Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models

by

William H. Greene
Department of Economics
Stern School of Business
New York University
44 West 4th Street
New York, NY 10012-1126

Phone 212-998-0876
Internet: wgreene@stern.nyu.edu

March, 1994

Abstract: We present several modifications of the Poisson and negative binomial models for count data to accommodate cases in which the number of zeros in the data exceed what would typically be predicted by either model. The excess zeros can masquerade as overdispersion. We present a new test procedure for distinguishing between zero inflation and overdispersion. We also develop a model for sample selection which is analogous to the Heckman style specification for continuous choice models. An application is presented to a data set on consumer loan behavior in which both of these phenomena are clearly present.

JEL Classification: C12, C13 - Econometric and Statistical Methods and Estimation
Field Designation: Cross Section Econometrics

We are grateful for the able research assistance of Jin Yoo. Errors in the paper are our responsibility.

1 Introduction

The Poisson regression model forms the basis for a large proportion of the received empirical literature involving discrete outcomes and count data. However, real data considerations and the shortcomings of the basic model, itself, have led researchers to employ a wide variety of alternative specifications. Modifications of the Poisson model have been suggested to accommodate:

· over- and underdispersion, which is a violation of the Poisson restriction that the variance of the observed random variable equal its mean,

· unobserved individual heterogeneity, for example, in panel data (Hausman, et al. (1984)), which mandates the introduction of a disturbance term into the Poisson specification much like that which appears in conventional regression models, and which induces overdispersion,

· `non-poissonness,' (Johnson and Kotz (1969)) which is reflected in an overabundance or underabundance of certain specific values, usually zero.[1]

Another issue which arises occasionally (e.g., Heilbron (1989), Smith (1990)), but remains to be examined in detail is an extension of the Poisson regression model to

· sample selection, which will likely produce distortions in the inference drawn from count data by conventional methods similar to those which arise in the analysis of continuous choice models.

The literature on the Poisson regression model often discusses separately *specification* and *estimation* of the model and its variants and *specification testing* in the context of the basic model. The focus of this paper is primarily the first of these. We will present two modifications of the Poisson model, a model for handling `excess zeros,' and a specification for modeling sample selection in the spirit of Heckman (1979). Since excess zeros will masquerade as overdispersion, we are also interested in the first two points. Our first model extends an existing literature. In addition, we will present a method of testing this extension of the model against the base (Poisson or negative binomial) case. The test procedure can also be applied to the problem of testing for overdispersion in the Poisson model, so, in passing, we will add another method to the set of tools that have already been proposed for this problem. Our second model is a sample selection model which does not appear to have been treated elsewhere.

[1]The third of these could show up as if it were the first or second. Terza and Wilson (1990) introduce a variant of our zero altered Poisson model specifically to allow for overdispersion by (at least in principle) disconnecting the Poisson mean and variance.

This paper proceeds as follows: Section 2 will review some of the existing literature on the Poisson model and tie together some widely dispersed but related contributions. We begin with a cursory presentation of the basic Poisson specification. Section 3.1 will describe the specification and estimation of our `zero inflated Poisson' (*ZIP*, Lambert (1992)) regression. Restrictive variants of the *ZIP* model have appeared elsewhere in the literature. In addition to presenting a more general model, we will propose a new method of testing the specification against the basic Poisson model. Section 3.2 will describe a framework for analyzing sample selection in the context of the Poisson model. Although this model has been hinted at in various places, it appears not to have been formalized previously. This section will detail a sample selection model and provide methods of parameter estimation and computation of appropriate asymptotic covariance matrices for the estimates. The *ZIP* and sample selection models are combined in Section 3.3. In section 4, we present an application of the techniques to an aspect of consumer behavior, default on credit card loans. We will use the Poisson model to examine the number of major derogatory reports to a credit reporting agency for a group of credit card applicants. The overwhelming majority of applicants have `clean' (at least in this respect) credit histories, so there is a prevalence of zeros in the data. Hence the *ZIP* model is appropriate. We will apply the model to a general population and to a heavily screened subpopulation (those whose applications for credit were accepted), in which the screen clearly produces the sort of nonrandomness found in settings in which Heckman's selection model is usually applied. The statistical results suggest unambiguously that models more general than the basic Poisson regression are called for. Conclusions are drawn in Section 5.

2 Poisson and Negative Binomial Regression Models

The Poisson model arises in many contexts as the probability distribution for the discrete, nonnegative count of the number of occurrences of an event.[2] Applications of the basic model include:

· the number of failures of electronic components per unit of time,

· the number of individuals arriving at a serving station (bank teller, gas station, cash register, etc.) within a fixed interval,

· the number of homicides per year (Grogger (1990a)),

· the number of patents applied for and received (Hausman, et al. (1984)),

and so on.[3] The unconditional probability distribution for a Poisson random variable is given by

$$\text{Prob}[Y = y_i | t_i] = p(y_i | t_i) = \frac{e^{-(t_i\lambda)} (t_i \lambda)^{y_i}}{y_i!}, y_i = 0,1,... \quad (2.1)$$

where λ is the mean occurrence rate per unit of time and t_i is the length of the interval over which y_i is observed.[4] For our purposes, no generality will be lost by assuming that the time interval is one unit for each observation.[5] It is easily shown that the unconditional mean of y_i given a unit length interval is λ. The model lends itself conveniently to a regression framework by defining the conditional mean function,

$$E[y_i | x_i, t_i=1] = \lambda_i = e^{\beta' x_i},$$

where here and in what follows, $\boldsymbol{x}_i$ will denote the full set of regressor variables for y_i.[6] The exponentiation

[2]See Johnson and Kotz (1969) for an extensive survey on the unconditional model. A useful overview is given by Cameron and Trivedi (1986).

[3]Other applications in the literature include Gray and Jones (1991) (citation counts), King (1986) (count data in political science), Kostiuk and Follman (1989) (success rates of military recruiters), Papke (1986) (industry "births" in different states), and Portney and Mullahy (1986) (air quality and the incidence of respiratory illness). See, as well, Agresti (1984), Arvan (1989), Cooil (1991), Coughlin, et al. (1988), Flowerdew and Aitken (1982), Frome (1983), Frome et al. (1973), Gart (1964), and Okoruwa, et al. (1988) for a variety of specifications and uses of the Poisson model.

[4]See, e.g., Hoffman and Milligan (1990).

[5]In the context of the regression model to be analyzed here, the case of differing time intervals is handled by including the time variable in the linear index function of the model with a coefficient of 1.0. We will return briefly to this point below. For an application, see McCullagh and Nelder (1983, pp. 136-140).

[6]See Cameron and Trivedi (1986), El Sayyad (1973), Engel (1984), Holgate (1964), Jorgenson (1961), Lawless (1987a, 1987b) Maddala (1983), McCullagh Nelder (1983), and Simpson (1987).

insures a positive mean.[7] Estimation and inference for the Poisson model are considered below.

The Poisson distribution has the convenient, albeit restrictive property that

$$E[y_i] = Var[y_i] = \lambda_i,$$

(where, for the moment, we have subsumed the conditioning variables, $\boldsymbol{x}_i$ in the subscript). The equality of the mean and variance is the subject of the literature on over- and underdispersion in the Poisson model. Although a number of modifications have been proposed, the most frequently cited alternative is the negative binomial regression model,

$$p(y_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta)\, y_i!} u_i^{\theta} (1 - u_i)^{y_i}, \ \theta > 0, \ y_i = 0,1,\ldots$$

$$u_i = \frac{\theta}{\theta + \lambda_i}. \qquad (2.2)$$

where

This has $E[y_i] = \lambda_i$

and $Var[y_i] = \lambda_i[1 + (1/\theta)\lambda_i] = \lambda_i(1 + \alpha\lambda_i)$

The negative binomial model has been formulated with overdispersion,

$$\frac{Var[y_i]}{E[y_i]} = 1 + \alpha\, E[y_i] > 1,$$

as an end in itself,[8] or as a consequence of incorporating unobserved individual heterogeneity (e.g., Hausman, et al. (1984)). Let

$$E[y_i \mid \varepsilon_i] = \lambda_i \varepsilon_i$$

where ε_i is a disturbance distributed as gamma with mean 1 and variance α;

[7]Note that the normalization needed to accommodate an observation specific interval length is handled by $t\lambda_i = \exp(\ln t_i + \boldsymbol{\beta}'\boldsymbol{x}_i)$. Thus, $\ln t_i$ is simply included in the regression with a coefficient of one. Henceforth, we will omit further reference to the normalization and, for simplicity, just assume that t_i equals one.

