Reference: Microeconometrics using Stata (Rev. ed.)
by A. Colin Cameron and Pravin K. Trivedi

Some Points About Count Models

Tests for the Over-dispersion for count data involve tests of the form

$$H_0 : Var(y|X) = E(y|X) \quad \text{(Equi-dispersion)}$$

$$H_1 : Var(y|X) = E(y|X) + \alpha\, E(y|X)^2 \quad \text{(over-dispersion)}$$

or equivalently,

$$H_0 : \alpha = 0$$

$$H_1 : \alpha > 0$$

In the case of over-dispersion, we should use the Negative Binomial model rather than the Poisson model.

The NBII model assumes the variance of counts formulation to be

$$Var(y|X) = \exp(X_i'\beta) + \alpha[\exp(X_i'\beta)]^2$$

In contrast, the NBI model assumes the variance of counts formulation to be

$$Var(y|X) = (1+\alpha)\exp(X_i'\beta)$$

see the stata program fertdata_Table_8-7.do

that reproduces the results for table 8.7 in the W & B textbook. Also see the programs MUS17_Count_Hurdle_TF.do and MUS17-Zero_Inflated_TF.do which conduct tests for over-dispersion and Excess zeroes.

In terms of modeling excess zeroes two types of mixture models are used. One is called the Hurdle model. The other is called the Zero-Inflated Model.

Hurdle Model

The Hurdle model relaxes the assumption that the zeroes and the positives come from the same data generating process like the Poisson or Negative Binomial distributions. The zeroes are determined by the density $f_1(\cdot)$, so that $\Pr(y=0) = f_1(0)$ and $\Pr(y>0) = 1-f_1(0)$. Suppressing regressors for notational simplicity we have as the Hurdle model density

$$f(y) = \begin{cases} f_1(0) & \text{if } y=0 \\ \dfrac{1-f_1(0)}{1-f_2(0)} f_2(y) & \text{if } y \geq 1 \end{cases}$$

This specializes to the standard one-part model if $f_1(\cdot) = f_2(\cdot)$. Although, the motivation for this model is to handle excess zeroes, it is also capable of modeling too few zeroes.

A Hurdle model has the interpretation that reflects a two-stage decision-making process, each part being a model of one decision. The two parts of the decision are functional independent. Therefore, ML estimation of the Hurdle model can be achieved by separately maximizing the two terms in the likelihood, one corresponding to the zeroes and the other to the positives. This is straightforward. The first part uses the full sample, but the second part uses only the positive count observations. (For example, see the STATA programs MUS17_Count_Hurdle_TF.do and Table8.8.do.)

For certain types of activities, such a specification is easy to rationalize. For example, in a model that explains the amount of smoking per day, the survey may include both smokers and non-smokers. One model determines whether one smokes, and a second model determines the number of cigarettes (or packs of cigarettes) smoked given that at least one is smoked.

The Hurdle model changes the conditional mean specification. Under the Hurdle model the conditional mean is

$$E(y|X) = Pr(y_1 > 0|X_1) \cdot E_{y_2>0}(y_2|y_2>0, X_2)$$

The two terms on the right are determined by the two respective parts of the model. Because of the form of the MEs, $\partial E(y|X)/\partial X_j$ is more complicated.

Zero-Inflated Models

The Zero-Inflated model was originally proposed to handle data with excess zeroes relative to the Poisson model. Like the Hurdle Model, it supplements a count density, $f_2(\cdot)$, with a binary process with a density of $f_1(\cdot)$. If the binary process takes on a value of 0, with a probability of $f_1(0)$, then $y = 0$. If the binary process takes on a value of 1, with probability of $f_1(1)$, then y takes on the count values 0, 1, 2, ... from the count density $f_2(\cdot)$. This lets zero counts occur in two ways: a realization of the binary process and a realization of the count process when the binary random variable takes on a value of 1.

Suppressing regressors for notational simplicity, the zero inflated model as a density

$$f(y) = \begin{cases} f_1(0) + \{1 - f_1(0)\} f_2(0) & \text{if } y = 0 \\ \{1 - f_1(0)\} f_2(y) & \text{if } y \geq 1 \end{cases}$$

As in the case of the Hurdle model, the probability $f_1(0)$ may be a constant or may be parameterized through a binomial model like the logit or probit. Once again, the set of variables in the $f_1(\cdot)$ density need not be the same as those in the $f_2(\cdot)$ density.

For the Poisson and NB models, the count process has the conditional mean $\exp(X_i'\beta)$ and the corresponding with-zeroes model can be shown to have the conditional mean

$$E(y|X) = \{1 - f_1(0|X_1)\} \cdot \exp(X_2'\beta)$$

where $1 - f_1(0|X_1)$ is the probability that the binary process variable equals 1. The MEs are complicated by the presence of regressors in both parts of the model, as for the hurdle model. But if the binary process does not depend on regressors, so that $f_1(0|X_1) = f_1(0)$, then the parameters, β_2, can be directly interpreted as semielasticities, as for regular Poisson and NB models.