Notes on the Zero-Inflated Poisson Regression Model

David Giles

Department of Economics, University of Victoria

March, 2010

The usual starting point for modeling count data (*i.e.*, data that take only non-negative integer values) is the Poisson distribution, whose p.m.f. is given as:

$$\Pr.[Y = y] = \exp(-\lambda)\lambda^{y} / y! \quad ; \quad y = 0, 1, 2, \ldots\ldots..$$

As is well-known, λ (> 0) is both the mean and variance of this distribution, so it is described as "equi-dispersed". In contrast, many data are "over-dispersed", in that their variance exceeds their mean, so this reduces the usefulness of the Poisson distribution. Allowing the variance to be modeled in turn by a gamma distribution, leads to familiar Negative Binomial (NegBin II) distribution. The latter can capture over-dispersion in the data.

In linear regression we "explain" the (conditional) mean of the dependent variable as a function of parameters and covariates, so it is natural to do the same here, and introduce covariates into the model by assigning:

$$\lambda = \exp(x'\beta) \quad ,$$

where use of the exponential function ensures that $\lambda > 0$, as is obviously required. Maximum likelihood estimation of the parameters is then straightforward, especially as the log-likelihood function is strictly concave (as it is also for the NegBin II model).

In terms of the ensuing discussion, it is important to recognize that the Poisson model, and standard variants that allow for over-dispersion, cannot describe multi-modal data. (More correctly, if λ is *integer*, then the Poisson distribution has modes at λ and $(\lambda - 1)$, but never at non-adjacent values. If λ is non-integer, the single mode occurs at $[\lambda]$.)

The *zero-inflated* Poisson (ZIP) regression model is a modification of this familiar Poisson regression model that allows for an over-abundance of zero counts in the data. This phenomenon

is widely encountered in practice. Standard references for this model include Mullahy (1986), Heilbron (1989), and Lambert (1992). Excellent discussions are also provided by Cameron and Trivedi (1998) and Winkelmann (2000).

The essential idea is that the data come from two regimes. In one regime (R_I) the outcome is always a zero count, while in the other regime (R_{II}) the counts follow a standard Poisson process. Suppose that

$$\Pr.[y_i \in R_I] = \omega_i ; \qquad \Pr.[y_i \in R_{II}] = (1-\omega_i) \quad ; \qquad i = 1, 2, \ldots, n.$$

Then,

$$\Pr.[y_i = 0] = \omega_i + (1-\omega_i)\exp(-\lambda_i)$$

and

$$\Pr.[y_i = r] = (1-\omega_i)\exp(-\lambda_i)\lambda_i^r / r! \qquad ; \qquad r = 1, 2, 3, \ldots..$$

As before, covariates enter the model through the conditional mean, λ_i, of the Poisson distribution:

$$\lambda_i = \exp(x_i'\beta) \ ,$$

where x_i' is a $(1\times k)$ vector of the i^{th} observation on the covariates, and β is a $(k\times 1)$ vector of coefficients.

Clearly,

$$E[y_i \mid x_i] = (1-\omega_i)\lambda_i$$

and

$$Var[y_i \mid x_i] = (1-\omega_i)(\lambda_i + \omega_i\lambda_i^2)$$

and so this framework also accommodates over-dispersion of the data (if $\omega_i > 0$). This over-dispersion does not arise from heterogeneity, as is the case when the Poisson model is generalized to the Negative Binomial model. Instead, it arises from the splitting of the data into the two regimes. In practice, the presence of over-dispersion may come from one or both of these sources (Mullahy, 1986; Greene, 2003, p.750).

Following Lambert (1992), it is common, and convenient, to model ω_i using a Logit model, so

$$\omega_i = [\exp(z_i'\gamma)]/[1+\exp(z_i'\gamma)] \ ,$$

where z_i is a $(1\times p)$ vector of the i^{th} observation on some covariates, and γ is a $(p\times 1)$ vector of additional parameters. Of course, the elements of z_i may include elements of x_i, and a Probit (or other) specification may be substituted for the Logit specification.

If we have n independent observations in our sample, it is readily seen that the log-likelihood function may be written as

$$\log L(\beta,\gamma) = \sum_{y_i=0} \log[\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))] + \sum_{y_i\neq 0} [y_i x_i'\beta - \exp(x_i'\beta) - \log(y_i!)] - \sum_{i=1}^{n} \log[1+\exp(z_i'\gamma)]$$

(For example, see Cameron and Trivedi, 1998, p.126.)

To code up the above log-likelihood function for use in EViews (or any other package that requires a single observation on the log-density, and then sums over all n, assuming independence of the observations), we need to take account of the different ranges of summation. The third term in the log-likelihood requires no modification as the range of summation is for all n. To deal with the ranges of summation in the first two terms, we can construct a dummy variable, D_i, which takes the value unity if $y_i = 0$, and zero otherwise. The i^{th} observation on the log-likelihood would then be coded as:

$$\log L_i(\beta,\gamma) = D_i \log[\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))] + (1-D_i)\,[y_i x_i'\beta - \exp(x_i'\beta) - \log(y_i!)] - \log[1+\exp(z_i'\gamma)]$$

Finally, note that the Negative Binomial regression model may be extended to allow for zero-inflation of the data in a corresponding and straightforward manner. In addition, Giles (2007) shows how "inflated" counts at several values of the dependent variable may be modeled using a "multinomial inflated Poisson" (MIP) model; and Giles (2010) applies the Hermite distribution to achieve the same objective.

References:

Cameron, A. C. and P. K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Giles, D. E. A. (2007), Modeling inflated count data. In L. Oxley, and D. Kulasiri, (eds.), MODSIM 2007 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, Christchurch, N.Z., 919-925.

Giles, D. E. A. (2010), Hermite regression analysis of multi-modal count data, mimeo., Department of Economics, University of Victoria.

Greene, W. E. (1993), *Econometric Analysis*, 5th ed., Prentice Hall, Upper Saddle River NJ.

Heilbron, D. (1989), Generalized linear models for altered zero probabilities and overdispersion in count data, Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco.

Lambert, D. (1992), Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 34, 1-14.

Mullahy, J. (1986), Specification and testing of some modified count data models, *Journal of Econometrics*, 33, 341-365.

Winkelmann, R. (2000), *Econometric Analysis of Count Data*, 3rd ed., Springer, Berlin.