[8]See Cameron and Trivedi (1986) (their model `NegBin II') and King (1989b).

$$g(\varepsilon_i) = \frac{\theta^{\theta}}{\Gamma(\theta)} e^{-\theta\varepsilon_i} \varepsilon_i^{\theta-1}, \varepsilon_i > 0, \theta = \frac{1}{\alpha}.$$

This produces

$$f(y_i|\varepsilon_i) = \frac{e^{-\exp(\beta' x_i + \varepsilon_i)} \left[\exp(\beta' x_i + \varepsilon_i)\right]^{y_i}}{y_i!}$$

The marginal distribution is

$$f(y_i) = \int_0^{\infty} f(y_i|\varepsilon_i) g(\varepsilon_i)\, d\varepsilon_i,$$

which is the negative binomial model given earlier.[9]

Maximum likelihood estimation of parameters of the Poisson regression model is straightforward. The log-likelihood function and its first and second derivatives are

$$\log L = \sum_{i=1}^{N} \log f(y_i) = \sum_{i=1}^{N} \left[-\lambda_i + y_i \log \lambda_i - \log y_i!\right], \quad (2.3)$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{N} (y_i - \lambda_i)\, x_i = \sum_{i=1}^{N} e_i\, x_i, \quad (2.4)$$

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} \lambda_i x_i x_i'$$

The Hessian is always negative definite, which makes Newton's method a convenient way to compute the *MLE* of β. Alternatively, the model is a nonlinear regression, so β can be estimated consistently by nonlinear ordinary least squares, or efficiently by nonlinear generalized (iteratively reweighted) least squares.[10]

The log-likelihood and gradient for the negative binomial model are[11]

[9]Hausman et al. (1984) present an extensive application of the model from this perspective. Recent modifications include Wedel, et al. (1993), Wasserman (1983), and Winkelmann and Zimmermann (1991a, 1991b, 1991c).

[10]Note that the moment condition in (2.4) is the same as that for the classical regression model and prescribes nonlinear GLS as the efficient GMM estimator.

[11]We have manipulated the function to eliminate the gamma integrals. This simplifies programming and marginally speeds up computation of the estimates. See Greene (1991). We will use the indicator function, **1**(condition) = 1 if the condition is true and 0 if not, at various points below.

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{N} u_i e_i x_i \qquad (2.6)$$

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{N} \left[\left(\frac{1}{\theta} \right) \log(1+\theta\lambda_i) + y_i \log u_i + \theta \log(1-u_i) - y_i \log\left(\frac{y_i}{\lambda_i} \right) \right] \qquad (2.7)$$

The function is a bit less well behaved than the Poisson log-likelihood owing to the need to keep θ positive, but is easily handled by a gradient method incorporating a line search. Among the interesting aspects of this model is the robustness of the Poisson *MLE* of β in the presence of heterogeneity (overdispersion).[12]

The literature on testing the Poisson restriction of equal mean and variance (over- or underdispersion) is vast.[13] The problem of testing for homogeneity does not fit into the classical Neyman-Pearson methodology because the restricted case lies on the boundary of the parameter space; $\alpha = 0$, or $\theta \rightarrow +\infty$.[14] The mechanics of the existing procedures for testing for heterogeneity or overdispersion are tangential to the subject of this paper. The interested reader is referred to the literature cited earlier for details. However, the specification test procedure for our zero altered model which is proposed below is readily adapted to a test for heterogeneity (see Section 3.1). An evaluation of this testing procedure versus the alternatives is left for further work.

Several modifications of the Poisson regression model beyond the negative binomial specification are of direct relevance to this study. They deal primarily with the observed frequencies in the data or the functional form of the conditional mean function.

Problems of censoring and truncation are common, and arise from the same sources as in the more familiar regression settings. In survey data, for example, respondents are sometimes given a limit category, `*C* or more' for some large value (censoring). In other settings, such as surveys of users of recreational facilities (e.g., Smith (1990), respondents who report zero are sometimes discarded from the sample. These complications can be built directly into the basic Poisson or negative binomial model in same way that they

[12]Gourieroux, et al. (1984) and White (1982).

[13]See, e.g., Breslow (1984, 1990), Cameron and Trivedi (1990), Chesher (1984), Collings and Margolin (1985), Cox (1983), Dean and Lawless (1989), Ganio and Schafer (1991), Gurmu(1991), King (1989), Lee (1986), Mullahy (1986,1990), Potthoff and Whittinghill (1966), and Wasserman (1983).

[14]Note that this is equivalent to the problem of testing for a zero variance, as arises in the random effects classical regression model. See Breusch and Pagan (1980).

are handled in classical normal regression model in the form of the tobit and truncated regression models.[15] For the first example given, the log-likelihood function for a model incorporating censoring of the form suggested is

$$\log\text{-}L = \sum_{i=1}^{N}\left[1\right].$$

The counterpart for a model with truncation at zero, as in the second example, would be

$$\log\text{-}L = \sum_{i=1}^{N}\left[\log p(y_i) - \log\left(1 - p(0)\right)\right]. \tag{2.9}$$

The gradients and Hessians, albeit tedious, are straightforward and appear in general form in Greene (1991).[16]

We will be interested in two rather sparsely analyzed variations on the Poisson/negative binomial models in this study. First, there are situations in which number of occurrences of a specific value (usually zero) exceeds what would be predicted by the Poisson model. The problem was analyzed by Cohen (1954) and is described in some detail in Johnson and Kotz (1969). Various modifications have been suggested which involve a rudimentary parameterization of the `non-Poissonness' of the distribution. For example, Heilbron (1989), who labels this the `zero altered Poisson,' or *ZAP*(λ,ρ) model, and Mullahy (1986), among others, who calls this a `hurdle' model,[17] suggest

$$\text{Prob}[y_i = 0] = \rho$$

$$\text{Prob}[y_i = k] = \left[\frac{1-\rho}{1-e^{-\lambda_i}}\right]\frac{e^{-\lambda_i}\lambda_i^{k}}{k!},\ k = 1, 2, \ldots$$

[15]See Greene (1991) for details. Applications are given by Terza (1985), Grogger and Carson (1988, 1991), Cohen (1954, 1960), Creel and Loomis (1990, 1991), van Praag (1993), and Shaw(1988).

[16]Full details for the censored Poisson model appear in Terza(1985). Greene(1991) gives results for models with truncation and for the negative binomial model.

[17]See Cragg (1971) and Lin and Schmidt (1984).

Heilbron's interpretation of the model is as a modification of the Poisson model to add mass to the zero point,[18] while Mullahy's hurdle interpretation (which is closer to ours) treats the modification as a binary data generating process. Thus, "[t]he idea underlying the hurdle formulations is that a binomial probability model governs the binary outcome of whether a count variate has a zero or a positive realization."[19,20] Note that the positive part of the distribution is the truncated Poisson model which appears in (2.9).

In the *ZAP* model, observations which surpass the `hurdle' are positive. Our interest here is in a setting in which the zeros are observed as well, with greater frequency than would otherwise be predicted by the Poisson model. The surfeit of zeros results from a mixture of two processes, both of which produce zeros. One generates the regime choice as a binary outcome, while the other generates the count variable, which may equal zero as well. In one regime, the zero value is automatic, while in the other, it is but one possible outcome. Consider, for example, answers to the survey question "[h]ow many children do you have?" Respondents would be of two types, some who have no intention of ever having children and some who may have some children or may not *yet* have any children at the time the question is asked, but might later. The model we propose is a straightforward modification of what Mullahy and Heilbron (following Johnson and Kotz) label the `with zeros' (*WZ*) model,

$$\text{Prob}[y_i = 0] = \psi + (1 - \psi) f(0)$$

$$\text{Prob}[y_i = j] = (1 - \psi) f(j), \; j = 1, 2, \ldots,$$

where ψ is a parameter between 0 and 1.[21] This formulation has the virtue of simplicity, though the inequality constraint on ψ does complicate the computation of the maximum likelihood estimates.

For our purposes, the primary shortcoming of Heilbron/Mullahy's specification is that there are no covariates in ψ, so that the construction of a behavioral splitting model (regime generation process) remains.

[18]Terza and Wilson (1990) adopt this formulation solely to induce overdispersion.

[19]Mullahy (1986, p. 345).

[20]The Hurdle model has a close resemblance to Schmidt and Witte's (1989) `splitting' model. They model a binary censoring indicator in the context of various survival models with a probit or logit specification. The counterpart to Mullahy's zero and truncated Poisson model is their survival or hazard function.

[21]As Heilbron notes, some negative values of ψ are admissable, though the interpretation of ψ as a mixing parameter will be lost in this case.

This interpretation of the model is suggested, more or less in the work of Lambert (1992), upon which much of our current study is built. Once again, her primary motivation is `non Poissonness,' although as will be clear below, one of her specifications provides our main building block.

The second framework proposed here will be on the subject of sample selection modelling. There appears to have been little progress on this aspect of the model.[22] We note that since the Poisson model is a bona fide regression model, the problem of sample selection poses itself naturally. In Smith's (1988) application, which drops neatly into this framework, we have a sample of observations which has been culled from a larger sample specifically on the basis of their use of recreational sites. The second purpose of this study is to offer one possible specification for handling the problem.

3 Modified Poisson and Negative Binomial Models

Our models for a zero augmented count model and for sample selection are built on the preceding in a straightforward fashion. The first is an extension of Lambert's *ZIP* model. The selection model takes the approach of modifying the joint discrete distribution of the random variables and the conditional mean function of the count variable, rather than relying on a transformation to normality to produce the conditional distribution of a latent continuous variable.

