

Count Data Models With Selectivity

Rainer Winkelmann*

University of Canterbury

Christchurch, New Zealand

May 6, 1996

Abstract

This paper shows how truncated, censored, hurdle, zero inflated and underreported count models can be interpreted as models with selectivity. Until recently, such count data models have commonly imposed independence between the count generating mechanism and the selection mechanism. Such an assumption is unrealistic in most applications, and various models with endogenous selectivity (correlation between the count and the selection equations) are presented. The methods are illustrated in an application to labor mobility where the dependent variable is the number of individual job changes during a ten year period.

1 Introduction

In this paper, I discuss the common structure of model specification and parameter estimation in count regression models with selectivity. In these models, two mechanisms are at play: one process controls the distribution of counts and another process controls selection. A broad class of recently introduced models can be analysed within this framework: truncated count models (Grogger and Carson, 1991), censored count models (Terza, 1985), hurdle count models (Mullahy, 1986), zero inflated count models (Lambert, 1992), and count models with endogenous reporting (VanPraag and Vermeulen, 1993, Winkelmann

**Address for correspondence:* Department of Economics, University of Canterbury, PB4800, Christchurch, New Zealand, Fax: +64 3 364 2635; r.winkelmann@econ.canterbury.ac.nz

and Zimmermann, 1993). A critical assumption adopted by this previous literature is the assumption of independence between the selection mechanism and the count mechanism.

Here, this assumption will be relaxed and given free to test. To fix ideas, consider a collection of counts $\{y_i^*\}$ and covariates $\{x_i\}$ and $\{z_i\}$. x_i is a $k_1 \times 1$ vector and z_i is a $k_2 \times 1$ vector. The set of covariates may be disjoint or overlapping. Assume that $y_i^*|x_i, u_i$ has Poisson distribution function

$$f(y_i^*|x_i, u_i) = \frac{e^{-\tilde{\lambda}} \tilde{\lambda}^{y_i^*}}{y_i^*!}$$

where $\tilde{\lambda} = E(y_i^*|x_i, u_i) = \exp(x_i' \beta + u_i)$ and u_i represents unobserved heterogeneity.¹ However, y_i^* is observed only conditional on selection with distribution function $f(y_i|c_i)$ where c_i is the selection variable. c_i has distribution function $f(c_i) = g(z_i' \gamma, \varepsilon_i)$ and u_i and ε_i are correlated with coefficient of correlation ρ . Endogenous selectivity is present for $\rho \neq 0$. The joint probability function of y_i and c_i is given by

$$f(y_i, c_i|x_i, z_i, u_i) = f(y_i|c_i, x_i, u_i) \int_{\varepsilon} f(c_i|z_i, \varepsilon_i) g(\varepsilon_i|u_i) d\varepsilon_i$$

and the likelihood function has the form

$$L(\theta; y_i, c_i, x_i, z_i) = \prod_{i=1}^n \int_u f(y_i, c_i|x_i, z_i, u_i) g(u_i) du_i$$

Under regularity conditions, maximizing the log-likelihood with respect to $\theta = [\beta, \gamma]$ yields a consistent and asymptotically normal estimator $\hat{\theta}$.

Models with correlation between selection and counts that fit this framework have been previously considered by Crepon and Duguet (1993), Greene (1994), and Terza (1995). The models are discussed in Section 3. Also, a new model for reporting with endogenous selectivity, or “endogenous reporting” is introduced. In this model, actual events are only reported (or counted) if a selection variable (the net benefit from reporting) is positive. An important area of application is in the study of labor mobility. Assume a researcher has access to data on the number of individual job changes during an extended period of time. Any single observed job change requires two separate events: a job offer is received,

¹While most models proposed so far in the literature come in one of two varieties, a “Poisson variety” and a “negative binomial variety”, identification problems arise here if f is the negative binomial distribution. This issue is discussed below in greater detail.

and the job offer is accepted. It is customary in search models to model offer arrival rates as a Poisson process (e.g. Mortensen, 1986. See also Ebmer, 1990). If the acceptance decision is modelled through a logit or probit model, and if the acceptance decision is correlated with arrival rates (as it should according to the standard search model) the resulting model structure for accepted offers is one of “endogenous reporting”.

The paper is structured as follows. Section 2 defines types of selectivity and demonstrates how truncated, censored, hurdle, zero inflated and underreported count models can be analysed within such a framework. Section 3 discusses models that allow for correlation between the mechanism that generates counts and the selection mechanism. Section 4 presents an application to estimating the determinants of labor mobility and Section 5 concludes.

