Part III

Count and Discrete Dependent Variable Models

10

Count Data Models

Count data consist of non-negative integer values and are encountered frequently in the modeling of transportation-related phenomena. Examples of count data variables in transportation include the number of driver route changes per week, the number of trip departure changes per week, drivers' frequency of use of intelligent transportation systems (ITS) technologies over some time period, number of vehicles waiting in a queue, and the number of accidents observed on road segments per year.

A common mistake is to model count data as continuous data by applying standard least squares regression. This is not correct because regression models yield predicted values that are non-integers and can also predict values that are negative, both of which are inconsistent with count data. These limitations make standard regression analysis inappropriate for modeling count data without modifying dependent variables.

Count data are properly modeled by using a number of methods, the most popular of which are Poisson and negative binomial regression models. Poisson regression is the more popular of the two, and is applied to a wide range of transportation count data. The Poisson distribution approximates rare-event count data, such as accident occurrence, failures in manufacturing or processing, and number of vehicles waiting in a queue. One requirement of the Poisson distribution is that the mean of the count process equals its variance. When the variance is significantly larger than the mean, the data are said to be overdispersed. There are numerous reasons for overdispersion, some of which are discussed later in this chapter. In many cases, overdispersed count data are successfully modeled using a negative binomial model.

10.1 Poisson Regression Model

To help illustrate the principal elements of a Poisson regression model, consider the number of accidents occurring per year at various intersections

in a city. In a Poisson regression model, the probability of intersection i having y_i accidents per year (where y_i is a non-negative integer) is given by

$$P(y_i) = \frac{EXP(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

where $P(y_i)$ is the probability of intersection i having y_i accidents per year and λ_i is the Poisson parameter for intersection i, which is equal to the expected number of accidents per year at intersection i, $E[y_i]$. Poisson regression models are estimated by specifying the Poisson parameter λ_i (the expected number of events per period) as a function of explanatory variables. For the intersection accident example, explanatory variables might include the geometric conditions of the intersections, signalization, pavement types, visibility, and so on. The most common relationship between explanatory variables and the Poisson parameter is the log-linear model,

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i) \text{ or, equivalently } LN(\lambda_i) = \boldsymbol{\beta}\mathbf{X}_i,$$

where $\mathbf{X}_i$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimable parameters. In this formulation, the expected number of events per period is given by $E[y_i] = \lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i)$. This model is estimable by standard maximum likelihood methods, with the likelihood function given as

$$L(\boldsymbol{\beta}) = \prod_i \frac{EXP\left[-EXP(\boldsymbol{\beta}\mathbf{X}_i)\right]\left[EXP(\boldsymbol{\beta}\mathbf{X}_i)\right]^{y_i}}{y_i!}. \tag{10.3}$$

The log of the likelihood function is simpler to manipulate and more appropriate for estimation,

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[-EXP(\boldsymbol{\beta}\mathbf{X}_i) + y_i\boldsymbol{\beta}\mathbf{X}_i - LN(y_i!)\right] \tag{10.4}$$

As with most statistical models, the estimated parameters are used to make inferences about the unknown population characteristics thought to impact the count process. Maximum likelihood estimates produce Poisson parameters that are consistent, asymptotically normal, and asymptotically efficient.

To provide some insight into the implications of parameter estimation results, elasticities are computed to determine the marginal effects of the independent variables. Elasticities provide an estimate of the impact of a variable on the expected frequency and are interpreted as the effect of a 1% change in the variable on the expected frequency λ_i. For example, an elasticity of –1.32

is interpreted to mean that a 1% increase in the variable reduces the expected frequency by 1.32%. Elasticities are the correct way of evaluating the relative impact of each variable in the model. Elasticity of frequency λ_i is defined as

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik}, \tag{10.5}$$

where E represents the elasticity, x_{ik} is the value of the kth independent variable for observation i, β_k is the estimated parameter for the kth independent variable and λ_i is the expected frequency for observation i. Note that elasticities are computed for each observation i. It is common to report a single elasticity as the average elasticity over all i.