3.1. ZIP Models

Lambert (1992) proposes the following modification of the Heilbron/Mullahy *WZ* model, which she labels the `zero inflated Poisson' or *ZIP* model:[23]

$$\begin{aligned} y_i &\sim 0 && \text{with probability } q_i \\ y_i &\sim \text{Poisson}(\lambda_i) && \text{with probability } 1 - q_i, \end{aligned} \tag{3.1}$$

where $\log\lambda_i = \boldsymbol{\beta}'\boldsymbol{x}_i$ as before, and

$$q_i = \frac{e^{\gamma' w_i}}{1 + e^{\gamma' w_i}} \quad \text{or} \quad q_i = \frac{e^{\tau\beta' x_i}}{1 + e^{\tau\beta' x_i}} = \frac{\lambda_i^{\tau}}{1 + \lambda_i^{\tau}}$$

[22]Heilbron notes (p. 29) "For non-Poisson counts, there is no transformation of Y that would make sensible the application of a normal theory sample selection model. Further, it seems difficult to formulate an appealing sample selection model for count data." Smith (1990) makes note of the utility of a selection model for counts of uses of recreational sites, but states that it is "beyond the scope of his study." Bockstael, et al. (1990) note the issue in passing (p. 41) but treat the counts as realizations of a continuous choice variable, and make no further mention of the problem. Shaw's (1988) model is somewhat related to this problem, but his analysis centers on direct truncation, not sample selection as we are considering it here. This appears to be the extent of the received commentary on the subject.

[23]We have changed the notation a bit, but the substance of the model is identical.

Lambert labels the latter the *ZIP*(τ) model.[24] Thus, the *ZIP* model generalizes Mullahy's *WZ* model by parameterizing a formal probability model for ψ. Although Mullahy's interpretation of a primary regime generating process for ψ is consistent with the *ZIP* model, in fact, Lambert's primary interest appears to be in non-Poissonness, i.e., the shape of the distribution.[25] We will propose some extensions of Lambert's *ZIP* and *ZIP*(τ) models. First, we will consider an alternative formulation of the splitting variate - the determination of q_i.[26] Second, we will extend the *ZIP* model to the negative binomial model. The extension is a natural one. Lambert does mention the possibility of augmenting the mass at zero for other discrete distributions (p. 12), however, our interest here goes beyond merely specifying an alternative distribution. The presence of excess zeros in the data will likely lead to a conclusion of overdispersion. In our case, we are interested in heterogeneity as the source of the overdispersion. A zero inflated negative binomial (*ZINB* or *ZINB*(τ)) model will enable us to distinguish between the effect of the splitting mechanism and the overdispersion induced by individual heterogeneity. In this connection, we are interested in a procedure which will enable us to test the zero inflated model against the simple Poisson model or against the negative binomial model. The latter will allow us to make a statement as to whether the excess zeros are the consequence of the splitting mechanism or are a symptom of unobserved heterogeneity. We note that the same test that we propose here provides an as yet unexamined method of testing the specification of the negative binomial model against the Poisson model, independently of the splitting mechanism.

We consider a process whereby the observed random variable y_i is generated as

$$y_i = z_i y_i^*$$

where z_i is a binary (0/1) variable and y is distributed as Poisson(λ_i) or negative binomial (λ_i,θ). The *ZIP* model is, by this construction, a model of `partial observability' - only the product of the two latent variables,

[24]Lambert gives her formulation in terms of -τ, but since τ is unrestricted in sign or magnitude, no generality is lost by using our slightly more convenient parameterization.

[25]The introduction does state, however, "One interpretation is that slight, unobserved changes in the environment cause the process to move back and forth between a perfect state in which defects [the Poisson variate] are extremely rare and an imperfect state in which defects are possible but not inevitable." (Lambert, p. 1.) This is, of course, consistent with Mullahy's description, though, it understates the case a bit since in the perfect state predicted by the model, defects are impossible.

[26]Lambert does mention in passing some alternative formulations for q_i in the *ZIP*(τ) model (p. 3), but confines attention to the initial logit model.

z_i and y, is observed.[27] Thus,

$$\text{Prob}[y_i = 0] = \text{Prob}[z_i = 0] + \text{Prob}[z_i = 1, y_i^* = 0] = q_i + (1 - q_i)\, f(0)$$

$$\text{Prob}[y_i = k] = (1 - q_i)\, f(k), k = 1, 2, \ldots$$

where $f(\cdot)$ is the Poisson or negative binomial probability distribution for y. For an application, consider the response to the question how many trips have you taken to a certain sport fishing site? The answer to this question comes at two levels. There are individuals who would not visit a sport fishing site almost regardless of circumstances because this activity does not interest them, whereas there are others for whom the number of visits might follow more conventional patterns amenable to a Poisson or negative binomial regression - but might, once again, be zero. The binary part of the model, i.e., the splitting mechanism, lends itself conveniently to a probit or logit specification, though we need not limit it to those two choices. Likewise, the conditional count variable can have a Poisson or negative binomial (or, in principle, some other) distribution. Clearly, any combination of models for a binary outcome, z_i, and count variable, y, might be considered. We will limit our attention to the familiar probit and logit models for z_i and the Poisson and negative binomial models for y. We will also consider both the *ZIP* and *ZIP*(τ) specifications.

3.1.1 Estimation

Estimation of the parameters of the *ZIP* model is fairly straightforward. Lambert suggested the *EM* algorithm, but our experience has been that a straightforward gradient approach with a line search is more efficient and poses no unusual calibration problems. To formulate the log-likelihood and gradient for the *ZIP* models, let

$$q_i = F(\boldsymbol{\gamma}'\boldsymbol{w}_i) \text{ for the } ZIP \text{ model}$$

and

$$q_i = F(\tau\boldsymbol{\beta}'\boldsymbol{x}_i) \text{ for the } ZIP(\tau) \text{ model,}$$

where $F(\cdot)$ is either the cumulative normal probability, Φ, for the probit model or the cumulative logistic probability, $\Lambda(\cdot)$ for the logit model. Let $f(\cdot)$ denote either the Poisson(λ_i) or the negative binomial (λ_i,θ) probability density function. (This produces eight possible models.) Then, the probability density function for the observed random variable, y_i, is

[27] See Poirier (1980) and Abowd and Farber (1982).

$$p(y_i) = p_i = (1 - q_i)f(y_i) + 1(y_i = 0)q_i,$$

so the log-likelihood is simply

$$\log\text{-}L = \sum_{i=1}^{N} \log p(y_i). \tag{3.2}$$

To obtain the gradient, let $\boldsymbol{\beta}^*$ equal either $\boldsymbol{\beta}$ for the Poisson model or $(\boldsymbol{\beta},\theta)$ for the negative binomial model. Then, each term in Σ_i $(\partial \log p_i/\partial \boldsymbol{\beta}^*)$ is

$$\frac{\partial \log p_i}{\partial \beta^*} = \frac{1}{p_i}\left[(1-q_i)f(y_i)\left(\frac{\partial \log f(y_i)}{\partial \beta^*}\right) + \left(1(y_i=0) - f(y_i)\right)\left(\frac{\partial q_i}{\partial \beta^*}\right)\right]. \tag{3.3}$$

The derivatives of $\log f(y_i)$ were given in Section 2. Also, $\partial q_i/\partial \boldsymbol{\beta}^*$ will equal $\mathbf{0}$ in the *ZIP* model, or $\tau \boldsymbol{x}_i q_i'$ for the *ZIP*(τ) model with a trailing zero for θ if $f(y_i)$ is the negative binomial model, since θ does not enter q_i. (The inner derivative, q_i', is either the standard normal density, φ_i for the probit model, or $\Lambda_i(1-\Lambda_i)$ for the logit model.) Finally, the parameters of the *ZIP* model are either $\boldsymbol{\gamma}$, a vector, in the *ZIP* model or τ, a scalar, in the *ZIP*(τ) model. Denoting these generically as $\boldsymbol{\gamma}$, we have

$$\frac{\partial \log p_i}{\partial \gamma} = [1\ (y_i = 0) - f(y_i)]\ \frac{q_{i'}}{p_i}\ (\beta' x_i) \tag{3.4}$$

for the *ZIP*(τ) model. For the *ZIP* model, $\boldsymbol{\beta}'\boldsymbol{x}_i$ is replaced with $\boldsymbol{w}_i$, the vector of covariates. The second derivatives are fairly complicated. In our applications, we have used the BHHH estimator instead as a convenient expedient. Finally, in the *ZIP* model, the hypothesis that some or all of the parameters in $f(y_i)$ equal those in q_i might be of interest. Estimation subject to the equality constraints is straightforward, and carrying out the test via a Wald or likelihood ratio procedure can be done by conventional procedures.