2 Patterns of Selectivity

Throughout this analysis, selectivity is introduced in the form of a binary indicator variable c_i . c_i may indicate censoring or truncation of the dependent variable, non-reporting of an event, or, alternatively, enter the model as a regressor. We adopt the convention that truncation, censoring, or non-reporting occurs if $c_i = 0$. In what follows, two generic models for c_i can be distinguished:

$$c_i = \begin{cases} 1 & \text{if } y_i^* \in A \\ 0 & \text{if } y_i^* \notin A \end{cases} \quad (1)$$

that is, c_i is uniquely determined through the count dependent variable y_i^* . The two most commonly encountered situations are:

1. A is the set of positive integers.
2. A is the set $\{0, \dots, a\}$ where a is some positive integer.

In the second type of models, c_i depends on a $k_2 \times 1$ vector of covariates z_i and a conformable parameter vector γ , and

$$c_i = \begin{cases} 1 & \text{if } z_i' \gamma + \varepsilon_i \geq 0 \\ 0 & \text{if } z_i' \gamma + \varepsilon_i < 0 \end{cases} \quad (2)$$

The probability distribution of c_i depends on the corresponding distribution functions for y_i^* in (1) and ε_i in (2). For instance, assume that c_i is defined as in (1) and y_i^* is Poisson distributed with parameter $\lambda_i = \exp(x_i'\beta)$. For $A = \{1, 2, \dots\}$

$$P(c_i = 1) = 1 - \exp(-\lambda_i)$$

and for $A = \{0, 1, \dots, a\}$

$$P(c_i = 1) = F(a)$$

where F is the cumulative distribution function of y_i^* .

From a modelling perspective, selection models of the second type that define c_i as in (2) are the more interesting case, because they avoid the mechanical relationship between y_i^* and c_i in (2). In fact, the selection variable and the count dependent variable may be determined totally independently. Alternatively, they may be correlated as a result of common explanatory variables in z and x , or they may be correlated even conditional on x and z . The two most common assumptions for the distribution of ε_i are the normal and the logistic distributions. In the former case,

$$P(c_i = 1) = \Phi(z_i'\gamma/\sigma)$$

where Φ denotes the cumulative density function of the standard normal distribution and σ the standard deviation. For the logistic distribution, the corresponding probability is given by

$$P(c_i = 1) = \frac{\exp(z_i'\gamma)}{1 + \exp(z_i'\gamma)}.$$

2.1 Truncation and Censoring

Truncated-at-zero count data models (Creel and Loomis, 1990, Grogger and Carson, 1991) are based on a selection variable defined as in (1) with $A = \{1, 2, \dots\}$. Individuals are observed (and $y_i = y_i^*$, where y_i^* is either Poisson or negative binomial distributed) if $c_i = 1$. In the Poisson case, the observed data distribution is given by

$$\begin{aligned} g(y_i|x_i, c_i = 1) &= \frac{P(y_i^*, c_i = 1|x_i)}{P(c_i = 1|x_i)} \\ &= \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!(1 - \exp(-\lambda_i))} \quad y_i = 1, 2, \dots \end{aligned}$$

This model is appropriate whenever inclusion in the sample requires at least one occurrence. The generic situation is that of a survey in which participants are asked about the number of participation occasions. This model has been applied to the study of the number of recreational trips per year where the sample was drawn at the recreational site (Shaw, 1988), and to the number of fishing trips during the 22 week Alaskan fishing season (Grogger and Carson, 1991).

Censored count data models (See Terza (1985), Brännäs (1992), and Caudill and Mixon (1995)) are based on a selection variable defined as in (2) where $A = \{0, \dots, a\}$. Unlike for continuous data Tobit models, the type of censoring that is typically encountered in count data models is right-censoring. It arises frequently in survey questionnaires where the highest category is “x or more” counts (See e.g. Merkle and Zimmermann, 1992). It holds that

$$y_i = \begin{cases} y_i^* & \text{for } c_i = 1 \\ a & \text{for } c_i = 0 \end{cases} \quad (3)$$

and the probability function of observed counts y is given by

$$g(y_i|x_i, c_i) = f(y_i)^{c_i} [1 - F(a)]^{1-c_i}$$

2.2 Hurdle and Zero Inflated Count Data Models

Both hurdle and zero inflated count data models address a situation that is frequently encountered in applications using count data: the observed data display a higher fraction of zeros, or non-occurrences, than can possibly be explained through any fitted standard (Poisson or negative binomial) regression model. In this situation, the selection variable c_i allows for a separate treatment of zeros and strictly positive outcomes. There are two ways of doing this. In the hurdle Poisson or hurdle negative binomial models (Mullahy, 1986), the observation mechanism

$$y_i = \begin{cases} 0 & \text{if } c_i = 0 \\ y_i^* > 0 & \text{if } c_i = 1 \end{cases}$$

gives rise to a probability function

$$g(y_i) = (1 - p_i)^{1-c_i} [p_i \tilde{f}(y_i)]^{c_i}$$

where $\tilde{f}(y^*) = f(y^*)/(1 - f(0))$ is a truncated-at-zero probability function and $p_i = P(c_i = 1)$. It is not meaningful to define c_i as in (1) because it would mean that the hurdle model is identical to the original model $f(y_i^*)$. While selection mechanism (2) could be used theoretically, it has been common practice to define directly $P(c_i = 1) = 1 - f(0; x_i' \beta_2)$ and $P(y_i^*) = f(y_i^*; x_i' \beta_1)$. In this way, the standard count model can be tested against the hurdle model by testing the parametric restriction $\beta_1 = \beta_2$ using standard methods. Note also, that the common covariates x_i introduce correlation between the hurdle step and the count mechanism,

