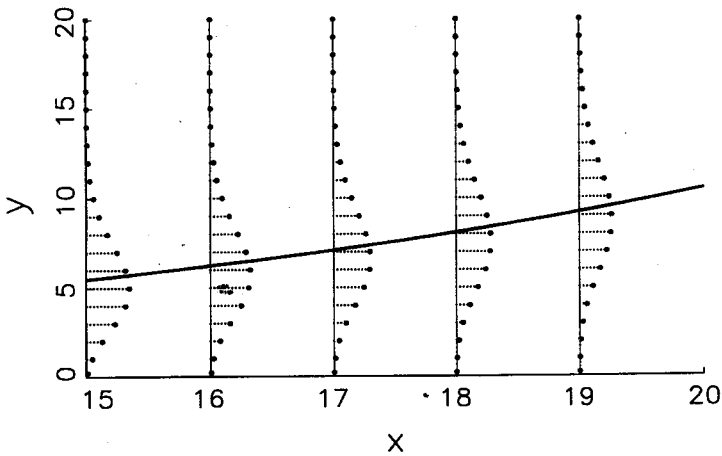
$$\mu_i = E(y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i \boldsymbol{\beta}) \quad [8.2]$$

Taking the exponential of $\mathbf{x}\boldsymbol{\beta}$ forces the expected count μ to be positive, which is required for the Poisson distribution. While other relationships between μ and the x 's are possible, such as $E(y | \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, they are rarely used.

Panel A of Figure 8.3 illustrates the PRM for a single independent variable x . The relationship $\mu = \exp(-.25 + .13x)$ is shown by a solid line. Since y is a count, it can only have nonnegative integer values. These values are represented by dotted lines which should be thought of as coming out of the page. The height of the line indicates the probability of a count given x . Specifically,

$$\Pr(y_i | \mathbf{x}_i) = \frac{\exp(-\mu_i) \mu_i^{y_i}}{y_i!}$$

*μ is conditional mean,
depends on
the relationship.
 $\mu = \exp(\mathbf{x}\boldsymbol{\beta})$
↑
estimated by
model.*

Panel A: $E(y|x)$ for $x=0$ to 25Panel B: $E(y|x)$ for $x=15$ to 20**Figure 8.3.** Distribution of Counts for the Poisson Regression Model

For example, at $x = 0$, $\mu = \exp(-.25) = .78$. Using this value of μ , the probabilities for various counts are (*Verify these values.*)

$$\begin{aligned} \Pr(y = 0 | \mu) &= .46 & \Pr(y = 1 | \mu) &= .36 \\ \Pr(y = 2 | \mu) &= .14 & \Pr(y = 3 | \mu) &= .04 \end{aligned}$$

Other probabilities can be computed similarly.

The distribution of counts around the conditional mean of y in panel A of Figure 8.3 reflects the characteristics of the Poisson distribution that were discussed using Figure 8.1. Indeed, I constructed Figure 8.3 so that the means at x equal to 0, 5, 10, and 20 correspond to the means in the earlier figure. You can see that as μ increases: (1) the conditional variance of y increases; (2) the proportion of predicted 0's decreases; and (3) the distribution around the expected value becomes approximately normal.

The figure also shows why the PRM can be thought of as a non-linear regression model with errors equal to $\varepsilon = y - E(y|x)$. While the conditional mean of ε is 0, the errors are heteroscedastic since $\text{Var}(\varepsilon|x) = E(y|x) = \exp(\mathbf{x}\beta)$. Note, however, that if your data are limited to a range of x where the relationship is approximately linear, the LRM is a reasonable approximation to the PRM. This is shown in panel B which expands that portion of the figure in panel A between $x = 15$ and $x = 20$. The relationship between μ and x is nearly linear, the errors are approximately normal, and there is only slight heteroscedasticity.

8.2.1. Estimation

The likelihood function for the PRM is

$$L(\beta | y, X) = \prod_{i=1}^N \Pr(y_i | \mu_i) = \prod_{i=1}^N \frac{\exp(-\mu_i) \mu_i^{y_i}}{y_i!} \quad [8.3]$$

where $\mu = \exp(\mathbf{x}\beta)$. After taking the log, numerical maximization can be used. The gradients and Hessian of the log likelihood are given by Maddala (1983, p. 52). Since the likelihood function is globally concave, if a maximum is found it will be unique.

