ECOM40006/ECOM90013 Econometrics 3
Department of Economics
University of Melbourne

Count Data Models

Semester 1, 2025

Version: May 8, 2025

Contents

1 Introduction

Count data models are relevant for a response variable (y) that can only take values that are non-negative integers. Often, there is a predominance of zeros and most numbers are relatively small integers. Examples are:

- The number of visits a person makes to a national park in a year.
- The number of visits to a doctor in a year.

- The number of car accidents in a year.

The definitive texts for this material are Cameron and Trivedi (1998) and Winkelmann (2008), but useful treatments can also be found in, *inter alia*, Cameron and Trivedi (2005, Chapter 20), Wooldridge (2002, Chapter 19), and Winkelmann and Boes (2006, Section 8.3.2).

2 The Poisson Distribution and Its Properties

The simplest distribution used to model count data is the Poisson distribution, whose probability function is given by

$$f(y_i) = P(Y_i = y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, \qquad \lambda > 0;\ y_i = 0, 1, 2, \ldots$$

To see that this probability mass function sums to unity observe that

$$\sum_{y=0}^{\infty} f(y) = \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} \times e^{\lambda} = 1, \tag{1}$$

where we have used the well-known series expansion for the exponential function

$$e^{\lambda} = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}.$$

The unknown parameter λ is equal to both the mean and variance of y. That is, $\mathrm{E}[Y] = \mathrm{Var}[Y] = \lambda$. To illustrate this result for $\mathrm{E}[Y]$ observe that

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_{y=0}^{\infty} y \times f(y) = \sum_{y=0}^{\infty} \left[y \times \frac{e^{-\lambda}\lambda^y}{y!} \right] \\
&= \underbrace{0 \times \frac{e^{-\lambda}\lambda^0}{0!}}_{y=0} + \sum_{y=1}^{\infty} \left[y \times \frac{e^{-\lambda}\lambda^y}{y!} \right] \\
&= \underbrace{0 \times \frac{e^{-\lambda} \times 1}{1}}_{=0} + \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^y}{(y-1)!} \\
&= \lambda \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^{y-1}}{(y-1)!}
\end{aligned} \tag{2}$$

Writing $j = y - 1$ we see that

$$\mathrm{E}[Y] = \lambda \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} = \lambda, \tag{3}$$

where the final equality follows from (1).

The variance is a little more work, although similar in spirit. First,

$$\mathrm{E}\left[Y^2\right] = \sum_{y=0}^{\infty} \left[y^2 \times \frac{e^{-\lambda}\lambda^y}{y!} \right]$$

$$= \underbrace{0^2 \times \frac{e^{-\lambda}\lambda^0}{0!}}_{y=0} + \sum_{y=1}^{\infty}\left[y^2 \times \frac{e^{-\lambda}\lambda^y}{y!}\right]$$
$$= \sum_{y=1}^{\infty}\left[y \times \frac{e^{-\lambda}\lambda^y}{(y-1)!}\right]$$

Again, writing $j = y - 1$, we see that

$$\begin{aligned}\mathrm{E}\left[Y^2\right] &= \sum_{j=0}^{\infty}\left[(j+1) \times \frac{e^{-\lambda}\lambda^{j+1}}{j!}\right] \\ &= \lambda \underbrace{\sum_{j=0}^{\infty}\left[j \times \frac{e^{-\lambda}\lambda^j}{j!}\right]}_{=\mathrm{E}[Y] \text{ from } (2)} + \lambda \underbrace{\sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^j}{j!}}_{=1 \text{ from } (1)} \\ &= \lambda^2 + \lambda,\end{aligned}$$

where the final equality follows from (3). Second,

$$\mathrm{Var}\left[Y\right] = \mathrm{E}\left[Y^2\right] - \left(\mathrm{E}\left[Y\right]\right)^2 = \lambda^2 + \lambda - (\lambda)^2 = \lambda,$$

as required.

3 The Poisson Regression Model

In a regression framework, the probability of a count depends on a set of explanatory variables, such as income, age, gender, etc. To allow for this dependence, we make λ a *strictly positive* function of these variables. The most common function that is used is

$$\lambda_i = \exp\{x_i'\beta\} \qquad \text{or} \qquad \ln \lambda_i = x_i'\beta$$

so that

$$f(y_i) = \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i > 0;\ y_i = 0, 1, 2, \ldots \tag{4}$$

and

$$\mathrm{E}\left[Y_i\right] = \mathrm{Var}\left[Y_i\right] = \lambda_i = \exp\{x_i'\beta\}.$$

This model is often called the Poisson regression model (PRM). Observe that, because both the variance and mean depend on x_i, the model is intrinsically heteroskedastic.