The elasticity in Equation 10.5 is only appropriate for continuous variables such as highway lane width, distance from outside shoulder edge to roadside features, and vertical curve length. It is not valid for noncontinuous variables such as indicator variables that take on values of 0 or 1. For indicator variables, a pseudo-elasticity is computed to estimate an approximate elasticity of the variables. The pseudo-elasticity gives the incremental change in frequency caused by changes in the indicator variables. The pseudo-elasticity, for indicator variables, is computed as

$$E_{x_{ik}}^{\lambda_i} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)}. \tag{10.6}$$

Also, note that the Poisson probabilities for observation i are calculated using the following recursive formulas:

$$\begin{aligned} P_{0,i} &= EXP(-\lambda_i) \\ P_{j,i} &= \left(\frac{\lambda_i}{i}\right) P_{j-1,i}, \; j = 1, 2, 3, \ldots; i = 1, 2, \ldots n, \end{aligned} \tag{10.7}$$

where $P_{0,i}$ is the probability that observation i experiences 0 events in the specified observation period.

10.2 Poisson Regression Model Goodness-of-Fit Measures

There are numerous goodness-of-fit (GOF) statistics used to assess the fit of the Poisson regression model to observed data. As mentioned in previous chapters, when selecting among alternative models, GOF statistics should be considered along with model plausibility and agreement with expectations.

The likelihood ratio test is a common test used to assess two competing models. It provides evidence in support of one model, usually a full or complete model, over another competing model that is restricted by having a reduced number of model parameters. The likelihood ratio test statistic is

$$X^2 = -2[LL(\boldsymbol{\beta}_R) - LL(\boldsymbol{\beta}_U)], \tag{10.8}$$

where $LL(\boldsymbol{\beta}_R)$ is the log likelihood at convergence of the "restricted" model (sometimes considered to have all parameters in $\boldsymbol{\beta}$ equal to 0, or just to include the constant term, to test overall fit of the model), and $LL(\boldsymbol{\beta}_U)$ is the log likelihood at convergence of the unrestricted model. The X^2 statistic is χ^2 distributed with the degrees of freedom equal to the difference in the numbers of parameters in the restricted and unrestricted model (the difference in the number of parameters in the $\boldsymbol{\beta}_R$ and the $\boldsymbol{\beta}_U$ parameter vectors).

The sum of model deviances, G^2, is equal to zero for a model with perfect fit. Note, however, that because observed y_i is an integer while the predicted expected value $\hat{\lambda}_i$ is continuous, a G^2 equal to zero is a theoretical lower bound. This statistic is given as

$$G^2 = 2\sum_{i=1}^{n} y_i LN\left(\frac{y_i}{\hat{\lambda}_i}\right). \tag{10.9}$$

An equivalent measure to R^2 in ordinary least squares linear regression is not available for a Poisson regression model due to the nonlinearity of the conditional mean ($E[y \mid \mathbf{X}]$) and heteroscedasticity in the regression. A similar statistic is based on standardized residuals,

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}\right]^2}{\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]^2}, \tag{10.10}$$

where the numerator is similar to a sum of square errors and the denominator is similar to a total sum of squares.

Another measure of overall model fit is the ρ^2 statistic. The ρ^2 statistic is

$$\rho^2 = 1 - \frac{LL(\boldsymbol{\beta})}{LL(\mathbf{0})}, \tag{10.11}$$

where $LL(\boldsymbol{\beta})$ is the log likelihood at convergence with parameter vector $\boldsymbol{\beta}$ and $LL(\mathbf{0})$ is the initial log likelihood (with all parameters set to zero). The

perfect model would have a likelihood function equal to one (all selected alternative outcomes would be predicted by the model with probability one, and the product of these across the observations would also be one) and the log likelihood would be zero, giving a ρ^2 of one (see Equation 10.11). Thus the ρ^2 statistic is between zero and one and the closer it is to one, the more variance the estimated model is explaining. Additional GOF measures for the Poisson regression model are found in Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (1995b).