8.2.2. Interpretation

The way in which you interpret a count model depends on whether you are interested in the expected value of the count variable or in the distribution of counts. If interest is in the expected count, several methods can be used to compute the change in the expectation for a change in an independent variable. If interest is in the distribution of counts or perhaps just the probability of a specific count, the probability of a count for a given level of the independent variables can be computed. Each of these methods is now considered.

Changes in the Conditional Mean

For the PRM, the expected value of y for a given \mathbf{x} is

$$\checkmark \quad \mu = E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}) \quad [8.4]$$

The change in $E(y|\mathbf{x})$ can be assessed in several ways.

Partial Change in $E(y|\mathbf{x})$. The partial derivative of $E(y|\mathbf{x})$ with respect to x_k , sometimes called the marginal effect, can be computed using the chain rule:

$$\chi \quad \frac{\partial E(y|\mathbf{x})}{\partial x_k} = \frac{\partial \exp(\mathbf{x}\boldsymbol{\beta})}{\partial \mathbf{x}\boldsymbol{\beta}} \frac{\partial \mathbf{x}\boldsymbol{\beta}}{\partial x_k} = \exp(\mathbf{x}\boldsymbol{\beta})\beta_k = E(y|\mathbf{x})\beta_k$$

Since the model is nonlinear, the value of the marginal effect depends on both the coefficient for x_k and the expected value of y given \mathbf{x} . The larger the value of $E(y|\mathbf{x})$, the larger the rate of change in $E(y|\mathbf{x})$. Further, since $E(y|\mathbf{x})$ depends on the values of all independent variables, the value of the marginal depends on the levels of all variables. Often, the marginal effect is computed with all variables held at their means.

Since the PRM and the other count models in this chapter are nonlinear, the partial derivative cannot be interpreted as the change in the expected count for a unit change in x_k . Further, the partial with respect to a dummy variables does not make sense. For these reasons, this measure of change is less informative than the factor change or discrete change.

Factor and Percentage Change in $E(y|\mathbf{x})$. The factor or percentage change in the expected count can be computed simply from the parameters of the model. To see this (Section 3.8, p. 79, provides a more detailed derivation), Equation 8.4 can be rewritten as

$$E(y|\mathbf{x}, x_k) = \exp(\beta_0) \exp(\beta_0) \exp(\beta_1 x_1) \cdots \exp(\beta_k x_k) \cdots \exp(\beta_K x_K)$$

where $E(y|\mathbf{x}, x_k)$ makes explicit the value of x_k . If x_k changes by δ ,

$$\begin{aligned} E(y|\mathbf{x}, x_k + \delta) \\ = \exp(\beta_0) \exp(\beta_1 x_1) \cdots \exp(\beta_k x_k) \exp(\beta_k \delta) \cdots \exp(\beta_K x_K) \end{aligned}$$

The factor change in the expected count for a change of δ in x_k equals

$$\begin{aligned} \frac{E(y | \mathbf{x}, x_k + \delta)}{E(y | \mathbf{x}, x_k)} &= \frac{\exp(\beta_0) \exp(\beta_1 x_1) \cdots \exp(\beta_k x_k) \exp(\beta_k \delta) \cdots \exp(\beta_K x_K)}{\exp(\beta_0) \exp(\beta_1 x_1) \cdots \exp(\beta_k x_k) \cdots \exp(\beta_K x_K)} \\ &= \exp(\beta_k \delta) \end{aligned}$$

Therefore, the parameters can be interpreted as follows:

- For a change of δ in x_k , the expected count increases by a factor of $\exp(\beta_k \times \delta)$, holding all other variables constant.

For specific values of δ :

- *Factor change.* For a unit change in x_k , the expected count changes by a factor of $\exp(\beta_k)$, holding all other variables constant.
- *Standardized factor change.* For a standard deviation change in x_k , the expected count changes by a factor of $\exp(\beta_k \times s_k)$, holding all other variables constant.

Alternatively, the percentage change in the expected count for a δ unit change in x_k , holding other variables constant, can be computed as

$$100 \times \frac{E(y | \mathbf{x}, x_k + \delta) - E(y | \mathbf{x}, x_k)}{E(y | \mathbf{x}, x_k)} = 100 \times [\exp(\beta_k \times \delta) - 1]$$

Notice that the effect of a change in x_k does not depend on the level of x_k or on the level of any other variable.

