A Simple Two-Step Estimator for Count Data Models with Sample Selection

Chris D. Orme and Simon A. Peters
School of Economic Studies,
University of Manchester,
Manchester M13 9PL.

October 31, 2001

Abstract

Extending the earlier work of Greene (1997) and Terza (1998), it is shown how the mean of a count variable, suffering sample selection or endogenous treatment effects, can be arbitrarily well approximated using polynomial-type Heckman correction variables. This leads to a particularly simple estimator based on standard pseudo maximum likelihood methods. A small simulation study is presented to demonstrate the utility of the procedure.

1 Introduction

In Winkelmann (1998), various full information maximum likelihood (FIML) procedures for count data models are considered when the dependent variable suffers from selectivity or endogenous treatment effects. The basic idea is to introduce a heterogeneity term into the regression specification of the count model which is correlated, through a parameter ρ, with the error term in a probit model for the selection/dummy endogenous variable. Under the distributional assumptions made, the derived estimator is consistent and asymptotically efficient. However, some authors have noted that such FIML procedures can be operationally problematic; see, for example, Cameron and Trivedi (1997). A formal treatment of the sample selection problem was also given by Greene (1994) who suggested a simple Heckman-type two-step estimator. Both Cameron and Trivedi (1997) and Terza (1998) argued that this Heckman-type procedure is not based on a correct adjustment to the mean count and they, therefore, concluded that this estimator will be inconsistent. By deriving the correct mean function, Greene (1995) derived a two-step non-linear least squares (NLS) estimator, as did Terza (1998) who termed it a two-step method of moments (TSM) estimator. Terza (1998) also developed a non-linear weighted least squares (NWLS) estimator, which is asymptotically more efficient than the TSM; he also extended the

analysis to endogenous treatment effects. Interestingly, Greene (1995,1997,1998) points out that the exact mean function, for the observed count, has the same linear expansion, in ρ, as that implied by the use of the Heckman correction (in Greene, 1994), so that the latter might be assumed to be approximately correct. Although Winkelmann, Greene and Terza all provide illustrative applications, they offer no Monte Carlo evidence in support of their procedures.

In this paper new results are derived which show that the mean of the dependent count can be *arbitrarily* well approximated, under selectivity or endogenous treatment effects, by adding $r \geq 1$ Heckman-type correction variables to the regression specification. These correction variables are simply the truncated cumulants of a standard normal variate and include, in the case of sample selection, the familiar inverse Mill's ratio (the only correction variable used in Greene's (1994) approach). Moreover, it is shown that the quality of this approximation improves as more of these correction variables are added (although a few may be sufficient) since the approximation error is $o(|\rho|^r)$, which becomes small as r increases. This suggests a very simple two-step estimation procedure similar in spirit to Greene's original approach: first, estimate a probit model which determines selection, or endogenous treatment effects, and construct the correction variables (truncated cumulants); at the second stage, add these variables to the set of regressors and obtain parameter estimates based on *Poisson* maximum likelihood (ML) procedures. It is demonstrated that, to $O(|\rho|^r)$, this two-step estimator is consistent (i.e., the inconsistency is $o(|\rho|^r)$) although standard errors may need correcting for: (a) model misspecification and, (b) the generated regressor problem induced by adding the correction variables. Therefore the two-step estimator has the potential to offer a considerable improvement over pseudo maximum likelihood methods which ignore the sample selection and for which the inconsistency is $O(|\rho|)$.[1] Indeed, Monte Carlo evidence suggests that this procedure can work extremely well especially in comparison with Terza's approach, which can exhibit very poor behaviour. Since the quality of the approximation employed is the key to obtaining satisfactory inference, a method for assessing the adequacy of the approximation is also discussed.

The plan of the paper is as follows: in the next section we derive an approximation to the mean of a count variable, subject to selectivity or endogenous treatment effects, which has the same Taylor expansion in ρ, to any order required, as the true mean function given by Terza (1998). In Section 3, the two-step estimator is defined and its asymptotic distribution is given. The inconsistency of this estimator is shown to be at most $O(|\rho|^{r+1})$, thereby offering a potential improvement over Greene's (1994) estimator for which the inconsistency is $O(|\rho|^2)$. Details of the analyses are relegated to Appendices. Section 4 provides a summary of Monte Carlo experiments which investigates the efficacy of the proposed two-step procedure, relative to Greene's (1995) NLS estimator and Terza's (1998) NWLS estimator. Section 5 concludes.

[1] A similar methodology has recently been proposed by Chesher and Santos-Silva (2001) for estimating a heterogeneous logit model. However, in that case, the error in the approximating model can not be made arbitrarily small, as it can for the count data models considered here.