Typical quantities for which we seek estimates are:

- Probabilities for different values of y_i given a vector x_i

$$\Pr\left(Y_i = y_i\right) = \frac{\exp\{-\exp\{x_i'\beta\}\}[\exp\{x_i'\beta\}]^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots$$

- The average count, λ_i, for a given x_i

$$\lambda_i = \exp\{x_i'\beta\}.$$

- The marginal effect of a change in the k-th explanatory variable, x_{ik}, on the average count, λ_i, or the marginal mean effect, is

$$\frac{\partial \lambda_i}{\partial x_{ik}} = \beta_k \exp\{x_i'\beta\} = \beta_k \lambda_i.$$

Since

$$\beta_k = \frac{\partial \lambda_i / \partial x_{ik}}{\exp\{x_i'\beta\}} = \frac{\partial \lambda_i / \partial x_{ik}}{\lambda_i},$$

the coefficient β_k can be viewed as describing the relative change in λ_i for a marginal change in the k-th element of x. We note that it is constant for all i.

If we consider the quantity

$$\frac{\partial^2 \lambda_i}{\partial x_{ik} \partial x_{il}} = \beta_k \beta_l \exp\{x_i'\beta\} = \beta_k \beta_l \lambda_i \neq 0 \text{ whenever } \beta_k \neq 0 \text{ and } \beta_l \neq 0,$$

we see that, unlike the linear regression model, interactive effects are implied by the PRM even though the model contains no explicit interactive term of the form $x_{ik}x_{il}$.

4 Maximum Likelihood Estimation of the Poisson Regression Model

Including the above assumption and taking logs of the probability function yields

$$\ln f(y_i) = -\exp\{x_i'\beta\} + y_i x_i'\beta - \ln(y_i!).$$

Consequently, for a simple random same of size n, the log-likelihood function is given by

$$\ln \mathcal{L}(\beta; y) = \sum_{i=1}^{n} \left[-\exp\{x_i'\beta\} + y_i x_i'\beta - \ln(y_i!) \right]. \tag{5}$$

The first-order conditions for a maximum are

$$\begin{aligned}
0 &= \left. \frac{\partial \ln \mathcal{L}(\beta; y)}{\partial \beta} \right|_{\beta=\hat{\beta}} \\
&= \sum_{i=1}^{n} \left[-\frac{\partial \exp\{x_i'\beta\}}{\partial x_i'\beta} \frac{\partial x_i'\beta}{\partial \beta} + \frac{y_i \partial x_i'\beta}{\partial \beta} - \frac{\partial \ln(y_i!)}{\partial \beta} \right]_{\beta=\hat{\beta}} \\
&= \sum_{i=1}^{n} \left[-\exp\{x_i'\beta\} x_i + y_i x_i - 0 \right]_{\beta=\hat{\beta}} \\
&= \sum_{i=1}^{n} \left[y_i - \exp\{x_i'\hat{\beta}\} \right] x_i.
\end{aligned}$$

These conditions can be viewed as an orthogonality between the (generalized) residuals $\left[y_i - \exp\{x_i'\hat{\beta}\} \right]$ and x_i. The hessian

$$\left. \frac{\partial^2 \ln L(\beta; y)}{\partial \beta \partial \beta'} \right|_{\beta=\hat{\beta}} = \left. \frac{\partial \mathcal{S}(\beta; y, X)'}{\partial \beta} \right|_{\beta=\hat{\beta}} = \left. \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \left[y_i - \exp\{x_i'\beta\} \right] x_i' \right|_{\beta=\hat{\beta}}$$

$$= -\sum_{i=1}^{n} \exp\{x_i'\hat{\beta}\} x_i x_i'$$

is negative definite for all x_i and $_{\beta=\hat{\beta}}$. Thus, convergence to the maximum likelihood estimator is typically rapid and straight-forward.

5 Departures from the Assumptions Underlying the PRM

There is a variety of different ways in which the data may not exactly match the assumptions underlying the PRM. For many of these problems simple adjustments to the model can be made which address the problem.

Truncation Often samples are truncated because zeros are not observed. Zeros are not observed when only those individuals taking part in an activity are selected. For example, surveys taken on buses to find out how frequently people travel by bus; surveys taken in national parks to find out how frequently people visit a park. When the distribution is truncated such that zeros are not observed, the relevant probability function from which the likelihood function can be specified is

$$f(y_i|y_i \geq 1) = \frac{f(y_i)}{P(y_i \geq 1)} = \frac{f(y_i)}{1 - f(0)} = \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i!\,[1 - \exp\{-\lambda_i\}]}, \quad \begin{array}{l} \lambda > 0, \\ y = 1, 2, \ldots \end{array}$$

where

$$f(0) = P(y_i = 0) = \frac{\exp\{-\lambda_i\}\lambda_i^0}{0!} = \frac{\exp\{-\lambda_i\} \times 1}{1} = \exp\{-\lambda_i\}.$$

Observe that the adjustment of the probability function in this case is completely analogous to what was done for truncated regression models. Logging this density yields the contribution to the log-likelihood by each observation:

$$-\exp\{x_i'\beta\} + y_i x_i'\beta - \ln(y_i!) - \ln\left[1 - \exp\left(-\exp\{x_i'\beta\}\right)\right]. \tag{6}$$

Comparison of (6) with the corresponding component of (5) reveals that truncation is handled by augmenting the log-likelihood with an extra term.