Discrete Change in $E(y | \mathbf{x})$. The effect of a variable can also be assessed by computing the discrete change in the expected value of y for a change in x_k starting at x_S and ending at x_E :

$$\frac{\Delta E(y | \mathbf{x})}{\Delta x_k} = E(y | \mathbf{x}, x_k = x_E) - E(y | \mathbf{x}, x_k = x_S) \quad [8.5]$$

This can be interpreted as:

- For a change in x_k from x_S to x_E , the expected count changes by $\Delta E(y | \mathbf{x}) / \Delta x_k$, holding all other variables constant.

As was the case in earlier chapters, the discrete change can be computed in a variety of ways, depending on your purpose:

1. The total possible effect of x_k is found by letting x_k change from its minimum to its maximum.
2. The effect of a binary variable is obtained by letting x_k change from 0 to 1.
3. The effect of a unit change in x_k is computed by changing from \bar{x}_k to $\bar{x}_k + 1$, while the centered discrete change can be computed by changing from $(\bar{x}_k - 1/2)$ to $(\bar{x}_k + 1/2)$.
4. The effect of a standard deviation change in x_k is computed by changing from \bar{x}_k to $\bar{x}_k + s_k$, while centered change is computed by changing from $(\bar{x}_k - s_k/2)$ to $(\bar{x}_k + s_k/2)$.

Unlike the factor or percentage change, the magnitude of the discrete change depends on the levels of all variables in the model.

Predicted Probabilities

The parameters can also be used to compute the probability distribution of counts for a given level of the independent variables. For a given \mathbf{x} , the probability that $y = m$ is

$$\widehat{\Pr}(y = m | \mathbf{x}) = \frac{\exp(-\widehat{\mu}) \widehat{\mu}^m}{m!} \quad [8.6]$$

where $\widehat{\mu} = \exp(\mathbf{x}\hat{\beta})$. The predicted probabilities can be computed for each observation for each count m that is of interest. Then the mean predicted probability for each count m can be used to summarize the predictions of the model:

$$\overline{\Pr}(y = m) = \frac{1}{N} \sum_{i=1}^N \widehat{\Pr}(y_i = m | \mathbf{x}_i) = \frac{1}{N} \sum_{i=1}^N \frac{\exp(-\widehat{\mu}_i) \widehat{\mu}_i^m}{m!} \quad [8.7]$$

The mean probabilities, which are computed after controlling for independent variables, can be compared to the observed proportions of the sample at each count. This is now illustrated with the data on scientific productivity.

TABLE 8.1 Descriptive Statistics for the Doctoral Publications Example

Name	Mean	Standard Deviation	Minimum	Maximum	Description
ART	1.69	1.93	0.00	19.00	Articles during last 3 years of Ph.D.
LnART	0.44	0.86	-0.69	2.97	Log of (ART + .5)
FEM	0.46	0.50	0.00	1.00	1 if female scientist; else 0
MAR	0.66	0.47	0.00	1.00	1 if married; else 0
KID5	0.50	0.76	0.00	3.00	Number of children 5 or younger
PHD	3.10	0.98	0.76	4.62	Prestige of Ph.D. department
MENT	8.77	9.48	0.00	77.00	Articles by mentor during last 3 years

NOTE: $N = 915$.*Example of the Poisson Regression Model: Article Counts*

The failure of the univariate Poisson distribution to account for the distribution of article counts could be due to heterogeneity in the characteristics of the scientists. If scientists who differ in their rate of productivity are combined, the univariate distribution of articles will be overdispersed. Research by Long (1990) suggests that gender, marital status, number of young children, prestige of the graduate program, and the number of articles written by a scientist's mentor could affect a scientist's level of publication. Table 8.1 contains descriptive statistics for these variables. Table 8.2 presents estimates from the PRM and the negative binomial regression model (NBRM) that is considered in Section 8.3.

For purposes of comparison, I have also included results from the LRM. By taking the log of Equation 8.2, the PRM can be written as the log-linear model:

$$\ln \mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

This suggests that the PRM can be approximated by the LRM:

$$\ln y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

However, since $\ln(0)$ is undefined, it is necessary to add a positive constant c to y before taking the log. Values of c equal to .5 or .01 are frequently used. This suggests the regression model:

$$\ln(y_i + c) = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

King (1988) demonstrates that estimating this model results in biased estimates of the parameters of the corresponding PRM. However, as

TABLE 8.2 Linear Regression, Poisson Regression, and Negative Binomial Regression of Doctoral Publications