Example 10.1

Accident data from California (1993 to 1998) and Michigan (1993 to 1997) were collected (Vogt and Bared, 1998; Vogt, 1999). The data represent a culled data set from the original studies, which included data from four states across numerous time periods and over five different intersection types. A reduced set of explanatory variables is used for injury accidents on three-legged stop-controlled intersections with two lanes on the minor and four lanes on the major road. The accident data are thought to be approximately Poisson or negative binomial distributed, as suggested by previous studies on the subject (Miaou and Lum, 1993; Miaou 1994; Shankar et al., 1995; Poch and Mannering, 1996; Milton and Mannering, 1998; and Harwood et al., 2000). The variables in the study are summarized in Table 10.1.

Table 10.2 shows the parameter estimates of a Poisson regression estimated on the accident data. This model contains a constant and four

TABLE 10.1

Summary of Variables in California and Michigan Accident Data

Variable Abbreviation	Variable Description	Maximum/ Minimum Values	Mean of Observations	Standard Deviation of Observations
STATE	Indicator variable for state: 0 = California; 1 = Michigan	1/0	0.29	0.45
ACCIDENT	Count of injury accidents over observation period	13/0	2.62	3.36
AADT1	Average annual daily traffic on major road	33058/2367	12870	6798
AADT2	Average annual daily traffic on minor road	3001/15	596	679
MEDIAN	Median width on major road in feet	36/0	3.74	6.06
DRIVE	Number of driveways within 250 ft of intersection center	15/0	3.10	3.90

TABLE 10.2

Poisson Regression of Injury Accident Data

Independent Variable	Estimated Parameter	*t* Statistic
Constant	–0.826	–3.57
Average annual daily traffic on major road	0.0000812	6.90
Average annual daily traffic on minor road	0.000550	7.38
Median width in feet	– 0.0600	– 2.73
Number of driveways within 250 ft of intersection	0.0748	4.54
Number of observations	84	
Restricted log likelihood (constant term only)	–246.18	
Log likelihood at convergence	–169.25	
Chi-squared (and associated *p*-value)	153.85 (<0.0000001)	
R_p-squared	0.4792	
G^2	176.5	

variables: two average annual daily traffic (*AADT*) variables, median width, and number of driveways. The mainline *AADT* appears to have a smaller influence than the minor road *AADT*, contrary to what is expected. Also, as median width increases, accidents decrease. Finally, the number of driveways close to the intersection increases the number of intersection injury accidents. The signs of the estimated parameters are in line with expectation.

The mathematical expression for this Poisson regression model is as follows:

$$E[y_i] = \lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i)$$

$$= EXP\begin{pmatrix} -0.83 + 0.00008(AADT1_i) \\ +0.0005(AADT2_i) - 0.06(MEDIAN_i) + 0.07(DRIVE_i) \end{pmatrix}.$$

The model parameters are additive in the exponent or multiplicative on the expected value of y_i. As in a linear regression model, standard errors of the estimated parameters are provided, along with approximate *t*-values, and *p*-values associated with the null hypothesis of zero effect. In this case, the results show all estimated model parameters to be statistically significant beyond the 0.01 level of significance.

Example 10.2

Inspection of the output shown in Table 10.2 reveals numerous properties of the fitted model. The value of the log-likelihood function for the fitted model is –169.25, whereas the restricted log likelihood is –246.18. The restricted or reduced log likelihood is associated with a model with the constant term only. A likelihood ratio test comparing the fitted model

TABLE 10.3

Average Elasticities of the Poisson Regression Model Shown in Table 10.2

Independent Variable	Elasticity
Average annual daily traffic on major road	1.045
Average annual daily traffic on minor road	0.327
Median width in feet	–0.228
Number of driveways within 250 ft of intersection	0.232

and the reduced model results in $X^2 = 153.85$, which is sufficient to reject the fit of the reduced model. Thus it is very unlikely (p-value less than 0.0000001) that randomness alone would produce the observed decrease in the log likelihood function.

Note that G^2 is 186.48, which is only relevant in comparison to other competing models. The R^2_p is 0.48, which again serves as a comparison to competing models. A model with higher X^2, lower G^2, and a higher R^2_p is sought in addition to a more appealing set of individual predictor variables, model specification, and agreement with expectation and theory.

Table 10.3 shows the average elasticities computed for the variables in the model shown in Table 10.2. These elasticities are obtained by enumerating through the sample (applying Equation 10.5 to all observations) and computing the average of these elasticities.