The model is relatively easy to estimate since the likelihood function can be factored into two separate blocks, and this model has gained recent popularity with applications in health economics (Pohlmeier and Ulrich, 1995), labor economics (Arulampalam and Booth, 1996) and transportation economics (Dionne, Artis and Guillen, 1995), among others.

The zero inflated Poisson model is very similar to the hurdle model. Here

$$y_i = \begin{cases} 0 & \text{if } c_i = 0 \\ y_i^* & \text{if } c_i = 1 \end{cases}$$

with probability function

$$g(y_i) = (1 - p_i)^{1 - c_i} + p_i f(y_i)$$

The difference is that for $c_i = 1$, $y_i = y_i^*$ for the full range of y_i^* and not just for strictly positive values. Hence, in the zero inflated count data models there are two types of zeros: one type is obtained as $c_i = 0$; the other as $c_i = 1$ and $y_i^* = 0$. Which of the two models is more appropriate will depend on the particular application. Lambert (1992) introduced the zero inflated model together with selection equation (2), assuming that ε_i has a logistic distribution. She has applied this model to the occurrence of defects in manufacturing. Economic applications of zero inflated models are given in Greene (1994) on the frequency of loan defaults, in Crepon and Duguet (1994) on patents, and in Grootendorst (1995) on prescription drug utilization. The Lambert formulation of the model is restrictive in the sense that it assumes independence between the selection and the count processes. This assumption is relaxed by a class of models for endogenous selectivity to be discussed below.

2.3 Under-Reporting

In this class of models, selection occurs through the reporting mechanism. As before, y_i^* has distribution function $f(y_i^*)$. c_i^* gives the utility from reporting a particular event j (assumed to be constant for all events j):

$$c_i^* = z_i' \gamma + \varepsilon_i.$$

and an event j is reported, and $c_i = 1$, if

$$c_i^* > a.$$

where a denotes some threshold value. It follows that the number of observed counts:

$$y_i = \sum_{j=1}^{y_i^*} c_i.$$

is obtained through a convolution operation, and we can write in general

$$f(y_i) = \sum_{y_i^*=y_i}^{\infty} f(y_i^*) \frac{y_i^*!}{y_i!(y_i^* - y_i)!} (1 - F(\tilde{a}))^{y_i} (F(\tilde{a}))^{y_i^* - y_i} \quad (4)$$

where $F(\tilde{a}) = P(\varepsilon_i < a - z_i' \gamma)$. The model is completed by making specific assumptions on $f(y_i^*; x, \beta)$ and $F(a)$. The Poisson-logistic model (Winkelmann and Zimmermann, 1993) is obtained as a special case for $a = 0$, $F = \exp(z_i' \gamma) / (1 + \exp(z_i' \gamma))$ and $y_i^* \sim \text{Poisson}$ with mean $E(y_i^* | x_i) = \exp(x_i' \beta)$. Under these assumptions, (4) reduces to (See Winkelmann and Zimmermann, 1993, or Mukhopadhyay and Trivedi, 1994, for details)

$$f(y_i | x_i, z_i) = \frac{\exp(-\tilde{\lambda}_i) \tilde{\lambda}_i^{y_i}}{y_i!}$$

where

$$\tilde{\lambda}_i = \frac{\exp(x_i' \beta + z_i' \gamma)}{1 + \exp(z_i' \gamma)} \quad (5)$$

This Poisson model with modified mean function can be easily estimated by maximum likelihood. Analytical first and second derivatives are provided in Winkelmann and Zimmermann (1993). Alternatively, assume that y_i^* is negative binomial distributed. Then

$$f(y_i | x_i, z_i) = \frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha) \Gamma(y_i + 1)} \left(\frac{\alpha}{\tilde{\lambda}_i + \alpha} \right)^\alpha \left(\frac{\tilde{\lambda}_i}{\tilde{\lambda}_i + \alpha} \right)^{y_i}$$

where $\tilde{\lambda}_i$ is given in (5).

In both cases, interdependencies between y_i^* and c_i^* are modelled in terms of common explanatory variables. As before, the two processes are assumed stochastically independent, conditional on x and z .