Variable		LRM of LnART	PRM of ART	NBRM of ART
Constant	β	0.178	0.305	0.256
	t/z	1.65	2.96	1.82
FEM	β	-0.135	-0.225	-0.216
	t/z	-2.35	-4.11	-2.82
MAR	β	0.133	0.155	0.150
	t/z	2.04	2.53	1.79
KID5	β	-0.133	-0.185	-0.176
	β^{S_x}	-0.102	-0.141	-0.135
	t/z	-3.28	-4.61	-3.28
PHD	β	0.026	0.013	0.015
	β^{S_x}	0.025	0.013	0.015
	t/z	0.90	0.49	0.42
MENT	β	0.025	0.026	0.029
	β^{S_x}	0.241	0.242	0.276
	t/z	8.61	12.73	9.10
Dispersion	α	—	—	0.442
	z	—	—	8.45
Pr($y = 0$)		—	0.21	0.34
$-2 \ln L$		2215.32	3302.11	3121.92

NOTE: $N = 915$. β is an unstandardized coefficient; β^{S_x} is an x -standardized coefficient; t/z is a t -test of β for LRM and a z -test of β for the PRM and NBRM.

illustrated in Table 8.2, the estimates from the LRM can be of roughly the same size and significance as the coefficients from the PRM. This is more likely to be true when large counts are frequent.

The simplest way to interpret the results of the PRM is by using the factor changes in the expected count. For example, the coefficient for *FEM* can be interpreted as:

- Being a female scientist decreases the expected number of articles by a factor of .80 ($= \exp[-.225]$), holding all other variables constant.

Or, equivalently,

- Being a female scientist decreases the expected number of articles by 20% ($= 100[\exp(-.225) - 1]$), holding all other variables constant.

Similarly, the effect of the mentor's productivity can be interpreted as:

- For every additional article by the mentor, a scientist's mean productivity increases by 2.6%, holding all other variables constant.

The standardized coefficient can be interpreted as:

- For a standard deviation increase in the mentor's productivity, a scientist's mean productivity increases by 27%, holding all other variables constant. ✓

(Verify these numbers.)

The results just given refer to the multiplicative factor change in the expected count. It can also be informative to examine the additive change in the expected count. For example, the change in the expected count for a change in *FEM* from 0 to 1 can be computed using Equation 8.5. First, hold all variables at their means except for *FEM*. If *FEM* is 1, indicating a female scientist, the expected productivity is 1.43; if *FEM* is 0, indicating a male scientist, the expected productivity is 1.79. Therefore, we conclude:

- Being a female scientist decreases the expected productivity by .36 articles, holding all other variables at their means. ✓

Notice that the change of .36 articles from 1.79 to 1.43 corresponds to the 20% decrease that was computed using the measure of percentage change. (Verify these values using Table 8.1.)

The results of the PRM can also be interpreted in terms of predicted probabilities using Equation 8.7. In Figure 8.4, the observed proportions are shown by solid diamonds connected by solid lines. The mean pre-

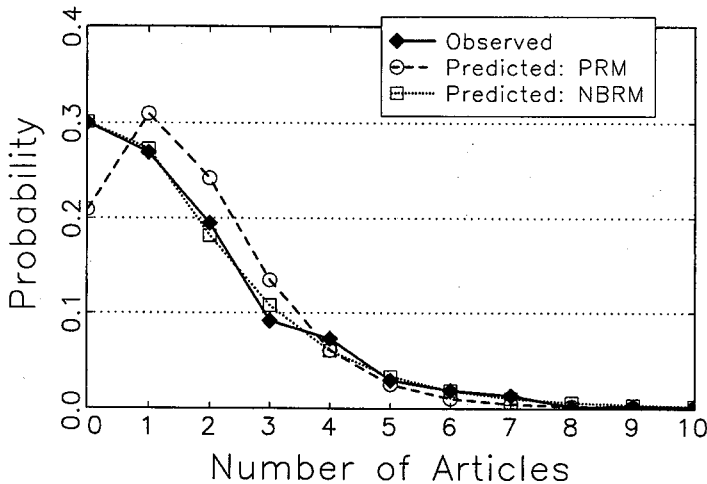


Figure 8.4. Comparisons of the Mean Predicted Probabilities From the Poisson and Negative Binomial Regression Models

dicted probabilities for the PRM are shown with open circles connected by dashed lines. (The predictions from the negative binomial regression, shown with squares, are discussed in the next section.) While the predictions from the PRM are somewhat better than those from the univariate Poisson distribution in Figure 8.2, the PRM still underpredicts 0's, overpredicts counts from 1 to 3, and slightly underpredicts counts in the upper tail. Large differences between the mean probabilities and the observed proportions suggest that a model is inappropriate. However, since an incorrect model can provide predictions that are close to the observed data, a close match is *not* clear evidence that a model is appropriate.