For the *ZIP* specification, a natural set of starting values for the parameters is provided by the probit or logit and independent Poisson or negative binomial estimates. In the *ZIP*(τ) case, the Poisson or negative binomial model can be used for the regression parameters. One could then choose a value for τ which would produce approximately the correct probability for zero. An alternative possibility would be to estimate τ by fitting a probit or logit model to the binary indicator $\mathbf{1}(y_i = 0)$ with the single covariate equal to the Poisson estimates of $\boldsymbol{\beta}'\boldsymbol{x}_i$ (only so as to get the right sign and approximately the right magnitude on τ; this is not a consistent estimator). Save for a few badly identified cases found by experimentation in which

no solution could be found, convergence of the DFP or Broyden algorithms appears to be routine.

3.1.2 Specification Testing

The *ZIP* model relaxes the assumption of equal mean and variance in the Poisson model. To derive the unconditional mean and variance of y_i, we first consider the Poisson case. The two conditional distributions are

$$f(y_i | z_i = 0) = 1, y_i = 0,$$

$$f(y_i | z_i = 1) = \text{Poisson}(\lambda_i), y_i = 0,1,\ldots$$

Then

$$E[y_i] = E_{z_i}[E[y_i | z_i]] = q_i 0 + (1 - q_i)\lambda_i = (1 - q_i)\lambda_i$$

and

$$\text{Var}[y_i] = E_{z_i}[\text{Var}[y_i|z_i]] + \text{Var}_{z_i}[E[y_i|z_i]]$$

$$= [q_i\, 0 + (1 - q_i)\lambda_i] + [q_i\,(0 - (1 - q_i)\lambda_i)^2 + (1 - q_i)\,(\lambda_i - (1 - q_i)\lambda_i)^2]$$

$$= \lambda_i\,(1 - q_i)\,[1 + \lambda_i\, q_i].$$

The unconditional Poisson model emerges if $q_i \to 0$. Also,

$$\frac{\text{Var}[y_i]}{E[y_i]} = 1 + \lambda_i\, q_i = 1 + \left[\frac{q_i}{1 - q_i}\right] E[y_i]$$

so the splitting phenomenon produces overdispersion in its own right. Thus, $q_i/(1-q_i)$ is the counterpart to α in the negative binomial model as regards overdispersion. The ratio increases with q_i, as might be expected, so the more likely is the zero state, the greater is the overdispersion. For the negative binomial model, the conditional means are the same, so the unconditional mean is unchanged. Only the term $E_z[Var[y_i|z]]$ changes, from $(1-q_i)\lambda_i$ to $(1-q_i)\lambda_i(1 + \alpha\lambda_i)$. Combining terms and simplifying produces the unconditional variance for the negative binomial model,

$$\text{Var}[y_i] = (1 - q_i)\,\lambda_i[1 + (q_i + \alpha)\,\lambda_i] \quad \text{or} \quad \frac{\text{Var}[y_i]}{E[y_i]} = 1 + \left(\frac{q_i + \alpha}{1 - q_i}\right) E[y_i].$$

This shows that the overdispersion arises from these two independent sources. Moreover, the effects are cumulative, since the term in parentheses is greater than α for all positive q_i.

There is a large literature on testing for overdispersion in the Poisson model.[28] With rare exception, the diagnostic statistics proposed and analyzed are based on second moments constructed from:

- deviations of estimated means and variances (e.g., Dean and Lawless(1989)),
- deviations of a regression slope from one or zero (e.g., Cameron and Trivedi (1990)),
- deviations of derivatives from zero (LM tests) (e.g., Mullahy (1986).

While tests such as these are clearly related to the model analyzed here, the potential lack of fit of the Poisson or negative binomial model to the observed data seems also to have potential utility as a diagnostic.

[28]See, for example, Breslow (1990), Cameron and Trivedi (1990), Collings and Margolin (1985), Ganio and Schafer (1992), Gurmu (1991), Mullahy (1990), and Potthoff and Whittinghill (1966).

For this purpose, Mullahy's (1990) result seems particularly useful. The test is a moment based test which essentially compares the number of zeros in the data with the average predicted probability of a zero. His recommended test statistic is

$$q = \frac{\sqrt{N}\,\bar{d}}{\sqrt{\hat{v}_d}} \quad \text{where} \quad \bar{d} = \frac{1}{N}\sum_{i=1}^{N}[1(y_i = 0) - \hat{p}_i(0)]$$

$$= P_0 - \bar{\hat{p}}(0)$$

$$\text{and} \quad \hat{p}_i(0) = e^{-\hat{\lambda}_i}.$$

The variance term in the denominator is computed as

$$\hat{v}_d = \frac{1}{N}\sum_{i=1}^{N} d_i^2 + g'Wg + 2g'Wh$$

$$g = \frac{1}{N}\sum_{i=1}^{N} \hat{\lambda}_i\, \hat{p}_i(0)\, x_i$$

$$h = \frac{1}{N}\sum_{i=1}^{N} d_i\,(y_i - \hat{\lambda}_i)\, x_i$$

$$W = \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\lambda}_i x_i\, x_{i'}\right]^{-1}.$$

Mullahy's statistic should produce a positive value under any form of heterogeneity. Indeed, as he notes, regardless of the form of heterogeneity, the actual proportion of zero outcomes will tend to exceed the proportion expected under a Poisson assumption.

The added mass at zero produces overdispersion even in the absence of heterogeneity, so the full specification will allow one to distinguish between heterogeneity and the *ZIP* specification. But, the models are not nested. For the probit and logit models, setting the *ZIP* parameters to zero does not produce the

restricted model; it produces q_i = ½. The restricted model requires q_i to vanish, but this requires τ in the *ZIP*(τ) model or some element of γ in the *ZIP* model to explode. None of these is amenable to the familiar *LR* or Wald tests. Vuong (1989, p. 318) has proposed a test statistic for nonested models which is well suited to this application:

$$V = \frac{\sqrt{N}\,\bar{m}}{s_m}, \text{ where } m_i = \log\left[\frac{f_1(y_i)}{f_2(y_i)}\right]$$

and f_1 and f_2 are two competing probability models (e.g., *ZINB* and negative binomial). V is the standard statistic for testing the hypothesis that $E[m_i]$ is zero. Vuong shows that asymptotically, V has a standard normal distribution. As Vuong notes, the test is directional. If $|V|$ is less than the predetermined critical value (e.g., 1.96), then the test result does not favor one model or the other. Otherwise, large positive values favor model 1 while large negative values favor model 2. Carrying out the test requires estimation of both models and computation of the sample of predicted probabilities.[29] Thereafter, for example, the test statistic is a simple t-statistic for testing for a zero mean of the variable

$$m_i = \log\left[\frac{\text{Prob}[y_i \_ \text{ZINB}]}{\text{Prob}[y_i \_ \text{negative binomial}}\right].$$

This test may be more convenient than Mullahy's. The primary advantage is that Vuong's statistic makes use of information about the entire distribution, not just the zero outcomes. We do note that unlike the more familiar test statistics cited earlier, both of these are based on goodness of fit, rather than overdispersion.

Vuong's statistic could also be used to test the restriction of the Poisson distribution on the negative binomial. Its power characteristics are less than obvious, though there seems to be no reason a priori to expect them to be inferior to those of Mullahy's test or the other moment based tests listed earlier. An investigation of this issue is outside the scope of this study, and is left for further work.

3.2 A Model for Sample Selection

Previous discussions (e.g., Smith, Bockstael et al., Heilbron) have hinted at the utility of a sample selection framework for the count data models, but left the derivation of the joint and conditional distributions needed to formalize one for further work. We propose to proceed, instead, by focusing on the

[29]The models and test statistics described here are supplied as procedures in LIMDEP (Greene (1991)).

conditional mean function, in keeping with Heckman's (1979) treatment of sample selection as a specification error. We specify the joint distribution of the two observed discrete random variables as

$$g(z_i) = \Phi[(2z_i-1)\boldsymbol{\gamma}'\boldsymbol{w}_i],\ z_i = 0,1,\ \Phi(\cdot) = \text{standard normal } CDF,$$

$$f(y_i \mid z_i=1) = \text{Poisson}(\lambda_i) \text{ or negative binomial}(\lambda_i,\theta), \text{ observed only when } z_i = 1,$$

where $$\log \lambda_i = \boldsymbol{\beta}'\boldsymbol{x}_i + \rho M_i$$

and $$M_i = \varphi(\boldsymbol{\gamma}'\boldsymbol{w}_i)/\Phi(\boldsymbol{\gamma}'\boldsymbol{w}_i).$$

M_i is the Mill's ratio used in two step estimation of the sample selection model in continuous choice settings.[30] Rather than focus on the joint and conditional latent normal distributions, we direct attention to the joint discrete distribution of the observed random variables, z_i and y_i. The role of normality in $g(z_i)$ is only to provide a functional form for its conditional mean function. As such, any proper *CDF* would suffice. The specification of the probability model is the same as the probit model based on a latent threshold model and the normal distribution. The choice of the Mill's ratio for the additional term in $E[y_i \mid z_i=1,\boldsymbol{x}_i]$ is likewise arbitrary. It is made here to preserve the analogy with more familiar continuous choice models.

