COUNT DATA REGRESSION MADE SIMPLE

A. Colin Cameron
Department of Economics, U.C.-Davis

SUMMARY

Count data regression is as simple as estimation in the linear regression model, if there are no additional complications such as endogeneity, panel data, etc. There is no reason to resort to adhoc alternatives such as taking the log of the count (with some adjustment for zero counts) and doing OLS.
The following summarizes results given, for example, in chapter 3 of Cameron, A. C. and P. K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge University Press.

THE POISSON MODEL

For *count data* y_i taking integer values 0, 1, 2, 3, ... the obvious model from statistics is the Poisson with parameter λ (the mean number of occurrences). The usual regression model sets

$$\mathsf{E}[y_i|\mathbf{x}_i] = \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\beta_1 + \beta_2 x_{2i} + \cdots \beta_k x_{ki}).$$

The regressors etc. are chosen in a manner similar to a linear regression model.
Many statistical packages estimate this model, often as a log-linear model as part of a generalized linear models module. The name log-linear model is also used as the model can be re-written as

$$\ln \mathsf{E}[y_i|\mathbf{x}_i] = \ln \lambda_i = \mathbf{x}_i'\boldsymbol{\beta} = \beta_1 + \beta_2 x_{2i} + \cdots \beta_k x_{ki}.$$

INTERPRETATION OF COEFFICIENTS

The interpretation of coefficients is different from that in the OLS model, due to the exponentiation. Some calculus and algebra show that

$$\frac{\partial \mathsf{E}[y_i|\mathbf{x}_i]}{\partial x_{ji}} = \exp(\beta_1 + \beta_2 x_{2i} + \cdots \beta_k x_{ki}) \times \beta_j = \mathsf{E}[y_i|\mathbf{x}_i] \times \beta_j.$$

Therefore, a one unit change in the j^{th} regressor leads to a change in the conditional mean by the amount $\mathsf{E}[y_i|\mathbf{x}_i] \times \beta_j$ (whereas in the linear model we would have simply β_j).
Another way of saying this is that a one unit change in j^{th} regressor leads to a **proportionate change** in $\mathsf{E}[y_i|\mathbf{x}_i]$ of β_j. (Since $\frac{\partial \mathsf{E}[y_i|\mathbf{x}_i]/\mathsf{E}[y_i|\mathbf{x}_i]}{\partial x_{ji}} = \beta_j$).
In some cases a regressor may first be transformed by the natural logarithm. Then β_j is an elasticity. For example, $\exp(\beta_1 + \beta_2 \ln x_{2i}) = \exp(\beta_1) x_{2i}^{\beta_2}$. If x_2 is a measure of exposure (such as population or time or miles travelled) we expect $\beta_2 = 1$.

STATISTICAL INFERENCE

The Poisson MLE has robustness to distributional misspecification similar to OLS in the linear regression model under normality: if $\mathsf{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, so the conditional mean is correctly specified, then the Poisson MLE estimate is consistent even if y_i is not Poisson distributed.
However, the usual Poisson MLE standard errors and t-statistics need to be adjusted. The Poisson model restricts the conditional variance to equal the conditional mean, called equidispersion. The data are called **overdispersed** if the variance exceeds the mean, and **underdispersed** if the

variance is less than the mean. Unless count data are equidispersed, teh usual Poisson MLE standard errors are wrong. This is similar to the OLS estimator being consistent if the errors are heteroskedastic, but an adjustment has to be made to the standard errors.
In statistics the standard correction (based on the generalized linear models framework) is as follows. Assume that the variance is an unknown multiple of the mean, so that $\mathsf{Var}[y_i|\mathbf{x}_i] = \alpha \times \lambda_i = \alpha \times \exp(\mathbf{x}_i'\boldsymbol{\beta})$, and data are equidispersed if $\alpha = 1$. Then the usual Poisson ML standard errors need to be multiplied by $\sqrt{\alpha}$ and t-statistics divided by $\sqrt{\alpha}$. An estimate of α is obtained after estimation of $\boldsymbol{\beta}$. Usually $\widehat{\alpha} = (n-k)^{-1} \times \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/\widehat{y}_i$ where $\widehat{y}_i = \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$.
In econometrics the standard correction is to generalize the White-heteroskedastic consistent estimate of standard errors from OLS to the Poisson. This places less structure on the form of heteroskedasticity than the model above, but in practice usually yields similar results. In Stata, for example, one uses the Poisson command with the robust option.
Such standard error corrections must be made for Poisson regression, as they can make a much bigger difference than similar heteroskedasticity corrections for OLS. Count data can be quite overdispersed, in which case uncorrected t's are much larger than the true corrected t-statistics.

ALTERNATIVE COUNT MODELS

A common more general model is the **negative binomial model**. This model can be used if data are overdispersed. It is then more efficient than Poisson, but in practice the efficiency benefits over Poisson are small. The negative binomial model should be used, however, if one wishes to predict probabilities and not just model the mean. The negative binomial model cannot be estimated if data are underdispersed.

Another more common general model is the **hurdle model**. This treats the process for zeros differently from that for the non-zero counts. In this case the mean of y_i is no longer $\exp(\mathbf{x}_i'\boldsymbol{\beta})$, so the Poisson estimator is inconsistent and the hurdle model should be used. This model can handle both overdispersion and underdispersion. Several econometrics packages include the hurdle model, which is presented, for example, in chapter 4.7 of Cameron and Trivedi.

COMPLICATIONS

Many programs handle panel data on counts. An understanding of fixed effects and/or random effects models for the linear regression models transfers over fairly simply to the count data case. Most other common complications, such as endogeneity, time series, measurement error and sample selection, require considerable skill for implementation in the count data case. These are presented in later chapters of Cameron and Trivedi.

OLS FOR NATURAL LOGARITHM OF y

A popular alternative is OLS regression of $\ln y$ on $\mathbf{x}$, so $\mathrm{E}[\ln y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, compared to count models that set $\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$.
While the log transformation for y Poisson can give something reasonably close to the normal distribution it is not as desirable, just as it is better to use logit or probit rather than OLS given binary data. And there are two problems:
1. If $y = 0$ then adhoc solutions are needed such as model $\ln(y+1)$, or model $\ln y$ except use $\ln 0.5$ when $y = 0$.
2. For prediction we want to predict $\mathrm{E}[y]$, but $\exp(\mathrm{E}[\ln y]) \neq \mathrm{E}[y]$ even though $\exp(\ln y) = y$.
One time when using $\ln y$ can be helpful is in exploratory data analysis to handle complications such as endogenous regressors for which count data software may not be readily available.