8.3. The Negative Binomial Regression Model

The Poisson regression model rarely fits in practice since in most applications the conditional variance is greater than the conditional mean. If the mean structure is correct, but there is overdispersion, the estimates from the PRM are consistent, but inefficient (Gourieroux et al., 1984). Further, the standard errors from the PRM will be biased downward, resulting in spuriously large z-values (Cameron & Trivedi, 1986, p. 31). For example, if there is overdispersion in the data analyzed in Table 8.2 (which later is shown to be the case), the z-tests may over estimate the significance of the variables.

A useful way to understand the limitation imposed by constraining the conditional variance to equal the conditional mean is to compare the PRM to the LRM. In the LRM, y given x is conditionally distributed with variance σ^2 , where σ^2 is estimated along with the β 's. Even though σ^2 is not of substantive interest, it allows the variance of the errors to be determined independently of the β 's. In the PRM, y has a conditional Poisson distribution with a variance that is a function of the x 's and the β 's: $\text{Var}(y|x) = \exp(x\beta)$. Our first extension of the PRM adds a parameter that allows the conditional variance of y to exceed the conditional mean. This is the negative binomial regression model, hereafter the NBRM. While the NBRM can be derived in several ways, I consider the most common motivation of the model in terms of unobserved heterogeneity. ✱

In the PRM, the conditional mean of y given x is known; $\mu = \exp(x\beta)$. In the NBRM, the mean μ is replaced with the random variable $\tilde{\mu}$:

$$\tilde{\mu}_i = \exp(x_i\beta + \varepsilon_i) \quad [8.8]$$

ε is a random error that is assumed to be uncorrelated with \mathbf{x} . You can think of ε either as the combined effects of unobserved variables that have been omitted from the model (Gourieroux et al., 1984) or as another source of pure randomness (Hausman et al., 1984). In the PRM, variation in μ is introduced through observed heterogeneity. Different values of \mathbf{x} result in different values of μ , but all individuals with the same \mathbf{x} have the same μ . In the NBRM, variation in $\tilde{\mu}$ is due both to variation in \mathbf{x} among individuals but also to unobserved heterogeneity introduced by ε . For a given combination of values for the independent variables, there is a distribution of $\tilde{\mu}$'s rather than a single μ .

The relationship between $\tilde{\mu}$ and our original μ follows readily:

$$\tilde{\mu}_i = \exp(\mathbf{x}_i \boldsymbol{\beta}) \exp(\varepsilon_i) = \mu_i \exp(\varepsilon_i) = \mu_i \delta_i \quad \checkmark$$

where δ_i is defined to equal $\exp(\varepsilon_i)$. Recall that the LRM was not identified until an assumption was made about the mean of the error (see Section 2.5.1). For similar reasons, the NBRM is not identified without an assumption about the mean of the error term. The most convenient assumption is that

$$E(\delta_i) = 1 \quad \text{or} \quad \varepsilon_i = \ln(\delta_i) = 0 \quad [8.9]$$

This assumption implies that the expected count after adding the new source of variation is the same as it was for the PRM:

$$E(\tilde{\mu}_i) = E(\mu_i \delta_i) = \mu_i E(\delta_i) = \mu_i$$

The distribution of observations given \mathbf{x} and δ is still Poisson:

$$\Pr(y_i | \mathbf{x}_i, \delta_i) = \frac{\exp(-\tilde{\mu}_i) \tilde{\mu}_i^{y_i}}{y_i!} = \frac{\exp(-\mu_i \delta_i) (\mu_i \delta_i)^{y_i}}{y_i!} \quad [8.10]$$

However, since δ is unknown we cannot compute $\Pr(y | \mathbf{x}, \delta)$ and instead need to compute the distribution of y given only \mathbf{x} . To compute $\Pr(y | \mathbf{x})$ without conditioning on δ , we average $\Pr(y | \mathbf{x}, \delta)$ by the probability of each value of δ . If g is the pdf for δ , then

$$\Pr(y_i | \mathbf{x}_i) = \int_0^\infty [\Pr(y_i | \mathbf{x}_i, \delta_i) \times g(\delta_i)] d\delta_i \quad [8.11]$$