Censoring Censoring occurs when the number of events greater than or equal to a particular value, say c, are aggregated into a single category designated as $y \geq c$. An example is a survey where respondents are asked to select the number of times they participate in some activity, and the numbers listed in the survey are, for example, 0, 1, 2, ..., 9, 10, 'more than 10'. In other examples there is a natural ceiling to the maximum number of counts, such as the number of wickets taken by a bowler in an innings of cricket. In these cases, the relevant probability function is

$$g(y_i) = \begin{cases} f(y_i), & \text{if } y_i < c, \\ P(y_i \geq c), & \text{if } y_i \geq c. \end{cases}$$

The function $f(y_i)$ is the usual Poisson probability function and contributes to the log-likelihood function only for observations where $y_i < c$. For observations where

$y_i \geq c$, the relevant probability that contributes to the likelihood function is

$$P(y_i \geq c) = 1 - F(c-1) = 1 - \sum_{j=0}^{c-1} \frac{\exp\{-\lambda_i\}\lambda_i^j}{j!}.$$

If we define the indicator function

$$I_{[y_i<c]}(y_i) = \begin{cases} 1, & \text{if } y_i < c, \\ 0, & \text{otherwise,} \end{cases}$$

then the log-likelihood becomes

$$\ln L(\lambda_i, y) = \sum_{i=1}^{n} \left\{ I_{[y_i<c]}(y_i) \ln f(y_i) + (1 - I_{[y_i<c]}(y_i)) \ln[1 - F(c-1)] \right\}.$$

Excess Zeros The data have more zeros than could be either predicted by or consistent with the PRM. In such cases we are said to have 'extra' or 'excess' zeros. One possible solution to the excess zeros problem is a *hurdle model*. The basic idea is to break the problem into two parts. This first part determines whether or not the outcome is a zero and the second part then models positive outcomes. Different distributions are used at each stage. For example, an initial decision to see a doctor may be made by the patient. Decisions about subsequent visits are likely to depend on the patient-doctor relationship.

Suppose that the hurdle model is constructed using two Poisson distributions. One distribution is used to describe two probabilities, the probability that y_i is zero and the probability that y_i is positive. The second distribution is used to model probabilities for all possible positive values of y_i. Let the density function that discriminates between zero and positive values be given by

$$f_0(y_i) = \frac{\exp\{-\theta_i\}\theta_i^{y_i}}{y_i!}, \qquad \theta_i > 0, y_i = 0, 1, 2, \ldots.$$

Thus,

$$P(Y_i = 0) = f_0(0) = \exp\{-\theta_i\}$$

and

$$P(Y_i > 0) = 1 - f_0(0) = 1 - \exp\{-\theta_i\}.$$

Because it does not include zeros, the Poisson model for $Y_i > 0$ is a truncated one with density function

$$f_1(y_i|y_i > 0) = \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i![1 - \exp\{-\lambda_i\}]}, \qquad \lambda_i > 0, y_i = 0, 1, 2, \ldots$$

The complete density function for the hurdle model then becomes[1]

$$f_h(y_i) = \begin{cases} \exp\{-\theta_i\}, & \text{if } y_i = 0, \\ \dfrac{[1 - \exp\{-\theta_i\}]}{[1 - \exp\{-\lambda_i\}]} \dfrac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i!}, & \text{if } y_i > 0. \end{cases} \tag{7}$$

[1]Note the use of the symbol $f_h(\cdot)$ to distinguish this probability function from that of the PRM, denoted simply $f(\cdot)$.

The result $f_h(y_i) = f_1(y_i|y_i > 0)P(y_i > 0)$ is used to obtain the second branch of this expression. If desired, θ_i and λ_i can depend on different explanatory variables, say,

$$\theta_i = \exp\{x'_{0i}\beta_0\} \qquad \text{and} \qquad \lambda_i = \exp\{x'_{1i}\beta_1\}.$$

Observe that if $\theta_i = \lambda_i$, or equivalently $\beta_0 = \beta_1$ in the case of regression models where $x_{0i} = x_{1i}$, then (7) collapses to (4). This hypothesis is readily tested using a likelihood ratio test. If the choice between zero and positive counts does not depend on some explanatory variables, then the above expression simplifies with $\exp\{-\theta_i\}$ being replaced by a single parameter, say p.

Consider the marginal probability effect (MPE) of a change to some regressor, w_{ik} say,

$$\frac{\partial P(y_i = j|x_i)}{\partial x_{ik}} = \frac{\partial f_h(y_i|x_i)}{\partial x_{ik}} = f_h(y_i)h(y_i)$$

where[2]

$$h(y_i) = \begin{cases} -\theta_i\beta_{0k}, & \text{if } y_i = 0, \\ (y_i - \lambda_i)\beta_{1k} + \dfrac{\exp\{-\theta_i\}\theta_i\beta_{0k}}{[1 - \exp\{-\theta_i\}]} - \dfrac{\exp\{-\lambda_i\}\lambda_i\beta_{1k}}{[1 - \exp\{-\lambda_i\}]}, & \text{if } y_i > 0. \end{cases}$$