Estimation of the model parameters can be approached in two ways. The counterpart to Heckman's two step estimator would be obtained by estimating $\boldsymbol{\gamma}$ first as the coefficient vector in a probit model with z_i as the dependent variable and $\boldsymbol{w}_i$ as the vector of covariates. Then, the constructed regressor, M_i, is included in the Poisson model using the selected ($z_i = 1$) sample. By virtue of the consistency of the probit estimator, estimates of the parameters $(\boldsymbol{\beta},\rho)$ are consistent. However, as in the continuous choice case, the conventional estimated asymptotic covariance matrix at this step is inappropriate. This two step estimator fits directly into the framework developed by Murphy and Topel (1985). Their Section 5.1, equation (34),

[30] See Heckman (1979) and Greene (1993).

$$\Sigma = R_2^{-1}[R_2 + R_3'R_1^{-1}R_3 - R_4'R_1^{-1}R_3 - R_3'R_1^{-1}R_4]R_2^{-1} \quad (3.5)$$

applies here. Let $\boldsymbol{x}$ denote $[\boldsymbol{x}_i', M_i]'$ and $M = -\varphi_i/(1-\Phi_i)$. For the Poisson model

(3.6)

(Note that $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are the uncorrected estimators of the asymptotic covariance matrices for the probit and Poisson coefficients respectively.) The counterpart for the negative binomial model requires that the term $(y_i - \lambda_i)\boldsymbol{x}_i$ in $\boldsymbol{R}_3$ and $\boldsymbol{R}_4$ be replaced with $\partial \log p(y_i)/\partial(\boldsymbol{\beta}', \theta)'$, which appears in (2.2) and (2.6)-(2.8), and that $\boldsymbol{R}_2$ be replaced with a consistent (unconditional) estimator of the asymptotic covariance matrix of the MLE for $[\boldsymbol{\beta}, \theta]$. We have used the BHHH estimator for this purpose.

Since ρ is an unrestricted parameter, a test for "selectivity" in this context is a bit simpler than in the continuous choice case; a simple asymptotic t-test is equivalent, conditioned, of course on the other assumptions already made. Note, though, that unlike its counterpart in the continuous choice case, ρ is not the correlation or covariance of two underlying disturbances. The correlation between z_i and y_i and between z_i and $E[y_i \mid \boldsymbol{x}_i, z_i]$ are both complicated functions of $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$. However, the force of the selection "bias" is embodied in ρ insofar as the model carries the effect of selectivity through nonzero values of ρ. Thus, the parameterization of the effect is no more complicated here than in the continuous choice case.

The second approach would be full information maximum likelihood. The log-likelihood function and gradient are

$$\text{log-L} = \sum_{i=1}^{N} 1(z_i = 0)\log(1 - \Phi_i) + 1(z_i = 1)\left[\log\Phi_i + y_i \log\lambda_i - \lambda_i - \log y_i!\right]$$

and

$$\frac{\partial \text{log-L}}{\partial \alpha} = \sum_{i=1}^{N}\left[1\right] w_i$$

$$\frac{\partial \text{log-L}}{\partial \begin{bmatrix}\beta \\ \rho\end{bmatrix}} = \sum_{i=1}^{N} 1(z_i = 1)(y_i - \lambda_i)x_i^*$$

The BHHH estimator is a convenient estimator for the asymptotic covariance matrix of the estimates given the complexity of the first derivatives. Note that if ρ equals zero, the log-likelihood reduces to the simple sum of the probit and Poisson log-likelihoods. For the negative binomial model, the corresponding probability in the log-likelihood and the second part of the gradient are changed to the formulas given in Section 2. Our experience has been that the log-likelihood is occasionally somewhat ill behaved because of the Mill's ratio term, but generally presents no unusual difficulties. Good starting values for the procedure are easy to obtain since the two step procedure produces consistent estimates. FIML estimation is a matter of efficiency, not consistency.

3.3 Combining the *ZIP* and Selection Models

Accommodating sample selection in the *ZIP* and *ZIP*(τ) models is straightforward using results already given. The regression defined by λ_i in (3.1) need only be modified by inclusion of M_i as in the previous section. Estimation conditioned on the selection can then proceed in the same two steps: (1) Probit estimation of the selection equation and computation of M_i, and (2) *LIML* (now) estimation of the *ZIP* or *ZIP*(τ) model as discussed in Section 3.1. This setup is essentially the same as the one discussed in the previous section. What remains is to complete the specification of the terms in the asymptotic covariance matrix in (3.5) and (3.6). $\boldsymbol{R}_1$ remains as before, the estimated asymptotic covariance matrix of the parameter estimates in the probit selection equation. $\boldsymbol{R}_2$ is the estimated asymptotic covariance matrix of the parameters of the *ZIP* model. $\boldsymbol{R}_3$ is

$$R_3 = \sum_{i=1}^{N} 1(z_i = 1) M_i (\gamma' w_i + M_i) \left(\frac{\rho}{M_i}\right)\left(\frac{\partial \log p_i}{\partial \rho}\right) w_i v_{i'},$$

where $\boldsymbol{v}_i$ is the vector of partial derivatives defined in (3.3) and (3.4). Note that $\partial \log p_i / \partial \rho$ is one component of $\boldsymbol{v}_i$, and that the expression is on the right is constructed for convenience by using

$$\frac{\partial \log p_i}{\partial M_i} = \frac{\rho}{M_i} \frac{\partial \log p_i}{\partial \rho},$$

since the term in $\log p_i$ involving M_i is ρM_i. Finally,

$$R_4 = \sum_{i=1}^{N} 1(z_i = 1) M_i w_i v_{i'}.$$

In principle, one could estimate the full model by *FIML*, by using, in place of (3.5),

$$\log\text{-}L = \sum_{i=1}^{N} \left[1\right],$$

where $p(y_i)$ is defined in the first paragraph of Section 3.1.1.

4 Application to Credit Reporting Data

To illustrate the techniques described above, we have applied them to measurement on an aspect of consumer credit behavior. Among the variables kept as part of the credit history of individuals by reporting agencies such as TRW is the number of major derogatory reports (*MDR*s) within a fixed recent period. An *MDR* is defined as a delinquency of sixty days or more on a credit account. For the large majority of individuals, as suggested by the data described below, this is zero. But, for the remainder, values typically reaching as many as five, with some observations reaching fourteen or more, are observed. Thus, at the outset, the *ZIP* models appear to be appropriate for these data. Moreover, as we have argued elsewhere (Greene (1992)), this sort of credit behavior also lends itself well to splitting models, wherein delinquent behavior (or, in that earlier study, loan default) is conveniently modelled in a probit sort of setting.

Our sample of 1319 individuals is drawn from a population of applicants for a major credit card.[31] We have observed the variables for the *ZIP* models for all applicants. Of those 1319 applicants, 1023 were

[31]The credit card vendor who provided our data has requested anonymity.

given approval. Since the characteristics that would lead to a credit approval are (one would presume) precisely those which would typically mark an individual as unlikely to have any *MDR*s, this subsample also fits neatly into the familiar framework of sample selection models.[32] Table 1 describes the observed values of the count variable.

Table 1. Frequencies of MDRs.

			0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Full sample	1060	137	50	24	17	11	5	6	0	2	1	4	1	0	1		
(pct.)			.804	.104	.038	.018	.013	.008	.004	.005	.000	.002	.001	.003	.001	.000	.001
Selected sample			915	90	13	4	1										
(pct.)			.894	.088	.013	.004	.001										

As expected, the clustering of individuals at zero is more pronounced among the cardholders.

[32]Credit card vendors outsource the evaluation of card applications to third parties who have automated the process of credit scoring. Their exact procedures are closely guarded secrets. However, according to the vendor who provided our data, it is known that these evaluators employ relatively simple linear discriminant procedures. Precisely which variables they use is unknown, but we have found that the number of MDRs is a highly significant covariate in any applicant acceptance equation. This suggests that a simultaneous equations approach might be appropriate for our model, but that extension is beyond the current study and is left for further work.

Table 2 lists descriptive statistics for the covariates in the Poisson regression and selection (cardholder acceptance) equation.[33]

Table 2. Descriptive statistics for independent variables.

Variable	Mean	Standard deviation	Minimum	Maximum	
Income	Income, in 10,000s				
		3.365	1.694	0.21	13.50
Age	Age, in years				
		33.21	10.14	17.00	83.50
Cur._Add.	Number of months residing at current address				
		55.268	66.272	0.000	540.000
Exp._Inc.	Average monthly expenditure divided by yearly income				
		0.0687	0.0946	0.0001	0.9063
Avg._Exp.	Average monthly credit card expenditure				
		185.06	272.22	0.00	3100.00
Own_Rent	Binary variable indicating home ownership				
		0.440	0.497	0.00	1.00
Self_Empl.	Binary variable, one for self employed				
		0.070	0.253	0.00	1.00
Depndt.	Number of dependents, not including the individual				
		0.994	1.248	0.00	6.00
Inc._per	Income per dependent				
		2.156	1.363	0.070	11.00
Major	Binary variable for whether the individual holds a major credit card				
		0.817	0.387	0.000	1.000
Active	Number of active credit card accounts				
		6.997	6.306	0.000	46.000

Table 3 contains estimates of the various *ZIP* and *ZIP*(τ) models specified in Section 3.1. Ignoring, for the moment, the *ZIP* specification, the Poisson model in the first column appears to be clearly inferior to the negative binomial model in the second. There are 1060 zeros in the observed data, which the negative binomial predicts far more closely. (The predicted number of zeros is computed as the sample sum of the estimates of Prob[$y_i = 0$] for all of the models.) By the Wald and likelihood ratio tests, it is clear that under

[33]The latter is greatly simplified from that used in Greene (1992).

any circumstances, the Poisson model would be rejected in favor of the negative binomial. Indeed, based only on the results considered thus far, it is not obvious that any of the ZIP specifications is likely to provide an improvement over the negative binomial specification.