To clarify what this important equation is doing, assume that δ has only two values, d_1 and d_2 . The counterpart to Equation 8.11 is

$$\begin{aligned} \Pr(y_i | \mathbf{x}_i) &= [\Pr(y_i | \mathbf{x}_i, \delta_i = d_1) \times \Pr(\delta_i = d_1)] \\ &\quad + [\Pr(y_i | \mathbf{x}_i, \delta_i = d_2) \times \Pr(\delta_i = d_2)] \end{aligned} \quad [8.12]$$

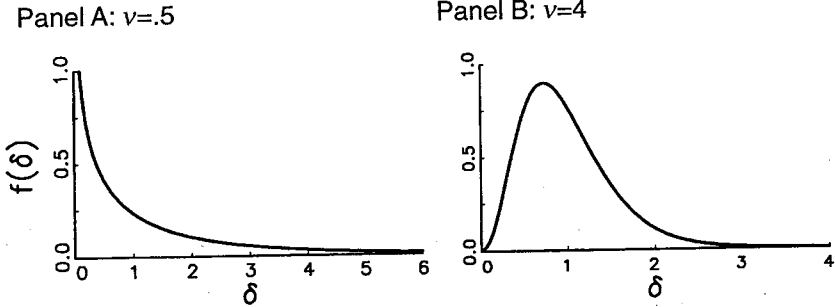


Figure 8.5. Probability Density Function for the Gamma Distribution

This equation weights $\Pr(y | \mathbf{x}, \delta)$ by $\Pr(\delta)$, and adds over all values of δ . Thus, Equation 8.12 computes the probability of y as a mixture of two probability distributions.

To solve Equation 8.11, we must specify the form of the pdf for δ . While several distributions have been considered, the most common assumption is that δ_i has a gamma distribution with parameter ν_i :

$$g(\delta_i) = \frac{\nu_i^{\nu_i}}{\Gamma(\nu_i)} \delta_i^{\nu_i-1} \exp(-\delta_i \nu_i) \quad \text{for } \nu_i > 0 \quad [8.13]$$

where the gamma function is defined as $\Gamma(\nu) = \int_0^\infty t^{\nu-1} e^{-t} dt$. It can be shown (Johnson et al., 1994, pp. 337–342) that if δ_i has a gamma distribution, then $E(\delta_i) = 1$, as required by Equation 8.9, and $\text{Var}(\delta_i) = 1/\nu_i$. The parameter ν also affects the shape of the distribution, as shown in Figure 8.5. As ν increases, the distribution becomes increasingly bell shaped and centered around 1.

The negative binomial, hereafter NB, probability distribution is obtained by solving Equation 8.11 using Equation 8.10 for $\Pr(y | \mathbf{x}, \delta)$ and Equation 8.13 for $g(\delta)$ (see Cameron & Trivedi, 1996, for details):

$$\Pr(y_i | \mathbf{x}_i) = \frac{\Gamma(y_i + \nu_i)}{y_i! \Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i + \mu_i} \right)^{\nu_i} \left(\frac{\mu_i}{\nu_i + \mu_i} \right)^{y_i}$$

The expected value of y for the NB distribution is the same as for the Poisson distribution:

$$E(y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i \mathbf{\beta}) = \mu_i \quad [8.14]$$

Typo?

but the conditional variance differs:

$$\text{Var}(y_i | \mathbf{x}) = \mu_i \left(1 + \frac{\mu_i}{\nu_i} \right) = \exp(\mathbf{x}_i \boldsymbol{\beta}) \left(1 + \frac{\exp(\mathbf{x}_i \boldsymbol{\beta})}{\nu_i} \right) \quad [8.15]$$

Since μ and ν are positive, the conditional variance of y in the NBRM must exceed the conditional mean $\exp(\mathbf{x}\boldsymbol{\beta})$. (What must happen to ν to reduce the variance to that of the PRM?)

The larger conditional variance in y increases the relative frequency of low and high counts. This is seen in Figure 8.6 where the Poisson and NB distributions are compared for means of 1 and 10. The NB distribution corrects a number of sources of poor fit that are often found when the Poisson distribution is used. First, the variance of the NB distribution exceeds the variance of the Poisson distribution for a given mean. Second, the increased variance in the NBRM results in substantially larger probabilities for small counts. In panel A, the probability of a zero count increases from .37 for the Poisson distribution to .50, .77, and .85 as the variance of the NB distribution increases. Finally, there are slightly larger probabilities for larger counts in the NB distribution.