Contrast this marginal effect with that for the PRM, based on (4), where

$$\frac{\partial P(y_i = j|x_i)}{\partial x_{ik}} = \frac{\partial f(y_i|x_i)}{\partial x_{ik}} = f(y_i)\tau(y_i)$$

where, in this case, $\lambda_i = \exp\{x'_i\beta\}$ and

$$\tau(y_i) = \begin{cases} -\lambda_i\beta_k, & \text{if } y_i = 0, \\ (y_i - \lambda_i)\beta_k, & \text{if } y_i > 0. \end{cases} \tag{8}$$

Note that (8) has been written to reflect the structure in (5) to highlight those difference, but not that we might equally have written

$$\tau(y_i) = (y_i - \lambda_i)\beta_k, \qquad y_i = 0, 1, 2, \ldots$$

If we think about this latter formulation it is clear that, when $y_i = 0$ that $y_i - \lambda_i < 0$ and so the sign of the MPE will be the opposite of that of $\text{sgn}\,(\beta_k)$.[3] However, as y_i increases there will come a point where $y_i - \lambda_i > 0$, whereupon the MPE will be the same as that of $\text{sgn}\,(\beta_k)$. Moreover, there can only be one point when the sign of the MPE changes. This is completely analogous to the single-crossing property that characterized the ordered probit model. Conversely, in the hurdle model, the interactions of y_i, λ_i and θ_i in the second branch of (5) are sufficiently complicated that it is possible for the MPE to change sign more than once; see, for example, Winkelmann and Boes (2006, Figure 8.10).

[2] If w_{ik} does not belong to either x_{0i} or x_{1i} (or both) then the corresponding coefficient, either β_{0k} or β_{1k} (or both), will be zero. A derivation of $h(y_i)$ is given in the Appendix.

[3] The notation $\text{sgn}\,(\beta_k)$ should be read to mean the sign of β_k.

One consequence of the way in which we developed the hurdle model is that the relationship between this model and the PRM is perhaps less clear than it might be. It can be shown that the hurdle model corresponds to a model of the form

$$f(y_i|x_i) = w_i d_i + (1 - w_i)\frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i > 0;\ y_i = 0, 1, 2, \ldots \tag{9}$$

where d_i is a dummy variable taking the value unity when $y_i = 0$ and zero otherwise, and w_i is the amount by which you want to perturb the probability of $Y_i = 0$ from $\exp\{-\lambda_i\}$, the probability implied by the PRM. Of course, if this probability changes there must be corresponding changes elsewhere so that the probabilities continue to sum to unity, which is the role played by the quantity $1 - w_i$. If we choose

$$w_i = \frac{\exp\{-\theta_i\} - \exp\{-\lambda_i\}}{1 - \exp\{-\lambda_i\}},$$

so that

$$1 - w_i = \frac{1 - \exp\{-\theta_i\}}{1 - \exp\{-\lambda_i\}},$$

then we obtain the hurdle model. We are, however, not constrained to make this choice and, for alternative choices, (9) is known as the *zero-inflated count data model*. For more about these models see Mullahy (1997).

Over-dispersion One of the implications of the PRM is that the mean and variance are equal. This property is known as *equi-dispersion*. In practice, data frequently imply that equi-dispersion is unlikely to hold in the population. In cases where the mean and variance of the population differ we say that there is either *under-dispersion* or *over-dispersion* when the variance is either smaller or larger than the mean, respectively. Of these two problems over-dispersion is encountered far more commonly and so it is that problem that we shall focus on. The arguments tend to work in essentially the same way for under-dispersion.

How might over-dispersion come about? Suppose that we are unable to characterize all of the variation in individual responses by our linear index because some of the sources of variation are unobservable. One way to characterize this *unobserved heterogeneity* is to replace $\lambda_i = \exp\{x_i'\beta\}$ by

$$\tilde{\lambda}_i = \exp\{x_i'\beta + \epsilon_i\} \tag{10}$$

where the disturbance term ϵ_i is meant to capture the unobserved heterogeneity. Note that simple factorization yields

$$\tilde{\lambda}_i = \exp\{x_i'\beta\}\exp\{\epsilon_i\} = \lambda_i u_i$$

where $u_i = \exp\{\epsilon_i\}$. That is, y_i is Poisson conditional on x_i and u_i. Typically we are seeking to model y_i conditional on x_i alone and so we need to get rid of the conditioning on u_i. In order to do this we need two standard results concerning conditional expectations; see, for example, Mood, Graybill, and Boes (1974, Chapter

4, Section 4.3). The first, called the law of iterated expectations, states that[4]

$$\mathrm{E}\left[Y\right] = \mathrm{E}_X\left[\mathrm{E}\left[Y|X=x\right]\right] \tag{11}$$

and the second tells us about the relationship between conditional and unconditional variances:[5]

$$\mathrm{Var}\left[Y\right] = \mathrm{E}_X\left[\mathrm{Var}\left[Y|X=x\right]\right] + \mathrm{V}_X\left[\mathrm{E}\left[Y|X=x\right]\right] \tag{12}$$

To apply these rules we equate Y to y_i given x_i and X to u_i. Hence, from (11),

$$\mathrm{E}\left[y_i|x_i\right] = \mathrm{E}_u\left(\mathrm{E}\left[y_i|x_i,u_i\right]\right) = \mathrm{E}_u(\tilde{\lambda}_i|x_i,u_i) = \mathrm{E}_u(\lambda_i u_i|x_i,u_i) = \lambda_i \mathrm{E}_u(u_i|x_i).$$