2 The Count Model and Selectivity

To keep the notation and algebra to a minimum, we deal first with the sample selection problem and then show how the results readily extend to the dummy endogenous regressor case.

Consider, then, a count random variable, y^*, which conditional on a $(q \times 1)$ vector of covariates, x and a random variable u, representing possible neglected heterogeneity (over-dispersion), has mean

$$E[y^*|x,u] = \exp(x'\beta + \sigma u) > 0 \tag{1}$$

where β and σ are unknown parameters (vector and scalar, respectively) and the regression specification, $x'\beta$, includes an intercept term. Conditional on x and u, (1) includes the Poisson and Negative Binomial models for y^*; see, for example, Winkelmann (1997). Under sampling, observations $\{y_i = y_i^*, x_i\}$, $i = 1, ..., n$, are only selected if $s_i = 1$ where

$$s_i = \mathbf{1}(z_i'\gamma + \varepsilon_i > 0), \qquad i = 1, ..., N, \tag{2}$$

where $\mathbf{1}(.)$ is the usual indicator function, with z_i being $(l \times 1)$.

2.1 Approximate mean function: Selectivity

As in Terza (1998), and rather than fully specifying a parametric model for y, we specify the mean of the observed count and use this is a basis for estimation. Following previous work (for example, Winkelmann, 1998) it is assumed that (u_i, ε_i) are *iid* standard bivariate normal with correlation ρ. Then $u = \rho\varepsilon + \sqrt{1-\rho^2}\upsilon$, with υ *iid* $\mathcal{N}(0,1)$ independent of ε. From (1) and (2),

$$E[y|x,z] = E_{\upsilon\varepsilon}\left[\exp\left(x'\beta + \sigma\rho\varepsilon + \sigma\sqrt{1-\rho^2}\upsilon\right) | \, s = 1, x, z\right],$$

where expectations are taken first with respect to υ and then with respect to ε, conditionally on $s = 1$. Thus, given x_i and z_i,

$$\begin{aligned} E[y|x_i,z_i] &= \exp(x_i'\beta^\dagger) E[\exp(\sigma\rho\varepsilon) | \varepsilon > -z_i'\gamma] \\ &= \exp\left(x_i'\beta^\dagger + k_i(\eta)\right) \end{aligned} \tag{3}$$

in which $\beta^\dagger = \beta - \frac{1}{2}\sigma^2(1-\rho^2)$, $\eta = \sigma\rho$ and $k_i(\eta)$ is simply the *cumulant generating function* (*cgf*) of a truncated standard normal variate; i.e.,[2]

$$k_i(\eta) = \frac{\eta^2}{2} + \ln\Phi(z_i'\gamma + \eta) - \ln\Phi(z_i'\gamma),$$

[2] It is readily shown that the moment generating function of ε, given $\varepsilon > z_i'\gamma$ is $E[\exp(\eta\varepsilon) | \varepsilon > z_i'\gamma] = \exp\left(\frac{\eta^2}{2}\right)\frac{\Phi(z_i'\gamma+\eta)}{\Phi(z_i'\gamma)}$.

where $\Phi(.)$ denotes the standard normal distribution function and $\phi(.)$ will denote the standard normal density function. An approximation, in $\eta = \sigma\rho$, to (3) is obtained by expanding $k_i(\eta)$ in a Taylor series about $\eta = 0$, giving

$$k_i(\eta) = \sum_{j=1}^{r} \frac{\eta^j}{j!}\kappa_{ij} + o(|\eta|^r), \tag{4}$$

where κ_{ij} denotes the j^{th} cumulant of ε, given $\varepsilon > -z_i'\gamma$. For example, $\kappa_{i1} = \xi_i$, where $\xi_i = \phi(z_i'\gamma)/\Phi(z_i'\gamma)$ (the inverse Mill's ratio) and $\kappa_{i2} = 1 - \xi_i(\xi_i + z_i'\gamma)$, etc. It will be useful to note that κ_{ij} is a function of $z_i'\gamma$; i.e., $\kappa_{ij} = \kappa_j(z_i'\gamma)$ where $\kappa_j(a)$ is the j^{th} cumulant of ε conditional on $\varepsilon > -a$. Substituting (4) into (3) yields

$$E[y|x_i, z_i] = \exp\left(x_i'\beta^\dagger + \sum_{j=1}^{r} \frac{\eta^j}{j!}\kappa_{ij} + o(|\eta|^r)\right). \tag{5}$$