Poisson regression is a powerful analysis tool, but, as with all statistical methods, it is used inappropriately if its limitations are not fully understood. There are three common analysis errors (Lee and Mannering, 2002). The first is failure to recognize that data are truncated. The second is the failure to recognize that the mean and variance are not equal, as required by the Poisson distribution. The third is the failure to recognize that the data contain a preponderance of zeros. These limitations and their remedies are now discussed.

10.3 Truncated Poisson Regression Model

Truncation of data can occur in the routine collection of transportation data. For example, if the number of times per week an in-vehicle navigation system is used on the morning commute to work, during weekdays, the data are right truncated at 5, which is the maximum number of uses in any given week. Estimating a Poisson regression model without accounting for this truncation will result in biased estimates of the parameter vector $\boldsymbol{\beta}$, and

erroneous inferences will be drawn. Fortunately, the Poisson model is adapted easily to account for such truncation. The right-truncated Poisson model is written as (see Johnson and Kotz, 1969)

$$P(y_i) = \left[\lambda_i^{y_i}/y_i!\right] \Bigg/ \left[\sum_{m_i=0}^{r}\left(\lambda_i^{m_i}/m_i!\right)\right], \tag{10.12}$$

where $P(y_i)$ is the probability of commuter i using the system y_i times per week, λ_i is the Poisson parameter for commuter i; m_i is the number of uses per week; and r is the right truncation (in this case, 5 times per week). A driver decision-making example of a right-truncated Poisson regression is provided by Mannering and Hamed (1990*a*) in their study of weekly departure delays.

10.4 Negative Binomial Regression Model

A common analysis error is a result of failing to satisfy the property of the Poisson distribution that restricts the mean and variance to be equal, when $E[y_i] = VAR[y_i]$. If this equality does not hold, the data are said to be under dispersed ($E[y_i] > VAR[y_i]$) or overdispersed ($E[y_i] < VAR[y_i]$), and the parameter vector is biased if corrective measures are not taken. Overdispersion can arise for a variety of reasons, depending on the phenomenon under investigation (for additional discussion and examples, see Karlaftis and Tarko, 1998). The primary reason in many studies is that variables influencing the Poisson rate across observations have been omitted from the regression.

The negative binomial model is derived by rewriting Equation 10.2 such that, for each observation i,

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i), \tag{10.13}$$

where $EXP(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α^2. The addition of this term allows the variance to differ from the mean as below:

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2. \tag{10.14}$$

The Poisson regression model is regarded as a limiting model of the negative binomial regression model as α approaches zero, which means that the selection between these two models is dependent on the value of α. The parameter α is often referred to as the overdispersion parameter. The negative binomial distribution has the form:

$$P(y_i) = \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha) + \lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{(1/\alpha) + \lambda_i}\right)^{y_i}, \tag{10.15}$$

where Γ(.) is a gamma function. This results in the likelihood function:

$$L(\lambda_i) = \prod_i \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha) + \lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{(1/\alpha) + \lambda_i}\right)^{y_i}. \tag{10.16}$$

When the data are overdispersed, the estimated variance term is larger than under a true Poisson process. As overdispersion becomes larger, so does the estimated variance, and consequently all of the standard errors of parameter estimates become inflated.

A test for overdispersion is provided by Cameron and Trivedi (1990) based on the assumption that under the Poisson model, $(y_i - E[y_i])^2 - E[y_i]$ has mean zero, where $E[y_i]$ is the predicted count $\hat{\lambda}_i$. Thus, null and alternative hypotheses are generated, such that

$$\begin{aligned} H_0&: \ VAR[y_i] = E[y_i] \\ H_A&: \ VAR[y_i] = E[y_i] + \alpha g(E[y_i]), \end{aligned} \tag{10.17}$$

where $g(E[y_i])$ is a function of the predicted counts that is most often given values of $g(E[y_i]) = E[y_i]$ or $g(E[y_i]) = E[y_i]^2$. To conduct this test, a simple linear regression is estimated where Z_i is regressed on W_i, where

$$\begin{aligned} Z_i &= \frac{(y_i - E(y_i))^2 - y_i}{E(y_i)\sqrt{2}} \\ W_i &= \frac{g(E(y_i))}{\sqrt{2}}. \end{aligned} \tag{10.18}$$

After running the regression ($Z_i = bW_i$) with $g(E[y_i]) = E[y_i]$ and $g(E[y_i]) = E[y_i]^2$, if b is statistically significant in either case, then H_0 is rejected for the associated function g.