Similarly, using (12),

$$\begin{aligned}\mathrm{Var}\left[y_i|x_i\right] &= \mathrm{E}_u\left[\mathrm{Var}\left[y_i|x_i,u_i\right]\right] + \mathrm{V}_u\left[\mathrm{E}\left[y_i|x_i,u_i\right]\right] \\ &= \mathrm{E}_u[\tilde{\lambda}_i|x_i] + \mathrm{V}_u[\tilde{\lambda}_i|x_i] \\ &= \lambda_i \mathrm{E}_u\left[u_i|x_i\right] + \lambda_i^2 \mathrm{V}_u\left[u_i|x_i\right]\end{aligned}$$

We see that it is necessary to make some assumptions about the conditional moments of u_i given x_i. The usual assumptions are $\mathrm{E}\left[u_i|x_i\right] = 1$, which is in the spirit of assuming $\mathrm{E}\left[\epsilon_i\right] = 0$, and $\mathrm{Var}\left[u_i|x_i\right] = \sigma_u^2$. Making these substitutions yields

$$\begin{aligned}\mathrm{E}\left[y_i|x_i\right] &= \lambda_i \\ \mathrm{Var}\left[y_i|x_i\right] &= \lambda_i + \lambda_i^2\sigma_u^2\end{aligned}$$

Because both $\lambda_i^2 > 0$ and $\sigma_u^2 > 0$ it follows that unobserved heterogeneity of the type described by (10) leads unequivocally to over-dispersion.

What are the consequences for estimation of over-dispersion? It can be shown that if the PRM maximum likelihood estimate is used when over-dispersion exists, then:

---

[4]Proof: If X and Y are joint distributed according to $f(x,y)$ then, by definition,

$$\mathrm{E}\left[Y\right] = \iint y\, f(x,y)\, dxdy = \iint y\, \frac{f(x,y)}{f(x)}\, f(x)\, dxdy = \iint y\, f(y|x)\, dy\, f(x)\, dx = \int \mathrm{E}\left[y|x\right]\, f(x)\, dx$$

which is the desired result.

[5]A proof of this result requires a little more work than the previous one. First, by definition,

$$\mathrm{Var}\left[Y|X=x\right] = \mathrm{E}\left[Y^2|X=x\right] - \left(\mathrm{E}\left[Y|X=x\right]\right)^2$$

and so

$$\mathrm{E}_X(\mathrm{Var}\left[Y|X=x\right]) = \mathrm{E}_X(\mathrm{E}\left[Y^2|X=x\right]) - \mathrm{E}_X[(\mathrm{E}\left[Y|X=x\right])^2] = \mathrm{E}\left[Y^2\right] - \mathrm{E}_X[(\mathrm{E}\left[Y|X=x\right])^2]$$

Adding and subtracting $(\mathrm{E}\left[Y\right])^2$ yields

$$\begin{aligned}\mathrm{E}_X(\mathrm{Var}\left[Y|X=x\right]) &= \mathrm{E}\left[Y^2\right] - (\mathrm{E}\left[Y\right])^2 - \left(\mathrm{E}_X[(\mathrm{E}\left[Y|X=x\right])^2] - (\mathrm{E}\left[Y\right])^2\right) \\ &= \mathrm{Var}\left[Y\right] - \left(\mathrm{E}_X[(\mathrm{E}\left[Y|X=x\right])^2] - (\mathrm{E}\left[Y\right])^2\right).\end{aligned}$$

Finally, writing $\mathrm{E}\left[Y\right] = \mathrm{E}_X\left[\mathrm{E}\left[Y|X=x\right]\right]$

$$\mathrm{E}_X(\mathrm{Var}\left[Y|X=x\right]) = \mathrm{Var}\left[Y\right] - \left(\mathrm{E}_X[(\mathrm{E}\left[Y|X=x\right])^2] - (\mathrm{E}_X\left[\mathrm{E}\left[Y|X=x\right]\right])^2\right) = \mathrm{Var}\left[Y\right] - \mathrm{V}_X[\mathrm{E}\left[Y|X=x\right]]$$

and the desired result follows on rearrangement.

- The estimates of β, and hence of λ_i, are still consistent.
- These estimates are inconsistent if truncation or censoring exists.
- Standard errors tend to be understated and t values inflated.

So it is clear that over-dispersion/under-dispersion is something that we should test for in these models.