In an independent study, VanPraag and Vermeulen (1993) have used this model for the study of consumer purchase behavior. In their application, purchases are reported, together with their money value, as long as they exceed a minimum threshold value a . Since both the number of events and the money values are observed, VanPraag and Vermeulen estimate the parameter vector $\theta = [\beta, \gamma]$ from the joint distribution function of $c_i = (c_{i1}, \dots, c_{iy})$ and y_i which is given by

$$\begin{aligned} g(y_i, c_i; \theta | x_i, z_i) &= \prod_{i=1}^{y_i} \frac{f(c_i | z_i' \gamma)}{1 - F(\tilde{a})} \sum_{y_i^* = y_i}^{\infty} f(y_i^*) \frac{y_i^*!}{y_i!(y_i^* - y_i)!} (1 - F(\tilde{a}))_i^y (F(\tilde{a}))_i^{y_i^* - y_i} \\ &= \prod_{i=1}^{y_i} f(c_i | z_i' \gamma) \sum_{y_i^* = y_i}^{\infty} f(y_i^*) \frac{y_i^*!}{y_i!(y_i^* - y_i)!} (F(\tilde{a}))_i^{y_i^* - y_i} \end{aligned}$$

where $f(y_i^*)$ is a Poisson or negative binomial probability function and $f(c_i)$ is the normal density.

3 Endogenous Selectivity

The essential elements for count models with endogenous selectivity have been provided in the introduction: the mean function

$$E(y_i^* | x_i, u_i) = \exp(x_i' \beta + u_i)$$

exhibits unobserved heterogeneity, that is, omitted regressors, and consequently the conditional mean varies more than can be explained by the observed covariates. A latent process

$$c_i^* = z_i' \gamma + \varepsilon_i$$

generates a binary indicator variable c_i where

$$c_i = \begin{cases} 1 & \text{if } c_i^* \geq 0 \\ 0 & \text{if } c_i^* < 0 \end{cases} \quad (6)$$

The main pay-off to this generalized framework is that correlation between y_i^* and c_i can be modelled through a joint distribution for $f(u_i, \varepsilon_i)$. Assume that u_i and ε_i are jointly normal distributed with mean vector zero and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix}$$

where ρ is the coefficient of correlation and σ^2 the variance of u_i . The variance of ε_i is normalized to one, since the selection equation (6) identifies γ only up to a scale factor. Note that this model implies a marginal regression $E(y_i^*|x_i) = \exp(x_i'\beta)v_i$, where v_i has a lognormal distribution with mean $\exp(0.5\sigma^2)$ and variance $\exp(2\sigma^2) - \exp(\sigma^2)$.

This set-up allows for endogenous selection in the sense that unobserved factors affecting c_i also affect y_i^* . Ignoring this correlation will lead to a misspecified model with the possibility of inconsistent parameter estimates. The effects of selectivity in count data models are similar to those found for continuous data (See e.g. Heckman, 1976), and the corresponding models will be presented in the next sections. Existing models for incidental truncation and censoring, and endogenous switching are summarized, and a new model with selective endogenous reporting is derived.

Before proceeding, it is necessary to consider the implications of alternative distributional choices for y_i^* in this class of models. It has been emphasized before that all previous models can be specified either with a Poisson or with a negative binomial distribution. By contrast, the negative binomial distribution is no longer suitable in the context of endogenous selectivity, since the resulting model suffers from overparametrization. To illustrate this point, consider a negative binomial distribution with $E(y_i|x_i, v_i) = \lambda_i v_i$ where $\lambda_i = \exp(x_i'\beta)$ and $v_i = \exp(u_i)$ and u_i represents additional unobserved heterogeneity. (Also, assume that $E(v_i) = 1$ which is not restrictive as long as x contains an intercept.) In the negative binomial model, $\text{Var}(y_i|x_i, v_i) = \lambda_i v_i + \alpha(\lambda_i v_i)^2$ where α is an additional dispersion parameter (See Cameron and Trivedi, 1986). Marginalizing with respect to v yields

$$E(y_i|x_i) = \lambda_i$$

and

$$\text{Var}(y_i|x_i) = E(\text{Var}(y_i|v_i)) + \text{Var}(E(y_i|v_i))$$

$$= \lambda_i + \lambda_i^2(\sigma_v^2 + \alpha\sigma_v^2 + \alpha)$$

Hence, the first two moments are not sufficient to identify both α and σ_v^2 ; while higher order moments can solve this problem, this approach is unsatisfactory and will cause numerical difficulties in practical applications.