Table 3. Parameter estimates for ZIP models.
Estimated standard errors in are parentheses. 1319 observations.

			Poisson				Negative Binomial			
			ZIP(τ)		ZIP		ZIP(τ)		ZIP	
	Poisson (1)	Neg.Bin (2)	Logit (3)	Probit (4)	Logit (5)	Probit (6)	Logit (7)	Probit (8)	Logit (9)	Probit (10)
Regression Model										
Constant -0.370*	-0.878*	0.897*	0.980*	1.111*	1.115*	-0.100	-0.100	0.550	0.489	
	(.174)	(.384)	(.140)	(.140)	(.145)	(.145)	(.351)	(.351)	(.551)	(.565)
Income	-0.025	-0.006	-0.022	-0.023	-0.032	-0.048	-0.106	-0.108	-0.047	-0.049
	(.028)	(.057)	(.023)	(.023)	(.023)	(.023)	(.051)	(.051)	(.058)	(.059)
Major	0.046	0.055	0.193*	0.191*	0.124	0.124	0.065	0.066	0.042	0.042
	(.105)	(.207)	(.080)	(.081)	(.076)	(.079)	(.185)	(.185)	(.201)	(.202)
Age	0.005	0.011	-0.005	-0.005	-0.006*	-0.006*	0.009	0.009	-0.015	-0.014
	(.004)	(.009)	(.003)	(.003)	(.001)	(.003)	(.009)	(,009)	(.009)	(.010)
Exp._Inc.	-18.0*	-9.29*	-14.0*	-14.2*	-9.20*	-9.18*	-8.90*	-8.92*	-8.63*	-8.66*
	(2.20)	(1.73)	(1.36)	(1.36)	(1.44)	(1.44)	(2.69)	(2.70)	(1.67)	(1.76)
Avg._Exp.	0.0014*	0.0006	0.0007	0.0007	0.0003	0.0003	0.0005	0.0005	0.0005	0.0005
	(.0006)	(.0007)	(.0005)	(.0005)	(.0005)	(.0005)	(.0006)	(.0006)	(.0006)	(.0007)
α		4.813*					1.718*	1.716*	2.803*	2.902*
		(.516)					(.229)	(.230)	(.781)	(.834)
Splitting Model										
Constant				1.957*	1.199*			3.148*	1.830	
					(.326)	(.193)			(1.017)	(.616)
Age					-0.024*	-0.015*			-0.112*	-0.065*
					(.009)	(.005)			(.040)	(.023)
Income					-0.073	-0.043			-0.194	-0.132
					(.087)	(.053)			(.324)	(.207)
Own_Rent				0.395	0.235*			0.898*	0.553	
					(.179)	(.105)			(.450)	(.286)
Self_Empl.					-0.081	-0.045			0.163	.098
					(.288)	(.174)			(.769)	.489)
Depndt					-0.203	-0.013			-0.054	-0.028
					(.125)	(.075)			(.342)	(.216)

Inc._per				-0.005	-0.004			-0.036	-0.023	
					(.121)	(.073)			(.396)	(.253)
τ			0.856*	0.520*			-0.088	-0.052		
			(.108)	(.066)			(.392)	(.244)		
Log-L	-1367.5	-1028.3	-1134.5	-1134.7	-1083.6	-1083.5	-1029.8	-1029.8	-1020.6	-1020.7
V			23.30	23.21	24.98	24.99	3.899	3.899	2.822	2.573
$\hat{N}_0$	980	1070	898	899	1064	1063	1074	1073	1047	1049

*Larger than two times the estimated asymptotic standard error.

Moving through the table, it is obvious that in all cases, the additional generality of the negative binomial model over the Poisson is called for, even when the clustering at zero is accounted for by the ZIP model. Compare, for examples, the diagnostic statistics in columns 3 and 7, 4 and 8, 5 and 9, and 6 and 10. Likewise, the *ZIP*(τ) model, which parameterizes the splitting phenomenon with a single parameter, appears to be inferior to the more general specifications. (The comparison is between columns 3 and 5 and between 4 and 6 for the Poisson model and between 7 and 9 then 8 and 10 for the negative binomial models.) The *ZIP* model appears to be preferable; one would expect that given its greater number of parameters. The models are not nested unless the variables in the splitting equation are the same as those in the regression, in which case, the restriction is that the parameters in the splitting equation all be the same proportion (τ) of their counterparts in the regression.[34]

For present purposes, the important question is whether the *ZIP* models in columns (7)-(10) provide any improvement over the basic negative binomial in column (2). The log-likelihood functions are uniformly higher, but as noted earlier, since the models are not nested, these are not directly comparable. The Vuong statistic, however, is consistent with the observation. In all cases, the statistic is well over two, which does imply that even with the striking improvement from column (1) (Poisson) to (2) (negative binomial), the *ZIP* specification sharpens the model even further. It would appear from the diagnostic statistics that although the negative binomial does capture the cluster at zero well, it does so at the expense of the fit to the rest of the distribution. The ZIP models in columns (9) and (10) predict the number of zeros as well as the unaugmented model (the former slightly underpredicts while the latter slightly overpredicts) but apparently capture the shape of the distribution of observed values better at the same time. Once again, the broader parameterization probably explains this.

Table 4 presents estimates of the sample selection models. The probit model is given in the first column. The coefficients are in line with expectations, with the only exception being the sign and magnitude of the coefficient on *Active* (the number of active credit card accounts). In fact, this variable is acting the way it should. Major credit card vendors are reluctant to open additional accounts for individuals with many cards, since the expected profit from the account diminishes when it must compete

[34]This is the model estimated by Lambert (1992), though she did suggest the more general specification used here.

Table 4. Parameter estimates for sample selection models. Estimated standard errors in are parentheses.

		Poisson			Negative Binomial		
		Base	ZIP(τ)	ZIP	Base	ZIP(τ)	ZIP
		Regression Model					
Constant	-4.594	-2.878	-3.090	-4.594	-2.904	-4.190	
		(.467)	(.903)	(.928)	(.552)	(.121)	(.783)
Income		-0.183	0.135	0.135	0.183	0.136	0.174
		(.056)	(.072)	(.084)	(.075)	(.087)	(.085)
Major		0.572	0.379	0.493	0.572	0.382	0.495
		(.278)	(.255)	(.315)	(.312)	(.277)	(.330)
Age		0.016	0.012	0.002	0.016	0.012	0.012
		(.009)	(.010)	(.019)	(.011)	(.010)	(.015)
Exp._Inc.		1.878	1.269	1.719	1.878	1.281	1.752
		(1.29)	(1.99)	(2.28)	(2.40)	(2.02)	(2.56)
Avg._Exp.		-.00002	.00003	-.00003	-.000007	.00003	.00002
		(.0004)	(.0006)	(.0006)	(.0007)	(.0006)	(.0008)
M_i		1.788	1.303	1.883	1.788	1.313	2.017
		(.296)	(.479)	(.466)	(.436)	(.538)	(.489)
α					1.211	0.001	0.875
					(.484)	(.285)	(.834)
	Probit	Splitting Model					
Constant	0.542		1.505			1.189	
	(.184)			(.903)			(1.52)
Age	-0.0086			-0.021			-0.016
	(0..5)			(.026)			(.043)
Income	0.092			-0.036			0.247
	(.053)			(.186)			(.420)
Own_Rent	0.349	Probit Predictions		-0.298			-1.054
	(.101)			(.316)			(.821)
Self_Empl.	-0.021	Predicted		-0.195			
	(.163)	Actual 0	1 Total	(.574)			(1.37)
		0 67	229 296				
Depndt	-0.131	1 17	1006 1023	0.053			-0.087
	(.069)			(.216)			(.500)
		Total 84	1235 1319				
Inc._per	-0.015		-0.271			-0.945	
	(.071)			(.298)			(1.16)
Major	0.212						
	(.103)						
Cur._Add.	-0.00041						
	(.0007)						
Active	-0.230						

	(.021)						
τ			-0.331			-0.320	
			(.277)			(.444)	
Log-L	-577.1	-394.2	-387.7	-385.1	-386.9	-387.7	-383.0
V			7.94	5.61		1.21	1.51
$\hat{N}_0$		918	927	917	923	917	912

with many other vendors.