Therefore, the correct mean of the observed (selected) count can be *approximated* as

$$E[y|x_i, z_i] \cong \exp\left(x_i'\beta^\dagger + \sum_{j=1}^{r} \frac{\eta^j}{j!}\kappa_{ij}\right),$$

which differs from the true mean function (3) by terms which are $o(|\eta|^r) = O(|\eta|^{r+1})$. However, since $|\rho| < 1$ and σ is (assumed) finite, the error in this approximation becomes negligible as r increases so that the quality of the approximation improves as more cumulants are added to the regression specification. For example, taking a second order approximation, $r = 2$, we might *model* the conditional mean as

$$E[y|x_i, z_i] \cong \exp(x_i'\beta + \delta_1\xi_i + \delta_2(1 - \xi_i(\xi_i + z_i'\gamma))).$$

Since y is still a count and the approximate mean specification is of the familiar Poisson form with the regression function being linear in parameters (given γ), the results of Gourieroux, Monfort and Trognon (1984) indicate that standard Poisson ML methods could be employed to correct the sample selection bias in the estimation of β. (This is not the case for Greene's (1995) NLS and Terza's (1998) NWLS estimators.)

2.2 Approximate mean function: Dummy Endogenous Regressors

The above analysis readily extends to the case of endogenous treatment effects; i.e., dummy endogenous regressors. In this case s_i, generated by (2), is the dummy endogenous variable included in the regression specification. Similar

manipulations to those employed above reveal that

$$\begin{aligned} E\left[y|x_i,z_i,s_i\right] &= E_{v\varepsilon}\left[\exp\left(x_i'\beta + s_i\alpha + \sigma\rho\varepsilon + \sigma\sqrt{1-\rho^2}v\right)|x_i,z_i,s_i\right] \\ &= E_{\varepsilon}\left[\exp\left(x_i'\beta^{\dagger} + s_i\alpha + \sigma\rho\varepsilon\right)|x_i,z_i,s_i\right] \\ &= \exp(x_i'\beta^{\dagger} + s_i\alpha + k_i^*\left(\eta\right)) \end{aligned}$$

where

$$k_i^*\left(\eta\right), = \frac{\eta^2}{2} + \ln\Phi\left((2s_i-1)(z_i'\gamma+\eta)\right) - \ln\Phi\left((2s_i-1)z_i'\gamma\right);$$

see, for example, Terza (1998). Expanding $k_i^*\left(\eta\right)$ about $\eta = 0$, we can therefore write

$$E\left[y|x_i,z_i,s_i\right] = \left(\exp x_i'\beta^{\dagger} + s_i\alpha + \sum_{j=1}^{r}\frac{\eta^j}{j!}\kappa_{ij}^*\right) + o\left(|\eta|^r\right)$$

where $\kappa_{ij}^* = \frac{\partial^j k_i^*(0)}{\partial\eta^j}$ and the remainder terms are $o\left(|\eta|^r\right)$, which go to zero as r increases.

Specifically, for $j = 1, 2$,

$$\begin{aligned} \kappa_{i1}^* &= \left(2s_i-1\right)\xi_i^*, \qquad \xi_i^* = \frac{\phi\left(z_i'\gamma\right)}{\Phi\left(\left(2s_i-1\right)z_i'\gamma\right)}, \\ \kappa_{i2}^* &= 1 - \xi_i^*\left(\xi_i^* + \left(2s_i-1\right)z_i'\gamma\right) \end{aligned}$$

which provide generalisations to the corrections employed in the sample selection case (where $s_i = 1$ for all i). Notice that $\kappa_{i1}^* = \left(2s_i-1\right)\xi_i^*$ is the first order generalised residual associated with the probit model (2). As before, it will be useful to think of κ_{ij}^* as $\kappa_j^*\left(z_i'\gamma\right)$, a function of $z_i'\gamma$.

3 Two-step Estimation

To make the procedure operational, one first estimates the probit model (2) to obtain $\hat{\gamma}$, which is then used to construct $\hat{\kappa}_{ij}$, or $\hat{\kappa}_{ij}^*$. These *generated* regressors are then added to the regression specification in a simple Poisson model whose parameters are estimated by maximum likelihood at the second step.[3]

Focussing on the sample selection problem, and given γ, the "Poisson" specification at the second step is *modelled* via its approximate mean as

$$\begin{aligned} E\left[y|x_i,\kappa_i\right] &= \exp\left(x_i'\beta + \kappa_i'\delta\right) \\ &= \exp(w_i'\theta), \quad \text{say,} \end{aligned} \tag{6}$$

[3] In the case $r = 1$ only $\hat{\kappa}_{i1} = \hat{\xi}_i$ is added, which is the procedure proposed by Greene (1997). Orme and Peters (2000) discuss similiar corrections in more general models of sample selection and demonstrate that such a procedure provides an approximate model for the observed data which is correct to $O(\eta)$, where $\eta = \sigma\rho$.