3.1 Incidental Censoring and Truncation

A model for endogenous censoring was introduced by Crepon and Duguet (1994). It has the same structure as the zero inflated Poisson model, augmented by correlated error terms. In particular

$$y_i = \begin{cases} y_i^* & \text{if } c_i = 1 \\ 0 & \text{if } c_i = 0 \end{cases}$$

and $c_i = 1$ for $\varepsilon_i > -z_i'\gamma$. Standard results on conditioning in the bivariate normal distribution can be used to obtain

$$\begin{aligned} \Phi^* &= P(c_i = 1|u_i, z_i) = P(\varepsilon > -z_i'\gamma|u_i) \\ &= \Phi \left[\frac{z_i'\gamma}{\sqrt{1-\rho^2}} + \frac{\rho u_i/\sigma}{\sqrt{1-\rho^2}} \right] \end{aligned} \tag{7}$$

where Φ is the cumulative density function of the standard normal distribution. Furthermore, for $y_i^* \sim \text{Poisson}$ with $\tilde{\lambda}_i = \exp(x_i'\beta + u_i)$ the probability function of y_i is given by

$$f(y_i|u_i, x_i, z_i) = \Phi^* \frac{\exp(-\tilde{\lambda}_i) y_i^{\tilde{\lambda}_i}}{y_i!} + (1 - \Phi^*)(1 - c_i)$$

This probability function depends on the unobserved u_i , and the observed data distribution is given by

$$f(y_i|x_i, z_i) = \int_{-\infty}^{\infty} \left\{ \Phi^* \frac{\exp(-\tilde{\lambda}_i) y_i^{\tilde{\lambda}_i}}{y_i!} + (1 - \Phi^*)(1 - c_i) \right\} f_u(u_i|z_i) du_i$$

While marginalizing with respect to u_i does not lead to a closed form solution, Crepon and Duguet suggest a feasible simulation method due to Gourieroux and Monfort (1993).

Alternatively, Gauss-Legendre quadrature can be used for an exact evaluation of the integral. Crepon and Duguet apply their model to a study of R&D productivity, where y_i^* gives the number of discoveries, y_i the number of patents applied for, and $c_i = 1$ if the firm decided to apply for patents in general.

A model for incidental truncation is proposed in Greene (1994). Greene models

$$y_i = \begin{cases} y_i^* & \text{if } c_i = 1 \\ \text{unobserved} & \text{if } c_i = 0 \end{cases}$$

where c_i and y_i^* are determined as above. In analogy to the conditional expectation in the bivariate normal model, where

$$E(y|z > 0) = \mu_y + \rho\sigma_y \left(\frac{\phi(\mu_z/\sigma_z)}{\Phi(\mu_z/\sigma_z)} \right)$$

and sample selection can be interpreted as an omitted variable problem. Greene suggests the “mean corrected” count data model

$$E(y_i|x_i, c_i = 1) = \exp \left(x_i'\beta + \tau \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)} \right)$$

which has to be estimated by a two-step procedure in analogy to Heckman (1979). A probit regression provides a consistent estimator $\hat{\gamma}$. The predicted selectivity term $\phi(z_i'\hat{\gamma})/\Phi(z_i'\hat{\gamma})$ is then used as a regressor in a second step count data regression. Greene applies his model to a study of the determinants of credit card default, where c_i indicates credit card approval.

3.2 Endogenous Switching

A model for endogenous switching has recently been proposed by Terza (1995). Terza considers the situation of an endogenous binary regressor c_i , which may measure, for instance, program participation. c_i is determined as in (6) and $f(y_i|x_i, c_i, u_i)$ denotes the conditional probability function of y_i with mean

$$E(y_i|x_i, c_i, u_i) = \exp(x_i'\beta + \alpha c_i + u_i)$$

The joint probability function of y_i and c_i is given by

$$f(y_i, c_i|x_i, z_i) =$$

$$\int_{-\infty}^{\infty} f(y_i | c_i, x_i, z_i, u_i) f_u(u | x_i, z_i) [c_i \Phi^*(u_i) + (1 - c_i)(1 - \Phi^*(u_i))] du_i$$

where Φ^* is defined as in (7). Computations of the integral using quadrature or other simulation methods provides no major difficulties. Terza applies this model to the study of trip frequencies, where the endogenous dummy variable is car ownership.

3.3 Endogenous Reporting

The anatomy of the model was laid out in Section 2.3. The principal shortfall of the previous approach is the assumption of independence between the count process and the binary reporting outcome. Consider, for instance, the study by Winkelmann and Zimmermann (1993), where the model is applied to data on labor mobility. y_i^* gives then the (unobserved) number of job offers, $\lambda_i = \exp(x_i' \beta)$ the offer arrival rate, p_i the acceptance probability and y_i the (observed) number of accepted offers. The explicit assumption is that a) the offer arrival rate is a deterministic function of observed covariates, and b) the offer arrival rate is independent of the acceptance probability. Yet, it is unreasonable to assume that all relevant variables are observed in practice. Moreover, economic models of efficient job search predict that the reservation wage depends on the offer arrival rate and hence a correlation between the two (See Mortensen, 1986, for instance). Therefore, a more general model that allows for endogenous under-reporting is required. Such a model is now introduced.

