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 ad hoc 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, 2013), *Regression Analysis of Count Data*, 1st and 2nd editions, 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 specifies for individual i, $i = 1, ..., n$,

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

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

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

SIMPLE 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 is associated with 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 $\frac{\partial \mathsf{E}[y_i|\mathbf{x}_i]/\mathsf{E}[y_i|\mathbf{x}_i]}{\partial x_{ji}} = \beta_j$)
- a **percentage change** in $\mathsf{E}[y_i|\mathbf{x}_i]$ of $100 \times \beta_j$
For example, if $\beta_j = 0.05$ then a **one unit change** in the j^{th} regressor is associated with a 5% change in the conditional mean.

In some cases a regressor may first be transformed by the natural logarithm.
Then β_j is an elasticity (since, for example, $\mathsf{E}[y_i|\mathbf{x}_{2i}] = \exp(\beta_1 + \beta_2 \ln x_{2i}) = \exp(\beta_1)x_{2i}^{\beta_2}$).
For example, if $\beta_j = 0.08$ then a **one percent change** in the j^{th} regressor is associated with 0.08% change in the conditional mean.
If x_2 is a measure of exposure (such as population or time or miles travelled) we expect $\beta_2 = 1$.

MARGINAL EFFECTS

Marginal effects are different from that in the OLS model, due to the exponentiation.
The **marginal effect** of changing the j^{th} regressor is

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

Unlike linear regression (where $\text{ME}_{ji} = \beta_j$) this varies with regressor values.

The **average marginal effect** is the average over the sample of the individual marginal effects

$$\text{AME}_{ji} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \text{ME}_{ji} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{2i} + \cdots \widehat{\beta}_k x_{ki}) \times \widehat{\beta}_j.$$

Alternative marginal effects include the **marginal effect at the mean** which computes the marginal effect at the sample average values of the regressors and the **marginal effect at representative values** of the regressors.
Additional complications arise if the model includes polynomials such as a quadratic $(\exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{2i}^2 + \cdots)$ or interactions, such as $(\exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i} x_{3i} + \cdots)$.
Marginal effects in nonlinear models such as the Poisson are best computed using specialized commands such as the margins command in Stata.

STATISTICAL INFERENCE

The Poisson MLE has robustness to distributional misspecification similar to OLS in the linear regression model under normality.
In particular, 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.
Furthermore the data y_i need not be counts. Poisson regression can be used for continuous data with $y_i \geq 0$ and is well-suited to data which is right skewed and for which it is more natural to think of changes in regressors leading to proportionate changes in y_i rather than level changes in y_i.

However, the usual Poisson MLE standard errors and t-statistics need to be adjusted whenever the data are not Poisson distributed. 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, the 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.
It is absolutely essential that 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.

For **independent observations**, 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, **heteroskedastic-robust standard errors** are obtained using the Poisson command with the vce(robust) option. In R one uses the sandwich package.

STATISTICAL INFERENCE (continued)

An additional complication arises if observations are not independent but are instead **clustered**, with individual-level observations correlated within cluster and independent across clusters. For cross-section data, examples are individuals in families with correlation within family and independence across families, and individuals in regions (such as village or state) with correlation within region and independence across region. For panel data we may have observations correlated over time for a given individual but uncorrelated across individuals. The standard correction is to use **cluster-robust standard errors** that cluster on the cluster unit. In Stata, for example, one uses the Poisson command with the vce(cluster) option. In R one uses the sandwich package. Cluster-robust standard errors are also heteroskedastic-robust.

ALTERNATIVE COUNT MODELS

A commonly-used more general model is the **negative binomial model**. This model can be used if data are overdispersed (but not if they are underdispersed). It is then more efficient than Poisson.

In practice the efficiency benefits over Poisson are small. And the Poisson model is much better able to handle complications such as endogenous regressors and panel data. The negative binomial model should be used, however, if one wishes to predict probabilities and not just model the mean. The most commonly-used negative binomial model is the NB2 model.

Another common more 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 of Cameron and Trivedi.

PANEL DATA

An understanding of fixed effects and/or random effects models for the linear regression models transfers over fairly simply to the count data case.

For panel data we have data $(y_{it}, \mathbf{x}_{it})$ where i denotes the individual and t denotes time. In the simplest case of a balanced panel each individual i, $i = 1, .., n$, is observed in all time periods $t = 1, ..., T$.

The simplest approach is the **population-averaged model** that estimates the same model as for cross-section data, with

$$\mathsf{E}[y_{it}|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}).$$

Then one can use a standard Poisson command but inference needs to be based on cluster-robust standard errors with cluster unit the individual.

A **random effects model** additionally introduces within-individual correlation, with

$$\mathsf{E}[y_{it}|\mathbf{x}_{it}] = \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}); \quad \alpha_i \sim (0,\sigma_\alpha^2).$$

The **random effects estimator** requires use of more specialized Poisson commands. In practice it is best to base inference on cluster-robust standard errors with cluster unit the individual.

PANEL DATA (continued)

A **fixed effects model** again introduces α_i, but does not specify a distribution for α_i. Instead α_i is viewed as an unobservable that is possibly correlated with $\mathbf{x}_{it}$. Then

$$\mathsf{E}[y_{it}|\mathbf{x}_{it}, \alpha_i] = \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}).$$

Given correlation between α_i and $\mathbf{x}_{it}$ Poisson regression of y_{it} on $\mathbf{x}_{it}$ will yield inconsistent estimates of $\boldsymbol{\beta}$.

Instead the **fixed effects estimator** of $\boldsymbol{\beta}$ is obtained by estimation of a transformed model that has eliminated α_i. Then one needs more specialized Poisson commands. In practice it is best to base inference on cluster-robust standard errors with cluster unit the individual.

The fixed effects estimator is generally less precise as it uses only within individual variation. Also while coefficients β_j can be again interpreted as semi-elasticities estimation of marginal effects is problematic as they depend on α_i which is not estimated. The fixed effects estimator should be only used if some regressors are felt to be endogenous, being correlated with an unobserved time-invariant individual-specific effect α_i.

OTHER COMPLICATIONS

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 the Cameron and Trivedi book.

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.

Furthermore there are two potential problems:

1. If $y = 0$ then ad hoc 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 $\exp(\mathrm{E}[\ln y|\mathbf{x}]) \neq \mathrm{E}[y|\mathbf{x}]$ even though $\exp(\ln y) = y$.

One occasion for which using linear regression for $\ln y$ can be helpful is in exploratory data analysis to handle complications such as endogenous regressors because count data software may not be readily available.