The regression and *ZIP* models for the selected data are, as expected, quite different from the full sample. As before, the base Poisson model (column (2)) would be rejected in favor of the base negative binomial model (column (5)), based on the likelihood ratio test. But, unlike the previous case, the negative binomial does not predict the zero outcome noticeably better than the Poisson model. Moreover, unlike our previous case, the *ZIP* models appear to be acting as surrogates for the negative binomial model (or vice versa). In columns (3) and (4), we see that the *V* statistic clearly favors the *ZIP* models over the simple Poisson model. However, as noted above, the data also support the negative binomial model over the Poisson. In the context of the negative binomial models (columns (6) and (7)), the *ZIP* specification appears to offer no additional fit to the data. The *V* statistics for the last two models are only slightly larger than 1.0, and do not favor the *ZIP* models over the base case. Note that the overdispersion parameter in the *ZIP* models is no longer statistically significant, and in the *ZIP*(τ) case, its magnitude is trivial. The log-likelihood values suggest the same conclusion, namely that in the selected sample, it is difficult to distinguish between the two types of heterogeneity built into the model.

The data do support the sample selection model in all cases. That is, the estimate of ρ is substantially larger than twice its standard error in all specifications. But, the estimate of ρ seems to have the wrong sign. Consider that

$$\begin{aligned}
\mathrm{Cov}[y_i, z_i] &= \mathrm{Cov}[z_i, E[y_i | z_i]] \\
&= \mathrm{Cov}[z_i, e^{\beta' x_i + \rho z_i M_i}] \\
&= E\left[(z_i - \Phi_i)\left(e^{\beta' x_i + \rho z_i M_i}\right)\right] \\
&= \Phi_i(1 - \Phi_i)e^{\beta' x_i}\left[e^{\rho M_i} - 1\right],
\end{aligned}$$

which, since M_i is positive, has the same sign as ρ. However, upon closer inspection, this is not the appropriate way to view the force of the model. Since ρ is the coefficient on M_i in the model, it carries the effect of increases in M_i on $E[y_i | \boldsymbol{x}_i, \boldsymbol{w}_i]$. Increases in M_i are associated with increases in the expected number of delinquencies. But, $dM_i/d\Phi_i$ is negative, so increases in the expected number of delinquencies are associated with decreases in the probability of obtaining the credit card, which is what one would expect.

This outcome persists in all specifications of the model that we examined. It is interesting that essentially the same result occurred in Boyes, et al. (1990) in their model of consumer bank loan default (a zero one outcome) and in Greene (1992) in his model of credit card default using this same data set. In both of these cases, a sample selection model based on a cardholder equation similar to the one above for the sampling procedure is applied to a model of loan default, which is clearly related to what we are examining here.

5 Conclusions

We have presented several modifications of the Poisson regression model. Most of these depart from received specifications, though our *ZIP* model is a bit more general than those that appear in some other applications.

The use of Vuong's statistic to test the specification seems not to have appeared in the received literature. It remains for further work to see if the power of this test is comparable to that of the more familiar procedures, which are targeted more directly to the variance of the distribution being modelled. The importance of this to the current study is that the *ZIP* specification produces overdispersion, but only as a consequence of its transformation of the Poisson model into some other distribution. Thus, a test which is targeted specifically toward the variance of the distribution would seem to be misdirected. In this sense, an analogy to the Bowman and Shenton(1972) (skewness and kurtosis) test for normality seems appropriate. We conjecture that the Vuong testing procedure offers some real potential for testing the distributional assumption in this discrete data context. In the cases examined, it appeared to perform well and in line with expectations. Whether it shows similar promise in continuous data settings remains to be seen.

Our sample selection model is constructed somewhat differently from the conventional continuous choice settings. To maintain the strict analogy, one might have departed from the following specification:

$$z_i^* = \alpha' w_i + u_i,\ u_i \sim N(0,1),$$

$$z_i = 1(z_i^* > 0)$$

$$y_i \_ \varepsilon_i \sim \text{Poisson}(\lambda_i \_ \varepsilon_i)$$

$$\log(\lambda_i \_ \varepsilon_i) = \beta' x_i + \varepsilon_i,\ (\varepsilon_i, u_i) \sim N_2(0,0,\sigma,1,\rho).$$

In the selected sample, it follows, then, that

$$E[\log(E[y_i \_ x_i]) \_ z_i = 1] = \beta' x_i + \rho\sigma M_i.$$

This may be a bit closer to the orthodoxy than our specification, but this remains to be seen. In the continuous choice case, the interesting results surround $E[y_i \mid \mathbf{x}_i, z_i=1]$, not $E[\log(E[y_i \mid \mathbf{x}_i]) \mid z_i=1]$, which is something very different. Moreover, the latter specification precludes both the Poisson and negative binomial distributions for the marginal distribution of y_i. Details on the nature of the marginal distribution, the exact form of $E[y_i \mid \mathbf{x}_i, z_i=1]$, and an estimation strategy for this model remain to be worked out. The two directions suggested here, our selection model or the orthodox approach detailed above, should provide appropriate directions for continued work on a wholly satisfactory approach to the sample selection problem for count data models.

References

Abowd, J., and H. Farber, `Job Queues and Union Status of Workers,' *Industrial and Labor Relations Review*, 35, 1982, pp. 354-367.

Agresti, A., *Analysis of Ordinal Categorical Data*, John Wiley and Sons, New York, 1984.

Arvan, L., `Optimal Labor Contracts with On-the-Job Search: Are Involuntary Layoffs Used as an Incentive Device to Make Workers Search Harder"' Journal of Labor Economics, 7, 1989, pp. 147-154.

Berndt, E., B. Hall, R. Hall, and J. Hausman, `Estimation and Inference in Nonlinear Structural Models,' *Annals of Economic and Social Measurement*, 3, 1974, pp. 653-666.

Bockstael, N., I. Strand, K. McConnell, and F. Arsanjani, `Sample Selection Bias in the Estimation of Recreation Demand Functions: An Application to Sportfishing,' *Land Economics*, 66, 1990, pp. 40-49.

Bowman, K., and L. Shenton, `Tests for Departures from Normality based on $\sqrt{\beta_1}$ and β_2,' *Biometrika*, 62, 1975, pp. 243-251.

Boyes, W., D. Hoffman, and Low, S., `An Econometric Analysis of the Bank Credit Scoring Problem,' *Journal of Econometrics*, 40, 1989, pp. 3-14.

Breslow, `Extra-Poisson Variation in Log-Linear Models,' *Applied Statistics*, 33, 1984, pp. 38-4.

Breslow, N., `Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models,' *Journal of the American Statistical Association*, 85, 1990, pp. 565-571.

Breusch, T., and A. Pagan, `The LM Test and Its Applications to Model Specification in Econometrics,' *Review of Economic Studies*, 47, 1980, pp. 239-254.

Cameron, C., and P. Trivedi, `Econometric Models Based On Count Data: Comparisons of Some Estimators and Tests,' *Journal of Applied Econometrics*, 1, 1986, pp. 29-54.

Cameron, C., and P. Trivedi, `Regression Based Tests for Overdispersion in the Poisson Model,' *Journal of Econometrics*, 46, 1990, pp. 347-364.

Chesher, A., `Testing for Neglected Heterogeneity,' *Econometrica*, 52, 1984, pp. 865-872.

Collings, B., and B. Margolin, `Testing Goodness of Fit for the Poisson Assumption When Observations Are Not Identically Distributed,' *Journal of the American Statistical Association*, 80, 1985, pp. 411-418.

Cohen, A., `Estimation of the Poisson Parameter from Truncated Samples and from Censored Samples,' *Journal of the American Statistical Association*, 49, 1954, pp. 158-168.

Cohen, A., `Estimation in the Truncated Poisson Distribution When the Zeros and Some Ones are Missing,' *Journal of the American Statistical Association*, 55, 1960, pp. 342-348.

Cooil, B., `Using Medical Malpractice Data to Predict the Frequency of Claims: A Study of Poisson Process Models,' *Journal of the American Statistical Association*, 86, 1991, pp. 285-295.

Coughlin, C., J. Terza, and N. Khalifah, `The Determinants of Escape Clause Petitions,' Working paper, Research Department, Federal Reserve Bank of St. Louis, 1988.

Cox, D., `Some Remarks on Overdispersion,' *Biometrika*, 70, 1983, pp. 269-274.

Cragg, J., `Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods, *Econometrica*, 35, 1967, pp. 89-110.

Creel, M., and J. Loomis, `Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California,' *American Journal of Agricultural Economics*, 72, 1990, pp. 434-441.

Creel, M., and J. Loomis, `Confidence Intervals for Welfare Measurements with Application to a Problem of Truncated Counts,' *Review of Economics and Statistics*, 73, 1991, pp. 370-373.

Dean, C., and F. Lawless, `Testing for Overdispersion in Poisson Regression Models,' *Journal of the American Statistical Association*, 84, 1989, pp. 467-472.

Efron, B., `Poisson Overdispersion Estimates Based on the Method of Asymmetric Maximum Likelihood,' *Journal of the American Statistical Association*, 87, 1992, pp. 98-106.