Let $y_i^* | u_i \sim$ have a count data distribution with mean

$$E(y_i^* | x_i, u_i) = \exp(x_i' \beta + u_i) . \quad (8)$$

As before, an event j is reported and $c_i = 1$ if the net utility from doing so is positive, i.e.

$$c_i^* = z_i' \gamma + \varepsilon_i > 0 .$$

and assume that u_i and ε_i are jointly normal distributed with correlation ρ . The number of reported counts is given by

$$y_i = \sum_{j=1}^{y_i^*} c_i .$$

To derive the probability function of y_i , consider first the case where u_i is given. As before

$$P(c_i = 1|u_i) = \Phi^*(u_i)$$

where Φ^* is defined as in (7). Moreover, conditional on u_i , x_i and z_i , c_i and y_i^* are independent. Assume that $y_i^*|u_i$ is Poisson distributed. It follows directly from results in section 2.3 that $y_i|u_i$ is Poisson distributed with mean

$$\tilde{\lambda}_i = \exp(x_i'\beta + u_i) \times \Phi^*(u_i) \quad (9)$$

while $y_i|x_i, z_i$ has distribution

$$g(y_i|x_i, z_i) = \int_{-\infty}^{\infty} \frac{\exp(-\tilde{\lambda}_i(u_i))\tilde{\lambda}_i(u_i)^{y_i}}{y_i!} f_u(u|z_i) du_i \quad (10)$$

or, in explicit notation

$$\begin{aligned} g(y_i|x_i, z_i; \beta, \gamma, \rho, \sigma) &= \int_{-\infty}^{\infty} \exp \left[-\exp(x_i'\beta + u) \Phi \left(\frac{z_i'\gamma - \rho u/\sigma}{\sqrt{1 - \rho^2}} \right) \right] \\ &\quad \times \left[\exp(x_i'\beta + u) \Phi \left(\frac{z_i'\gamma - \rho u/\sigma}{\sqrt{1 - \rho^2}} \right) \right]^{y_i} \times \frac{1}{y_i! \sigma} \phi(u/\sigma) du \end{aligned} \quad (11)$$

The parameters of the model, β , γ , ρ , and σ are estimated by maximum likelihood. The resulting log-likelihood function involves simple integrals that can be evaluated by Gauss-Legendre quadrature. The following applications use a quadrature routine provided by the programming package GAUSS that is based on 40 quadrature points over an area of +/- 8 standard deviations of u_i . The maximization has been performed with the GAUSS maxlik routine.

The model is quite general and encompasses a variety of interesting special cases that can be tested using parametric restrictions. For $\rho = 0$ the selection and count equations are independent. For $\rho = 0$ and $\sigma = 0$, the model reduces to a version of the Poisson-logistic regression model in Winkelmann and Zimmermann (1993) where the logit type expression for the reporting probability is replaced by a probit type expression. Positive values for σ indicate unobserved heterogeneity in the count regression. In particular, the implicit variance function for y_i^* is

$$\text{Var}(y_i^*|x_i) = \lambda_i + \alpha \lambda_i^2$$

where $\alpha = \exp(2\sigma_u^2) - \exp(\sigma_u^2)$.

4 An Application to Labor Mobility

The following empirical analysis of the determinant of labor mobility uses data from the *German Socio-Economic Panel* (See Wagner, Burkhauser and Behringer, 1993.) The subsample includes 1962 men between the age of 25 and 50 in 1974. The dependent variable is the number of direct job changes (*i.e.* changes without an intervening spell of unemployment) during the ten year period 1974-1984, which is collected retrospectively in the second (1985) wave of the panel. The age and gender selection facilitates the interpretation of the results by minimizing potential distortions due to labor force participation decisions (which cannot be explicitly accounted for given the limited information in the data).

As previously argued, the endogenous-reporting model is well suited to capture the structure of the labor mobility process. Worker receive job offers over time and consecutively accept or reject. The proposed model allows for a separate modeling of the offer function and the acceptance process. A particular offer is accepted if the wage exceeds the reservation wage. The reservation wage in turn measures the value of alternative uses of time. The literature has identified human capital variables like years of schooling and years of experience as main determinants of reservation wages (See Devine and Kiefer, 1991).

In contrast, the empirical literature on the determinants of offer arrival rates is thin. Ebmer (1990) is one of the few studies that has direct information on the number of job offers for individuals. However, his data come from official records by the Austrian employment exchange, a government agency, and measure arrival rates for unemployed individuals. For this group, Ebmer finds that the most important determinants of offer arrival rates are labor demand conditions like unemployment and vacancy ratios rather than individual characteristics.

Winkelmann (1994) provides a more comprehensive discussion of the substantive economic issues involved in modelling labor mobility. Here, I concentrate on issues of econometric methodology. Besides the human capital variables and a constant, five indicator variables capturing various aspects of employment and individual life situation are used. These are: union, German, single, qualified white collar worker, ordinary white collar

worker, qualified blue collar worker (ordinary blue collar worker is the reference category). The indicator variables are coded as 1 if true. The dependent variable, the number of direct job changes during the ten year period, has mean 0.54 and variance 1.18. 68 percent of all workers have recorded no job change during the period.