While the mean structure is fully specified by Equation 8.14, the variance is unidentified in Equation 8.15. The problem is that if ν varies by individual, then there are more parameters than observations. The most common identifying assumption is that ν is the same for all individuals:

$$\checkmark \quad \nu_i = \alpha^{-1} \quad \text{for } \alpha > 0$$

This assumption simply states that the variance of δ is constant. (We set the variance to α^{-1} rather than α to simplify the formulas that follow.) α is known as the dispersion parameter since increasing α increases the conditional variance of y . This is seen by substituting $\nu = \alpha^{-1}$ into Equation 8.15:

$$\text{Var}(y_i | \mathbf{x}) = \mu_i \left(1 + \frac{\mu_i}{\alpha^{-1}} \right) = \mu_i (1 + \alpha \mu_i) = \mu_i + \alpha \mu_i^2 \quad [8.16]$$

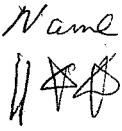
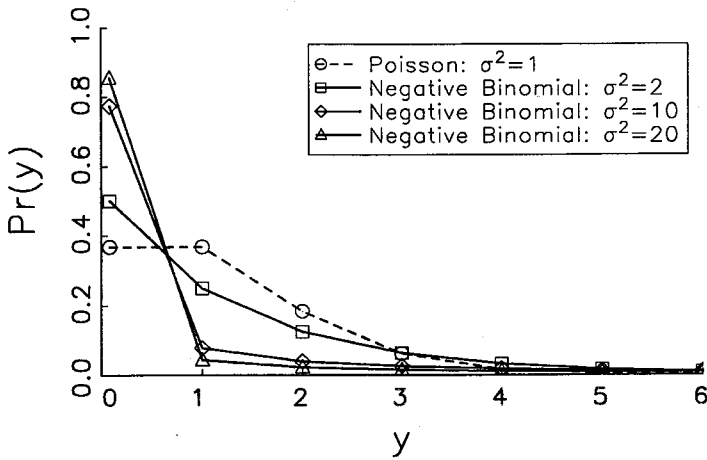
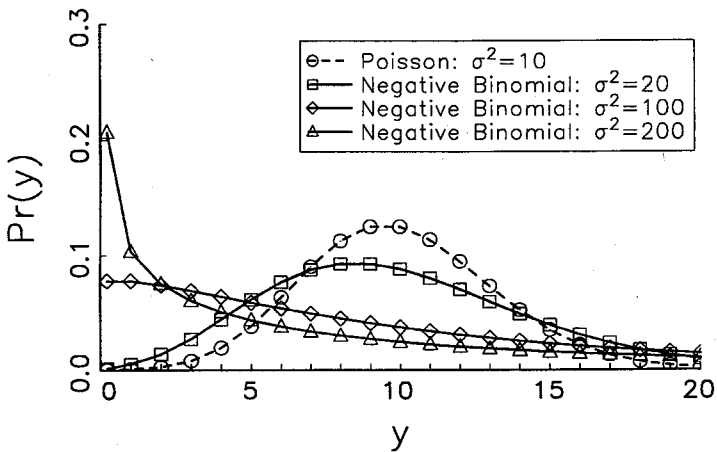
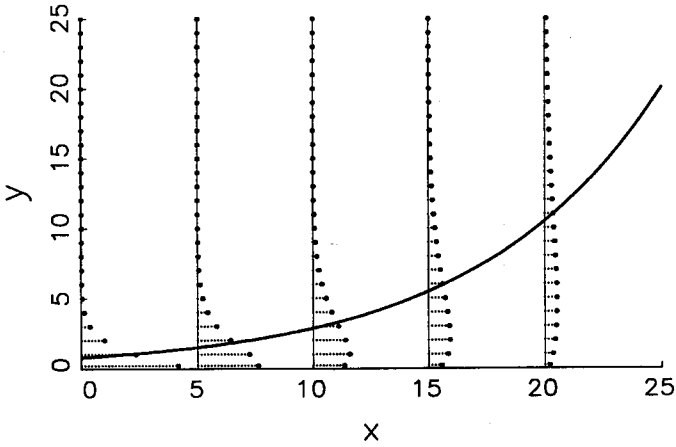
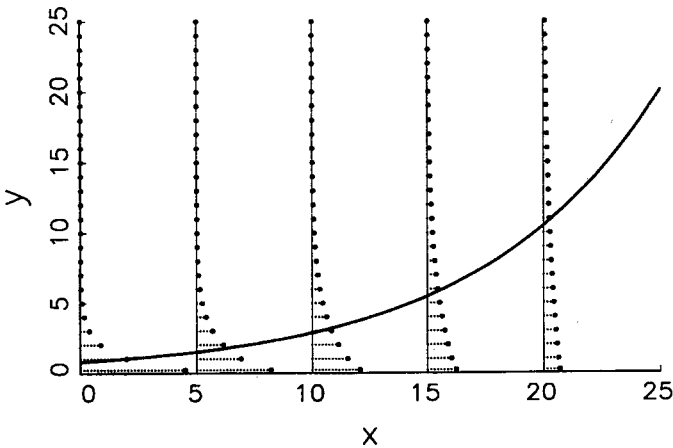
Pos (What happens if $\alpha = 0$?) Under this specification of ν , the conditional variance is quadratic in the mean, which has led Cameron and Trivedi (1986) to call this the Negbin 2 model. Name 