Example 10.3

Consider again the Poisson regression model on injury accidents estimated previously. Even though the model presented in Example 10.2 appears to fit the data fairly well, overdispersion might be expected. Using the regression test proposed by Cameron and Trivedi, with

TABLE 10.4

Negative Binomial Regression of Injury Accident Data

Independent Variable	Estimated Parameter	*t* Statistic
Constant	–0.931	–2.37
Average annual daily traffic on major road	0.0000900	3.47
Average annual daily traffic on minor road	0.000610	3.09
Median width in feet	– 0.0670	–1.99
Number of driveways within 250 ft of intersection	0.0632	2.24
Overdispersion parameter, α	0.516	3.09
Number of observations	84	
Restricted log likelihood (constant term only)	–169.25	
Log likelihood at convergence	–153.28	
Chi-squared (and associated *p*-value)	31.95 (<0.0000001)	

$g(E[y_i]) = E[y_i]$, $b = 1.28$ with a t statistic = 2.75 and with $g(E[y_i]) = E[y_i]^2$, $b = 0.28$ with a t statistic = 2.07. Based on the model output, both cases are statistically significant at the 5% level of significance, or 95% level of confidence. This result suggests that random sampling does not satisfactorily explain the magnitude of the overdispersion parameter, and a Poisson model is rejected in favor of a negative binomial model.

Example 10.4

With evidence that overdispersion is present, a negative binomial model is estimated using the accident data. The results of the estimated negative binomial regression model are shown in Table 10.4.

As with the Poisson regression model, the signs of the estimated parameters are expected and are significant. In addition, the overdispersion parameter is statistically significant, confirming that the variance is larger than the mean. The restricted log-likelihood test suggests that the fitted model is better than a model with only the constant term.

10.5 Zero-Inflated Poisson and Negative Binomial Regression Models

There are certain phenomena where an observation of zero events during the observation period can arise from two qualitatively different conditions. One condition may result from simply failing to observe an event during the observation period. Another qualitatively different condition may result from an inability ever to experience an event. Consider the following example. A transportation survey asks how many times a commuter has taken

mass transit to work during the past week. An observed zero could arise in two distinct ways. First, last week the commuter may have opted to take the van pool instead of mass transit. Alternatively, the commuter may never take transit, as a result of other commitments on the way to and from the place of employment. Thus two states are present, one a normal count-process state and the other a zero-count state.

At times what constitutes a zero-count state may be less clear. Consider vehicle accidents occurring per year on 1-km sections of highway. For straight sections of roadway with wide lanes, low traffic volumes, and no roadside objects, the likelihood of a vehicle accident occurring may be extremely small, but still present because an extreme human error could cause an accident. These sections may be considered in a zero-accident state because the likelihood of an accident is so small (perhaps the expectation would be that a reported accident would occur once in a 100-year period). Thus, the zero-count state may refer to situations where the likelihood of an event occurring is extremely rare in comparison to the normal-count state where event occurrence is inevitable and follows some known count process (see Lambert, 1992). Two aspects of this nonqualitative distinction of the zero state are noteworthy. First, there is a preponderance of zeros in the data — more than would be expected under a Poisson process. Second, a sampling unit is not required to be in the zero or near zero state into perpetuity, and can move from the zero or near zero state to the normal-count state with positive probability. Thus, the zero or near zero state reflects one of negligible probability compared to the normal state.

Data obtained from two-state regimes (normal-count and zero-count states) often suffer from overdispersion if considered as part of a single, normal-count state because the number of zeros is inflated by the zero-count state. Example applications in transportation are found in Miaou (1994) and Shankar et al. (1997). It is common not to know if the observation is in the zero state — so the statistical analysis process must uncover the separation of the two states as part of the model estimation process. Models that account for this dual-state system are referred to as zero-inflated models (see Mullahey, 1986; Lambert, 1992; Greene, 2000).