Four models are fitted to these data: a standard Poisson model as a benchmark, a Poisson model with normally distributed unobserved heterogeneity, the Poisson model with endogenous reporting and the zero inflated Poisson model with endogeneity. The results are given in Table 1.

The results show that neither occupational nor family status are statistically significant. However, union members, Germans, and workers with more experience have a significantly lower mobility rates as others. These results hold independently of the specification. However, the models are performing very differently in explaining the observed variation in the dependent variable. For instance, the Poisson model is rejected against the Poisson-log normal model with a likelihood ratio test statistic of 353.2. This is to be expected; it mirrors the common rejection of the Poisson model against the negative binomial (negbin) model in the presence of overdispersion. The model estimated here assumes a log-normally distributed multiplicative error rather than a (algebraically more convenient) gamma distributed error underlying the negbin model. The negbin model has been estimated as well in its two parametrizations giving log-likelihood values of -1978.6 and -1973.3, respectively (for a description of the models, see Cameron and Trivedi, 1986; the complete regression results are available upon request). It is therefore interesting to note that the negbin model, despite its widespread popularity, fits less well to the data than the Poisson-lognormal model used here. This suggests that the individual specific unobserved effect has a normal distribution rather than the exponential of a gamma distribution.

The third column of Table 1 give the results for the endogenous reporting model. In the chosen specification the constant, union status, nationality and the occupational indicators form the x -vector that affects the offer arrival rates. The human capital variables and marital status form the z -vector that determines acceptance. The interesting result is that the hypothesis of no correlation between offer arrival rates and acceptance decision ($\rho = 0$) can be rejected. In fact, the two are negatively correlated with $\hat{\rho} = -0.3$ (with standard error 0.12), as predicted by search theory. Moreover, the offer arrival equation displays

unobserved heterogeneity; the variance of the error term is estimated as $0.67^2 = 0.41$ with standard error 0.13, implying a variance of the multiplicative log-normal error of 0.76.

For the purpose of comparison, the fourth column of Table 1 gives the estimation results for an alternative model with endogenous selectivity, the zero-inflated Poisson model. This model has been recommended frequently on the grounds that it is able to accommodate excess zeros as observed in these data. However, in the present context, the zero inflated Poisson model is inferior to the endogenous reporting model. While the two models are not nested, they have the same number of parameters and a direct comparison of the log-likelihood clearly favors the reporting model. It does not surprise that in the zero inflated model, the estimated correlation between the count and the probability of observing at least one count is positive. However, it is interesting to note that the probability of observing a positive number of job changes (at least one) increases with years of schooling, while the years of schooling variable is insignificant in the other three models, where it either influences the overall number of counts (models (1) and (2)) or the acceptance probability. One possible explanation is that years of schooling influences job changes in a nonlinear way: More years of schooling increases the probability of a small (but positive) number of job changes, while at the same time reducing the probability of a high number of job changes.

5 Conclusions

This article presents a framework for the analysis of selectivity in econometric count data models. The key issue is whether or not the selection equation is allowed to be correlated with the count equation. Much of the previous literature on count data models with selection, such as truncated, censored, hurdle, zero inflated or underreported count models, has imposed an independence assumption. The validity of this assumption is questioned and relaxed. Correlation between the two processes is modelled through a joint normal distribution and a new model for endogenous reporting is applied to a dataset on labor mobility.

Table 1. Regression Results for the Number of Direct Job Changes^a

Variable	M O D E L			
	(1) Poisson	(2) Poisson log-normal	(3) Endogenous Reporting	(4) Zero-infl. Poisson
Constant ^b	0.501 (3.167)	0.102 (0.294)	0.755 (4.745)	0.036 (0.165)
Union ^b	-0.292 (-4.499)	-0.314 (-3.537)	-0.311 (-3.467)	-0.332 (-3.750)
German ^b	-0.368 (-4.842)	-0.369 (-2.925)	-0.354 (-3.642)	-0.493 (-4.529)
Qualified White Collar ^b	0.067 (0.513)	0.017 (0.105)	-0.007 (-0.164)	-0.269 (-1.882)
Ordinary White Collar ^b	0.184 (1.255)	0.220 (1.106)	0.192 (1.028)	-0.052 (-0.268)
Qualified Blue Collar ^b	0.147 (1.794)	0.095 (0.881)	0.103 (1.067)	-0.037 (-0.460)
Years of Schooling*10 ⁻¹	-0.137 (-1.005)	-0.139 (-0.577)	-0.092 (-0.868)	1.191 (4.416)
Experience*10 ⁻¹	-0.769 (-6.928)	-0.849 (-4.335)	-0.763 (-6.956)	-0.763 (-3.740)
Experience ² * 10 ⁻²	0.119 (3.269)	0.128 (2.147)	0.126 (3.650)	0.106 (1.866)
Single	-0.049 (-0.460)	-0.095 (-0.641)	-0.100 (-0.787)	-0.171 (-0.774)
σ		1.013 (22.104)	0.674 (5.370)	0.887 (10.492)
ρ			-0.295 (-2.528)	0.455 (3.460)
Log likelihood	-2044.5	-1867.8	-1865.0	-1889.0

Notes:

^a Data are from the German Socio-Economic Panel. Number of Observations: 1962.