To test for over-dispersion, we assume that

$$\text{Var}\left[y_i|x_i\right] = \lambda_i + \alpha g(\lambda_i)$$

where $g(\cdot)$ is a known function, usually taken as $g(\lambda_i) = \lambda_i$ or $g(\lambda_i) = \lambda_i^2$, and α is an unknown parameter. Then, we test $H_0 : \alpha = 0$ against $H_1 : \alpha > 0$. When testing for under-dispersion the alternative hypothesis is $H_1 : \alpha < 0$ and when testing for either under or over-dispersion the alternative hypothesis is $H_1 : \alpha \neq 0$. The steps for performing this test are:

1. Estimate the Poisson model and compute the fitted values $\hat{\lambda}_i = \exp\{x_i'\hat{\beta}\}$.
2. Run the following least squares regression

$$\frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i} = \alpha \frac{g(\hat{\lambda}_i)}{\hat{\lambda}_i} + error_i$$

3. Use the normal distribution to test the significance of α. Note that, when $g(\lambda_i) = \lambda_i$, the right hand side of the above equation is just a constant. When $g(\lambda_i) = \lambda_i^2$, it is equal to $\alpha\hat{\lambda}_i$. The numerator of the left hand side $[(y_i - \hat{\lambda}_i)^2 - y_i]$ is used to estimate $\text{Var}\left[y_i\right] - \lambda_i$. It is divided by $\hat{\lambda}_i$ to standardize the test statistic so that it has a $\text{N}(0, 1)$ distribution under H_0.

If we accept that over-dispersion is a problem then we must either contemplate alternative models that are sufficiently flexible to allow for the over-dispersion, which shall be done using the negative binomial distribution, or else find ways to make our inference robust to the problem. We shall explore both of these options in turn. Before we do that we need some more distribution theory.

6 A Digression on the Gamma Distribution

Before introducing the negative binomial distribution as an alternative to the Poisson model, we need some background on the gamma distribution. A random variable X has the gamma distribution if its probability density function is given by

$$f(x) = \frac{1}{b\Gamma(c)} \left(\frac{x}{b}\right)^{c-1} \exp\left(-\frac{x}{b}\right),$$

where b and c are positive constants and $\Gamma(\cdot)$ is the gamma function defined by

$$\Gamma(c) = \int_0^\infty t^{c-1} e^{-t} dt.$$

Using integration by parts, it is possible to show that the gamma function has the property $\Gamma(c) = (c-1)\Gamma(c-1)$, so that it can be thought of as a generalized factorial

function. The gamma function must be evaluated numerically unless c is an integer or $2c$ is an integer. When c is an integer, we can use the result $\Gamma(c) = (c-1)!$. (By convention we see that $\Gamma(1) = 0! = 1$, however, $\Gamma(0)$ is undefined, as is $\Gamma(-c)$ for any integer value of c.) When $2c$ is an integer, but c is not, we can use the recursion formula $\Gamma(c) = (c-1)\Gamma(c-1)$ and the fact that $\Gamma(1/2) = \sqrt{\pi}$. The mean and variance of X are $\mathrm{E}[X] = bc$ and $\mathrm{Var}[X] = b^2c$. Of interest for the negative binomial model for count data is a gamma random variable with mean $\mathrm{E}[X] = 1$ and $\mathrm{Var}[X] = \alpha$. In this case $bc = 1$ and $b^2c = \alpha$, so that $b = \alpha$ and $c = \alpha^{-1}$. The pdf becomes

$$f(x) = \frac{1}{\alpha\Gamma(\alpha^{-1})}\left(\frac{x}{\alpha}\right)^{\frac{1}{\alpha}-1}\exp\left(-\frac{x}{\alpha}\right).$$

7 The Negative Binomial Regression Model

To introduce the negative binomial model for count data, we begin with our earlier PRM with unobserved heterogeneity so that

$$f(y_i|\lambda_i) = \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i!}$$

with

$$\lambda_i = \mu_i\nu_i = [\exp\{x_i'\beta\}]\nu_i,$$

where ν_i is random variable with mean 1. Thus,

$$\mathrm{E}[\lambda_i \mid x_i] = \mu_i\,\mathrm{E}[\nu_i \mid x_i] = \exp\{x_i'\beta\}.$$

As before the term ν_i represents unobserved individual heterogeneity. Sample individuals with a given x_i can now have different λ_i, but the average of all the λ_i is given by $\exp\{x_i'\beta\}$. The marginal density function for (that is needed to construct the likelihood function for and the unobserved parameters in the density function for y_i (that is needed to construct the likelihood for β and the unobserved parameters in the density function for ν_i) is given by

$$f(y) = \int f(y|\nu)\,f(\nu)\,d\nu.$$

If ν has a gamma distribution with mean $\mathrm{E}[\nu] = 1$ and variance $\mathrm{Var}[\nu] = \alpha = c^{-1}$, then, making the appropriate substitutions into the above expression and evaluating the integral yields a function $f(y)$ that is a negative binomial distribution with probability density function[6]

$$f(y_i) = \frac{\Gamma(c+y_i)}{\Gamma(c)\Gamma(y_i+1)}\left(\frac{c}{c+\mu_i}\right)^c\left(\frac{\mu_i}{\mu_i+c}\right)^{y_i}.$$

The problem with this most elegant of developments is that it is something of a red herring and doesn't lead to the mean and variance claimed. They are best thought of a moment conditions used by a GMM estimator that really has very little to do with the Negative Binomial Distribution.

An alternative device that makes more clearly the link between the Negative Binomial and the Poisson Regression Models is to nest the two distributions in a more general family of distributions. This was done by Katz (1965) although his paper, being in a conference volume, can be a little hard to find. Subsequent econometric contributions along these lines include Lee (1986) and McCabe and Skeels (2020, especially the appendices).