To address phenomena with zero-inflated counting processes, the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models have been developed. The ZIP model assumes that the events, $Y = (y_1, y_2, \ldots y_n)$, are independent and the model is

$$y_i = 0 \text{ with probability } p_i + \left(1 - p_i\right)EXP\left(-\lambda_i\right)$$

$$y_i = y \text{ with probability } \frac{\left(1 - p_i\right)EXP\left(-\lambda_i\right)\lambda_i^y}{y!}. \tag{10.18}$$

where y is the number of events per period. Maximum likelihood estimates are used to estimate the parameters of a ZIP regression model and confidence intervals are constructed by likelihood ratio tests.

The ZINB regression model follows a similar formulation with events, $Y = (y_1, y_2, \ldots, y_n)$, independent and

$$y_i = 0 \text{ with probability } p_i + (1-p_i)\left[\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right)+\lambda_i}\right]^{1/\alpha}$$

$$y_i = y \text{ with probability } (1-p_i)\left[\frac{\Gamma\left(\left(\frac{1}{\alpha}\right)+y\right)u_i^{1/\alpha}(1-u_i)^y}{\Gamma\left(\frac{1}{\alpha}\right)y!}\right], y = 1,\ 2,\ 3... \tag{10.19}$$

where $u_i = (1/\alpha)/[(1/\alpha)+\lambda_i]$. Maximum likelihood methods are again used to estimate the parameters of a ZINB regression model.

Zero-inflated models imply that the underlying data-generating process has a splitting regime that provides for two types of zeros. The splitting process is assumed to follow a logit (logistic) or probit (normal) probability process, or other probability processes. A point to remember is that there must be underlying justification to believe the splitting process exists (resulting in two distinct states) prior to fitting this type of statistical model. There should be a basis for believing that part of the process is in a zero-count state.

To test the appropriateness of using a zero-inflated model rather than a traditional model, Vuong (1989) proposed a test statistic for non-nested models that is well suited for situations where the distributions (Poisson or negative binomial) are specified. The statistic is calculated as (for each observation i),

$$m_i = LN\left(\frac{f_1(y_i \mid \mathbf{X}_i)}{f_2(y_i \mid \mathbf{X}_i)}\right), \tag{10.20}$$

where $f_1(y_i \mid \mathbf{X}_i)$ is the probability density function of model 1, and $f_2(y_i \mid \mathbf{X}_i)$ is the probability density function of model 2. Using this, Vuong's statistic for testing the non-nested hypothesis of model 1 vs. model 2 is (Greene, 2000; Shankar et al., 1997)

$$V = \frac{\sqrt{n}\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} m_i\right]}{\sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(m_i - \overline{m}\right)^2}} = \frac{\sqrt{n}\left(\overline{m}\right)}{S_m}, \tag{10.21}$$

where $\overline{m}$ is the mean

$$\left((1/n)\sum_{i=1}^{n} m_i\right),$$

S_m is standard deviation, and n is the sample size. Vuong's value is asymptotically standard normal distributed (to be compared to z-values), and if $|V|$ is less than $V_{critical}$ (1.96 for a 95% confidence level), the test does not support the selection of one model over another. Large positive values of V greater than $V_{critical}$ favor model 1 over model 2, whereas large negative values support model 2. For example, if comparing negative binomial alternatives, one would let $f_1(.)$ be the density function of the ZINB and $f_2(.)$ be the density function of the negative binomial model. In this case, assuming a 95% critical confidence level, if $V > 1.96$, the statistic favors the ZINB, whereas a value of $V < -1.96$ favors the negative binomial. Values in between would mean that the test was inconclusive. The Vuong test can also be applied in an identical manner to test ZIP Poisson and Poisson models.