^b Variable marked with a ^b are among the x -variables in Models (3) and (4). Variables not marked with a ^b are among the z variables.

References

- Arulampalam, W. and A.L. Booth 1996, "Who gets over the training hurdle? A study of Training experiences of young men and women in Britain," mimeo, University of Essex.
- Brännäs, K. 1992, "Limited dependent Poisson regression," *The Statistician* 41, 413-423.
- Cameron, A.C. and P.K. Trivedi 1986, "Econometric models based on count data: comparisons and applications of some estimators and tests," *Journal of Applied Econometrics* 1:29-53.
- Caudill, S.B. and F.G. Mixon 1995, "Modeling household fertility decisions: Estimation and testing of censored regression models for count data," *Empirical Economics* 20, 183-196.
- Creel, M.D. and J.B. Loomis 1990, "Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California," *American Journal of Agricultural Economics* 72(2):434-441.
- Crepon, B. and E. Duguet 1994, "Research and development, competition and innovation: Pseudo maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity," INSEE working paper.
- Devine, T.J. and N.M. Kiefer 1991, *Empirical Labor Economics: The Search Approach*, Oxford University Press.
- Dionne, G., M. Artis and M. Guillen 1995, "On the repayment of personal loans under asymmetrical information: a count data model approach," Working Paper No. 9528, University of Montreal.
- Ebmer, R. 1990, "Placement service and offer-arrival rates," *Economics Letters* 34, 289-294.
- Greene, W.H. 1994, "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models", Stern School of Business, New York University, Working Paper No. 94-10.
- Grogger, J.T. and R.T. Carson 1991, "Models for truncated counts," *Journal of Applied Econometrics* 6:225-238.
- Grootendorst, P.V. 1995, "A comparison of alternative models of prescription drug utilization," *Health Economics* 4, 183-198.
- Heckman, J.J. 1976, "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement* 5/4, 475-492.
- Heckman, J.J. 1979, "Sample selection bias as a specification error," *Econometrica* 47, 153-161.

- Lambert, D. 1992, "Zero-inflated Poisson regression with an application to defects in manufacturing," *Technometrics* 34, 1-14.
- Merkle, L. and K.F. Zimmermann 1992, "The demographics of labor turnover: A comparison of ordinal probit and censored count data models," *Recherches Economiques de Louvain*, 58:283-306.
- Mortensen, D.T. 1986, "Job search and labor market analysis," in: O.Ashenfelter and R. Layard (eds.) *Handbook of Labor Economics* Vol.II, Amsterdam, North-Holland.
- Mukhopadhyay, K. and P.K. Trivedi 1994, "Estimation and inference in count regression with under-recording," mimeo, Indiana University.
- Mullahy, J. 1986, "Specification and testing in some modified count data models," *Journal of Econometrics* 33:341-365.
- Ozuna, T. and I.A. Gomez 1995, "Specification and testing of count data recreation demand functions," *Empirical Economics* 20, 543-550.
- Pohlmeier, W. and V. Ulrich 1995, "An econometric model of the two-part decision making process in the demand for health care," *Journal of Human Resources* 30, 339-361.
- Shaw, D. 1988, "On-Site samples regression," *Journal of Econometrics* 37:211-223.
- Terza, J.V. 1985, "A Tobit type estimator for the censored Poisson regression model," *Economics Letters* 18, 361-365.
- Terza, J.V. 1995, "Estimating count data models with endogenous switching and sample selection," Department of Economics, Pennsylvania State University, Working Paper No. 4-95-14.
- VanPraag, B.M.S. and E.M. Vermeulen 1993, "A count-amount model with endogenous recording of observations," *Journal of Applied Econometrics* 8, 383-395.
- Wagner, G.G., R.V. Burkhauser and F. Behringer 1993, "The English language public use file of the German Socio-Economic Panel," *Journal of Human Resources* 28, 429-433.
- Winkelmann, R. and K.F. Zimmermann 1993, "Poisson-Logistic Regression," Department of Economics, University of Munich, Working Paper No. 93-18.
- Winkelmann, R. 1994, *Count Data Models - Econometric Theory and an Application to Labor Mobility*, Springer (Lecture Notes in Economics and Mathematical Systems 410).
- Winkelmann, R. and K.F. Zimmermann 1995, "Recent developments in count data modeling: Theory and applications," *Journal of Economic Surveys* 9: 1-24.