[6] A derivation is available in Winkelmann and Boes (2006, p.286).

8 A GMM Approach to Over-Dispersion

One can simply specify moment conditions in the spirit of a GMM approach. In particular, we can specify the mean and variance

$$\mathrm{E}\left[y_i \mid x_i\right] = \mu_i = \exp\{x_i'\beta\} \qquad \text{and} \qquad \mathrm{Var}\left[y_i \mid x_i\right] = \mu_i + \alpha\mu_i^2,$$

as we had previously with unobserved heterogeneity. Thus, the negative binomial distribution is more general than the Poisson; the variance is no longer necessarily equal to the mean. The log-likelihood function is given by $\ln \mathcal{L} = \sum_{i=1}^{n} \ln f(y_i)$.

Sometimes the above specification is called a negative binomial 2 (NEGBIN2) model. The so-called negative binomial 1 (NEGBIN1) model is obtained by setting $\alpha = \delta/\mu_i$ so that $\mathrm{Var}\left[y_i\right] = \mu_i(1+\delta)$. Recall that, when discussing testing for over-dispersion we considered the model with $\mathrm{Var}\left[y_i\right] = \lambda_i + \alpha g(\lambda_i)$, with $g(\lambda_i) = \lambda_i$ or $g(\lambda_i) = \lambda_i^2$. These two variance specifications correspond to negative binomial models 1 and 2, respectively.

9 Robust Standard Errors

The alternative to generalized models is to generate robust standard errors for our maximum likelihood estimates (remember that $\hat{\beta}$ remains consistent in the absence of censoring or truncation). To proceed we shall assume that the mean number of counts is given by the exponential function

$$\mathrm{E}\left[y_i \mid x_i\right] = \exp\{x_i'\beta\}$$

and the log-likelihood function of the Poisson model

$$\ln \mathcal{L}(\beta; y, X) = \sum_{i=1}^{n} \left[-\exp\{x_i'\beta\} + y_i x_i'\beta - \ln(y_i!)\right]$$

is maximized to find an estimator $\hat{\beta}$. The estimator $\hat{\beta}$ is consistent and consistent estimates of its asymptotic standard errors can be obtained even when the distribution of counts is not Poisson. The formula for estimating the covariance matrix for $\hat{\beta}$ (from which standard errors are obtained) depends on what further assumptions about $\mathrm{Var}\left[y_i\right]$ are made. If we assume $\mathrm{Var}\left[y_i \mid x_i\right] = \mathrm{E}\left[y_i \mid x_i\right] = \exp\{x_i'\beta\}$, the same variance assumption as is implied by the Poisson distribution, then the estimated covariance matrix is

$$\mathrm{V}_1 = H^{-1}$$

where

$$H = -\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'}\bigg|_{\beta=\hat{\beta}} = \sum_{i=1}^{n} \exp\{x_i'\hat{\beta}\} x_i x_i'$$

If we allow for under or over-dispersion by assuming that $\mathrm{Var}\left[y_i\right] = \sigma^2 \exp\{x_i'\beta\}$, then the estimated covariance matrix is

$$\mathrm{V}_2 = \hat{\sigma}^2 H^{-1}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(y_i - \exp\{x_i'\hat{\beta}\}\right)^2}{\exp\{x_i'\hat{\beta}\}}$$

This estimator is known as the generalized linear model (GLM) variance estimator.

If no variance assumption for is made (the variance must be finite), then a consistent estimate of the covariance matrix is

$$\mathrm{V}_3 = H^{-1} B H^{-1}$$

where

$$B = \sum_{i=1}^{n} \frac{\partial \ln \mathcal{L}}{\partial \beta} \frac{\partial \ln \mathcal{L}}{\partial \beta'} \bigg|_{\beta=\hat{\beta}} = \sum_{i=1}^{n} \left(y_i - \exp\{x_i' \hat{\beta}\} \right)^2 x_i x_i'$$

This estimator is known as White's robust covariance matrix estimator.

10 R

Like we saw in the case of Binary Response Models, the Poisson Regression Model can be estimated using the `glm` command. What you need to know is that is that the family is `poisson` and that the link function is `log`. Thus, a generic command is of the form

```
mdl.glm=glm(dependent variable ~ list of explanators, family = poisson(link
= "log"))
mdl.glm.stats=summary.glm(mdl.glm)
```

where the second of these commands generates similar output to that which was discussed at length in the Binary Response Models handout and that discussion won't be repeated here.