El-Sayyad, G., `Bayesian and Classical Analysis of Poisson Regression,' *Journal of the Royal Statistical Society, Series B*, 35, 1973, pp. 445-451.

Engel, J., `Models for Response Data Showing Extra-Poisson Variation,' *Statistica Neerlandica*, 38, 1984, pp. 159-167.

Fin, T., and P. Schmidt, `A Test of the Tobit Specification Against an Alternative Suggested by Cragg,' *Review of Economics and Statistics*, 66, 1984, pp. 174-177.

Flowerdew, R., and M. Aitken, `A Method of Fitting the Gravity Model Based on the Poisson Distribution,' *Journal of Regional Science*, 22, 1982, pp. 191-202.

Frome, E., L. Kutner, and J. Beauchamp, `Regression analysis of Poisson-distributed Data,' *Journal of the American Statistical Association*, 68, 1973, pp. 935-940.

Frome, E., `The Analysis of Rates Using Poisson Regression Models,' *Biometrics*, 39, 1983, pp. 665-674.

Ganio, L., and D. Schafer, `Diagnostics for Overdispersion,' *Journal of the American Statistical Association*, 87, 1992, pp. 795-804.

Gart, J., `The Analysis of Poisson Regression With an Application in Virology,' *Biometrika*, 51, pp. 517-521.

Gourieroux, C., A. Monfort, and A. Trognon, `Pseudo Maximum Likelihood Methods: Applications to Poisson Models,' *Econometrica*, 52, 1984, pp. 701-720.

Gray, W., and C. Jones, `Are OSHA Health Inspections Effective? A Longitudinal Study in the Manufacturing Sector,' *Review of Economics and Statistics*, 73, 1991, pp. 504-508.

Greene, W., *LIMDEP Version 6.0, User's Manual*, Econometric Software, Inc., Bellport, New York, 1991

Greene, W., `Statistical Models for Credit Scoring,' Working Paper, Department of Economics, Stern School of Business, New York University, 1992.

Greene, W. *Econometric Analysis*, Macmillan, New York, 1993

Grogger, J., `The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts,' *Journal of the American Statistical Association*, 85, 1990a, pp. 295-303.

Grogger, J., `A Simple Test for Exogeneity in Probit, Logit, and Poisson Regression Models, *Economics Letters*, 33, 1990b, pp. 329-332.

Grogger, J., and R. Carson, `Models for Truncated Counts,' *Journal of Applied Econometrics*, 6, 1991, pp. 225-238.

Grogger, J., and R. Carson, `Models for Counts from Choice Based Samples,' Working paper, Department of Economics, University of California, San Diego, 1988.

Gurmu, S., `Tests for Detecting Overdispersion in the Positive Poisson Regression Model,' *Journal of Business and Economic Statistics*, 9, 1991, pp. 215-222.

Hausman, J., B. Hall, and Z. Griliches, `Economic Models for Count Data with an Application to the Patents-R&D Relationship,' *Econometrica*, 52, 1984, pp. 909-938.

Heckman, J., `Sample Selection Bias As a Specification Error,' *Econometrica*, 47, 1979, pp. 153-161.

Heilbron, D., `Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data,' Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco, 1989.

Hinde, J., `Compound Poisson Models,' in R. Gilchrist, (ed.), *GLIM82: Proceedings of the International Conference on Generalized Models*, Springer Varlag, New York, 1982.

Hoffman, S., and J. Milligan, `A Queueing Theory Approach to Daycare Standards,' Working paper, Department of Economics, University of Delaware, 1990.

Holgate, P., `Estimation for the Bivariate Poisson Distribution,' *Biometrika*, 51, 1964, pp. 241-245.

Johnson, N., and S. Kotz, *Distributions in Statistics - Discrete Distributions*, John Wiley and Sons, New York, 1969,

Jorgenson, D., `Multiple Regression Analysis of a Poisson Process,' *Journal of the American Statistical Association*, 56, 1961, pp. 235-245.

King, G., `Statistical Models for Political Science Event Counts,' Working paper, Department of Politics, new York University, 1985.

King, G., 'A Method of Estimating A Seemingly Unrelated Poisson Regression Model and Testing Cross Equation Hypotheses,' Working paper, Department of Politics, New York University, 1986.

King, G., `A Seemingly Unrelated Poisson Regression Model,' Sociological Methods and Research, 17, 1989a, pp. 235-255.

King, G., `Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator,' *American Journal of Political Science*, 33, 1989b, pp. 762-784.

Kostiuk, P., and D. Follman, Learning Curves, Personal Characteristics, and Job Performance,' *Journal of Labor Economics*, 7, 1989, pp. 129-146.

Lambert, D., `Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing,' *Technometrics*, 34, 1, 1992, pp. 1-14.

Lawless, J., `Negative Binomial and Mixed Poisson Regression,' *Canadian Journal of Statistics*, 15, 1987a, pp. 209-225.

Lawless, J., `Regression Methods for Poisson Process Data,' *Journal of the American Statistical Association*, 82, 1987b, pp. 808-815.

Lee, L., `Specification Tests for Poisson Regression Models,' *International Economic Review*, 27, 1986, pp. 689-706.

Maddala, G., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, New York, 1983.

McCullagh, P., and J. Nelder, *Generalized Linear Models*, Chapman and Hall, New York, 1983.

Murphy, K., and R. Topel, `Estimation and Inference in Two-Step Econometric Models,' *Journal of Business and Economic Statistics*, 3, 1985, pp. 370-379.

Mullahy, J., `Specification and Testing of Some Modified Count Data Models,' *Journal of Econometrics*, 33, 1986, pp. 341-365.

Mullahy, J., `Detecting Heterogeneity and Overdispersion in Poisson Models via Goodness of Fit,' Working paper, Department of Economics, Trinity College, 1990.

Okoruwa, A., J. Terza, and H. Nourse, `Estimating Patronization Shares for Urban Retail Centers: An Extension of the Poisson Gravity Model,' *Journal of Urban Economics*, 24, 1988, pp. 241-259.

Papke, L., `The Influence of Interstate Tax Differentials on the Birth of New Firms: Estimates of a Poisson Process,' Working paper, Department of Economics, MIT, 1986.

van Praag, B. and E. Vermulen, `A Count-amount Model with Endogenous Recording of Observations,' *Journal of Applied Econometrics*, 8, 1993, pp. 383-396.

Poirier, D., `Partial Observability in Bivariate Probit Models,' *Journal of Econometrics*, 12, 1980, pp. 209-217.

Portney, P., and J. Mullahy, `Urban Air Quality and Acute Respiratory Illness,' *Journal of Urban Economics*, 20, 1986, pp. 21-38.

Potthoff, R., and M. Whittinghill, Testing for Homogeneity II: The Poisson Distribution,' *Biometrika*, 53, 1966, pp. 183-190.

Shaw, D., `On-Site Samples: Regression Problems of Nonnegative Integers, Truncation, and Endogenous Stratification,' *Journal of Econometrics*, 37, 1988, pp. 211-223.

Schmidt, P., and A. Witte, ``Predicting Criminal Recidivism Using Split Population Survival Time Models,' *Journal of Econometrics*, 40, 1989, pp. 141-160.

Simpson, D., `Minimum Hellinger Distance Estimation for the Analysis of Count Data,' Journal of the American Statistical Association, 82, 1987, pp. 802-807.

Smith, V, `Selection and Recreation Demand,' *American Journal of Agricultural Economics*, 70, 1988, pp. 29-36.

Terza, J., `A Tobit Type Estimator for the Censored Poisson Regression Model,' *Economics Letters*, 18, 1985, pp. 361-365.

Terza, J., 'A Mixed Poisson-Multinomial Regression Model for the Analysis of Count Data,' Working paper, Department of Economics, University of Georgia, 1986.

Terza, J., and P. Wilson, Analyzing Frequencies of Several Types of Events: A Mixed Multinomial-Poisson Approach,' *Review of Economics and Statistics*, 72, 1990, pp. 108-115.

Vuong, Q., `Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses,' *Econometrica*, 57, 1989, pp. 307-334.

Wasserman, S., `Distinguishing Between Stochastic Models of Heterogeneity and Contagion,' *Journal of Mathematical Psychology*, 27, 1983, pp. 201-215.

Wedel, M., W. DeSarbo, J. Bult, and V. Ramaswamy, `A Latent Class Poisson Regression Model for heterogeneous Count Data,' *Journal of Applied Econometrics*,' 8, 1993, pp. 397-412.

White, H., `Maximum Likelihood Estimation of Misspecified Models,' *Econometrica*, 50, 1982, pp. 1-16.

Winkelmann, R., and K. Zimmermann, `A New Approach for Modeling Economic Count Data,' *Economics Letters*, 37, 1991a, pp. 139-143.

Winkelmann, R., and K. Zimmermann, `Count Data Models for Demographic Data,' Working paper, SELAPO, University of Munich, 1991b.

Winkelmann, R., and K. Zimmermann, `Inference in Misspecified Poisson Models,' Working paper, SELAPO, University of Munich, 1991c.