Because overdispersion will almost always include excess zeros, it is not always easy to determine whether excess zeros arise from true overdispersion or from an underlying splitting regime. This could lead one to erroneously choose a negative binomial model when the correct model may be a ZIP. For example, recall from Equation 10.14 that the simple negative binomial gives

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]]. \tag{10.22}$$

For the ZIP model it can be shown that

$$VAR\left[y_i\right] = E\left[y_i\right] + \left[1 + \frac{p_i}{1 - p_i} E\left[y_i\right]\right]. \tag{10.23}$$

Thus, the term $p_i/(1 - p_i)$ could be erroneously interpreted as α. For guidance in unraveling this problem, consider the Vuong statistic comparing ZINB and negative binomial (with $f_1(.)$ the density function of the ZINB and $f_2(.)$ the density function of the negative binomial model for Equation 10.20). Shankar et al. (1997) provide model-selection guidelines for this case based on possible values of the Vuong test and overdispersion parameter

TABLE 10.5

Decision Guidelines for Model Selection (using the 95% confidence level) among Negative Binomial (NB), Poisson, Zero-Inflated Poisson (ZIP), and Zero-Inflated Negative Binomial (ZINB) Models Using the Vuong Statistic and the Overdispersion Parameter α

		***t* Statistic of the NB Overdispersion Parameter α**	
		<|1.96|	**>|1.96|**
Vuong statistic for ZINB(f_1(.)) and NB(f_2(.)) comparison	<– 1.96	ZIP or Poisson as alternative to NB	NB
	>1.96	ZIP	ZINB

(α) *t* statistics. These guidelines, for the 95% confidence level, are presented in Table 10.5. Again, it must be stressed that great care must be taken in the selection of models. If there is not a compelling reason to suspect that two states are present, the use of a zero-inflated model may be simply capturing model misspecification that could result from factors such as unobserved effects (heterogeneity) in the data.

Example 10.5

To determine if a zero-inflated process is appropriate for the accident data used in previous examples one must believe that three-legged stop-controlled intersections with two lanes on the minor and four lanes on the major road are capable of being perfectly safe with respect to injury accidents. If convinced that perfectly safe intersections may exist, a zero-inflated negative binomial model may be appropriate. The results from such a model are shown in Table 10.6.

The first noteworthy observation is that the negative binomial model without zero inflation appears to account for the number of observed zeros as well as the zero-inflated model. The Vuong statistic of 0.043 suggests that the test is inconclusive. Thus there is no statistical support to select the ZINB model over the standard negative binomial model (see also Table 10.5 where the *t* statistic of α nearly satisfies the 95% confidence guidelines used). This intuitively makes sense because there is no compelling reason to believe that intersections, which are noted as accidents hot spots, would be in what might be considered a zero-accident state.

10.6 Panel Data and Count Models

The time-series nature of count data from a panel (where groups of data are available over time) creates a serial correlation problem that must be dealt

TABLE 10.6

Zero-Inflated Negative Binomial Regression of Injury Accident Data: Logistic Distribution Splitting Model

Variable Description	Estimated Parameter	*t*-Statistic
Negative Binomial Accident State		
Constant	–13.84	–4.33
Log of average annual daily traffic on major road	1.337	4.11
Log of average annual daily traffic on minor road	0.289	2.82
Number of driveways within 250 ft of intersection	0.0817	2.91
Overdispersion parameter, α	0.0527	1.93
Zero-Accident State		
Constant	–5.18	0.77
Number of observations	84	
Log likelihood at convergence (Poisson)	–171.52	
Log likelihood at convergence (negative binomial)	–154.40	
Log likelihood at convergence (zero-inflated negative binomial)	–154.40	
Poisson zeros: actual/predicted	29/15.6	
Negative binomial zeros: actual/predicted	29/25.5	
Zero-inflated negative binomial zeros: actual/predicted	29/25.5	
Vuong statistic for testing zero-inflated negative binomial vs. the normal-count model	0.043	

with in estimation. Hausman et al. (1984) proposed random- and fixed-effects approaches for panel count data to resolve problems of serial correlation. The random-effects approach accounts for possible unobserved heterogeneity in the data in addition to possible serial correlation. In contrast, the fixed-effects approach does not allow for unobserved heterogeneity.

Transportation applications of fixed- and random-effects Poisson and negative binomial models have been quite limited. However, two studies are noteworthy in this regard. Johansson (1996) studied the effect of a lowered speed limit on the number of accidents on roadways in Sweden. And Shankar et al. (1998) compared standard negative binomial and random-effects negative binomial models in a study of accidents caused by median crossovers in Washington State. The reader is referred to these sources and Hausman et al. (1984) for additional information on random- and fixed-effects count models.