References

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Econometric Society Monographs No. 30, Cambridge University Press, Cambridge. 2

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge. 2

Katz, L. (1965). Unified treatment of a broad class of discrete distributions. In *Classical and Contagious Discrete Distributions. Proceedings of the International Symposium held at McGill University, Montreal, Canada, August 15–August 20, 1963*, G. P. Patil, editor, 175–182, Statistical Publishing Society, Calcutta; Pergamon Press, Oxford. 11

Lee, L.-F. (1986). Specification test for Poisson regression models. *International Economic Review* 27(3), 689–706. 11

McCabe, B. P. and C. L. Skeels (2020). Distributions you can count on ... but what's the point. *Econometrics* 8(9), 1–36, doi:10.3390/econometrics8010009. 11

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., New York, third edition. 8

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* 12(3), 337–350. 8

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer-Verlag, Berlin, 5th edition, ISBN 978-3-540-77648-2, doi:10.1007/978-3-540-78389-3. 2

Winkelmann, R. and S. Boes (2006). *Analysis of Microdata*. Springer-Verlag, Berlin. 2, 7, 11

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts. 2

A Derivation of $h(y_i)$

We need to evaluate

$$\frac{\partial P(y_i = j|x_i)}{\partial x_{ik}} = \frac{\partial f(y_i|x_i)}{\partial x_{ik}} = \begin{cases} \dfrac{\partial \exp\{-\theta_i\}}{\partial x_{ik}}, & \text{if } y_i = 0, \\ \dfrac{\partial}{\partial x_{ik}} \dfrac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]} \dfrac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{y_i!}, & \text{if } y_i > 0. \end{cases}$$

Dealing with each of the branches in turn, we have first

$$\begin{aligned} \frac{\partial \exp\{-\theta_i\}}{\partial x_{ik}} &= \frac{\partial \exp\{-\theta_i\}}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial x_{0i}'\beta_0} \times \frac{\partial x_{0i}'\beta_0}{\partial x_{ik}} \\ &= -\exp\{-\theta_i\} \times \exp\{x_{0i}'\beta_0\} \times \beta_{0k} \\ &= -\exp\{-\theta_i\}\theta_i\beta_{0k}. \end{aligned}$$

By symmetry we have immediately that

$$\frac{\partial \lambda_i}{\partial x_{ik}} = \lambda_i\beta_{1k} \qquad \text{and} \qquad \frac{\partial \exp\{-\lambda_i\}}{\partial x_{ik}} = -\exp\{-\lambda_i\}\lambda_i\beta_{1k}.$$

Turning to the (much messier) second branch we have

$$\begin{aligned} &\frac{\partial}{\partial x_{ik}} \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]} \exp\{-\lambda_i\}\lambda_i^{y_i} \\ &= \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]} \exp\{-\lambda_i\} \frac{\partial \lambda_i^{y_i}}{\partial x_{ik}} \\ &\quad + \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]} \lambda_i^{y_i} \frac{\partial \exp\{-\lambda_i\}}{\partial x_{ik}} \\ &\qquad + \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{[1-\exp\{-\lambda_i\}]} \frac{\partial [1-\exp\{-\theta_i\}]}{\partial x_{ik}} \\ &\qquad\quad + [1-\exp\{-\theta_i\}] \exp\{-\lambda_i\}\lambda_i^{y_i} \frac{\partial [1-\exp\{-\lambda_i\}]^{-1}}{\partial x_{ik}} \\ &= \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]} \exp\{-\lambda_i\} y_i \lambda_i^{y_i-1} \frac{\partial \lambda_i}{\partial x_{ik}} \\ &\quad - \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]} \lambda_i^{y_i} \exp\{-\lambda_i\}\lambda_i\beta_{1k} \\ &\qquad - \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{[1-\exp\{-\lambda_i\}]} \frac{\partial \exp\{-\theta_i\}}{\partial x_{ik}} \end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]^2}\exp\{-\lambda_i\}\lambda_i^{y_i}\frac{\partial[1-\exp\{-\lambda_i\}]}{\partial x_{ik}} \\
=&\frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]}\exp\{-\lambda_i\}y_i\lambda_i^{y_i}\beta_{1k} \\
&- \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]}\lambda_i^{y_i}\exp\{-\lambda_i\}\lambda_i\beta_{1k} \\
&\quad + \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{[1-\exp\{-\lambda_i\}]}\exp\{-\theta_i\}\theta_i\beta_{0k} \\
&\quad\quad + \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]^2}\exp\{-\lambda_i\}\lambda_i^{y_i}\frac{\partial[\exp\{-\lambda_i\}]}{\partial x_{ik}} \\
=&\frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]}\exp\{-\lambda_i\}y_i\lambda_i^{y_i}\beta_{1k} \\
&- \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]}\lambda_i^{y_i}\exp\{-\lambda_i\}\lambda_i\beta_{1k} \\
&\quad + \frac{\exp\{-\lambda_i\}\lambda_i^{y_i}}{[1-\exp\{-\lambda_i\}]}\exp\{-\theta_i\}\theta_i\beta_{0k} \\
&\quad\quad - \frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]^2}\exp\{-\lambda_i\}\lambda_i^{y_i}\exp\{-\lambda_i\}\lambda_i\beta_{1k} \\
=&\frac{[1-\exp\{-\theta_i\}]}{[1-\exp\{-\lambda_i\}]}\exp\{-\lambda_i\}\lambda_i^{y_i} \\
&\times\left[(y_i-\lambda_i)\beta_{1k} + \frac{\exp\{-\theta_i\}\theta_i\beta_{0k}}{[1-\exp\{-\theta_i\}]} - \frac{\exp\{-\lambda_i\}\lambda_i\beta_{1k}}{[1-\exp\{-\lambda_i\}]}\right]
\end{aligned}
$$