Figure 8.7 shows the effect of the added variation in the NBRM. While the mean structure is identical to that used to illustrate the PRM, namely $E(y | \mathbf{x}) = \exp(-.25 + .13x)$, the distribution around the mean differs. In panel A, $\alpha = .5$ and the difference from the PRM is subtle. Compared

Panel A: $E(y)=1$ Panel B: $E(y)=10$ **Figure 8.6.** Comparisons of the Negative Binomial and Poisson Distributions

to Figure 8.3, the differences are barely noticeable at $x = 0$, but can be clearly seen for larger values of x . When α is increased to 1 in panel B, the effects are more pronounced. Note, for example, that the conditional mode for all values of x is now 0 and that the errors no longer appear normal as μ increases.

Panel A: NBRM with $\alpha=0.5$ Panel B: NBRM with $\alpha=1.0$ **Figure 8.7.** Distribution of Counts for the Negative Binomial Regression Model**8.3.1. Heterogeneity and Contagion**

The NB distribution can be derived in a variety of ways as shown by Feller (1971, pp. 57–58) and Johnson et al. (1992, pp. 203–207). The derivation used above is based on unobserved heterogeneity, which is represented by the error ε in Equation 8.8. This derivation dates to

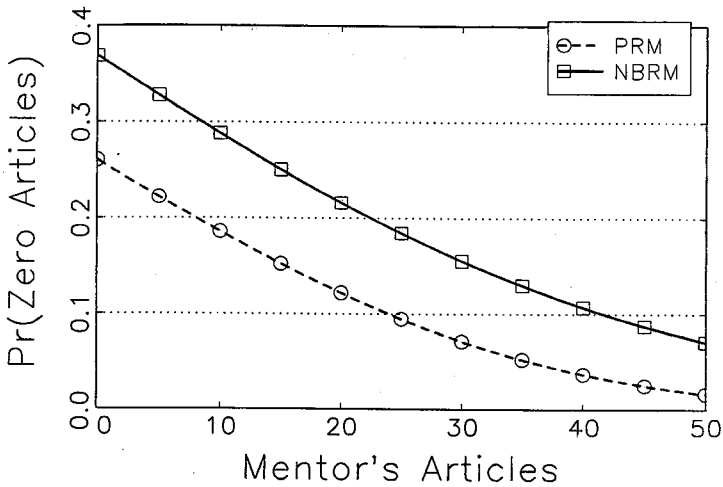


Figure 8.8. Probability of 0's From the Poisson and Negative Binomial Regression Models

publications, the higher proportion of predicted zeros for the NBRM is offset by the higher proportion of larger counts that are also predicted by this model.

8.3.5. Related Models

The NBRM is one of a class of models constructed by mixing the Poisson distribution with a second distribution using Equation 8.11. The mixture of the Poisson and gamma distributions is particularly convenient given the closed form of the resulting negative binomial distribution. In addition to the Negbin 2 models considered above, Cameron and Trivedi (1986) suggest a Negbin k model in which $\text{Var}(y | \mathbf{x}) = \mu + \alpha\mu^{2-k}$. If $k = 1$, then $\text{Var}(y | \mathbf{x}) = \mu + \alpha\mu$ which corresponds to replacing the assumption $\nu = \alpha^{-1}$ of the Negbin 2 model with $\nu = \mu/\alpha$. This is known as the Negbin 1 model. Other distributions and mixtures can also be used. Hinde (1982) considered a Poisson and normal mixture; Dean et al. (1989) used a Poisson and inverse Gaussian mixture. King (1989a) proposed a generalized event count (GEC) model that allows for either overdispersion or underdispersion. See Winkelmann (1994, pp. 112–120) for further details.