where $\delta' = \left(\eta, \frac{\eta^2}{2}, \frac{\eta^3}{3!}, ..., \frac{\eta^r}{r!}\right)$ and $\theta' = (\beta', \delta')$; $\kappa_i' = (\kappa_{i1}, ..., \kappa_{ir})$, and depends on γ, and $w_i' = (x_i', \kappa_i')$. It is assumed that there are no linear dependencies among the regressors, w_i'. Then, using Poisson ML methods, the proposed two-step estimator for θ is the unique solution to the $q + r$ equations

$$H(\theta, \hat{\gamma}) = \sum_{i=1}^{n} h_i(\theta, \hat{\gamma}) = 0 \tag{7}$$

where $h_i(\theta, \gamma) = (y_i - \exp(w_i'\theta))w_i$, δ is unconstrained, and $\hat{\gamma}$ denotes any consistent (eg, Probit) estimator for γ; i.e., the $\kappa_{ij} = \kappa_j(z_i'\gamma)$ are replaced with $\hat{\kappa}_{ij} = \kappa_j(z_i'\hat{\gamma})$ and treated as observed regressors. This solution will be denoted $\hat{\theta}$, and it is unique since $\dfrac{\partial H}{\partial \theta'}$ is negative definite for all θ. Observe that the probability limit of $\hat{\theta}$ is $\theta^* \neq \theta$, in general, since the assumed model specification embodied in (6) is incorrect. However, it is shown in Appendix 1 that $\theta^* = \theta^\dagger + o(|\rho|^r)$, where $\theta^{\dagger\prime} = \left(\beta^{\dagger\prime}, \delta'\right)$. (The probit MLE, $\hat{\gamma}$, which maximises $L^p(\gamma)$, the probit log-likelihood function over $N > n$ observations, is consistent for γ whatever the value of ρ.)

Standard tools, such as those employed by Newey and McFadden (1994), can be exploited to derive the limiting distribution of $\sqrt{n}\left(\hat{\theta} - \theta^*\right)$ under the assumption that $n \to \infty$ such that $n/N \to c$, a finite constant, and independent observations; details are outlined in Appendix 2. The result is

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}) \tag{8}$$

where, to $O(|\rho|^r)$,

$$\mathcal{V} = \mathcal{H}_\theta^{-1}\left\{\mathcal{G} + c\mathcal{H}_\gamma\left(\mathcal{I}_{\gamma\gamma}^p\right)^{-1}\mathcal{H}_\gamma'\right\}\mathcal{H}_\theta^{-1}$$

and

$$\begin{aligned}
\mathcal{H}_\theta &= p\lim\left[\frac{1}{n}\frac{\partial H(\theta^*, \gamma)}{\partial \theta'}\right] = \mathcal{H}_\theta', \\
\mathcal{H}_\gamma &= p\lim\left[\frac{1}{n}\frac{\partial H(\theta^*, \gamma)}{\partial \gamma'}\right] \\
\mathcal{G} &= p\lim\left[\frac{1}{n}\sum_{i=1}^{n} h_i(\theta^*, \gamma)\, h_i(\theta^*, \gamma)'\right] \\
\mathcal{I}_{\gamma\gamma}^p &= -p\lim\left[\frac{1}{N}\frac{\partial^2 L^p(\gamma)}{\partial\gamma\partial\gamma'}\right] = p\lim\left[\frac{1}{N}\frac{\partial L^p(\gamma)}{\partial\gamma}\frac{\partial L^p(\gamma)}{\partial\gamma'}\right].
\end{aligned}$$

The first term in $\mathcal{V}$ is $\mathcal{H}_\theta^{-1}\mathcal{G}\mathcal{H}_\theta^{-1}$, the "sandwich" covariance matrix (White (1982), Gourieroux, Monfort and Trognon (1984)), which arises since the *model* (6) is misspecified; the second term appears due to the fact that generated regressors, $\hat{\kappa}_{ij}$, are employed.

The calculations are particularly simple and lead to the following estimator for the asymptotic variance of $\hat{\theta}$. Let $\hat{W}$ be the $(n \times q+r)$ matrix with rows $\hat{w}_i' = \left(x_i', \hat{\kappa}_i'\right)$; Z the $(N \times l)$ matrix with rows z_i', $i = 1, ..., N$, and Z_1 be the $(n \times l)$ matrix with rows z_i' corresponding to selected sample defined by $s_i = 1$. Then, $\mathcal{H}_\theta, \mathcal{G}$ and $\mathcal{I}^p_{\gamma\gamma}$ can be consistenly estimated by $\frac{1}{n}\hat{W}'\hat{D}_1\hat{W}$, $\frac{1}{n}\hat{W}D_2W$ and $\frac{1}{N}Z'D_3Z$, respectively, where $\hat{D}_1 = diag\left(\exp(\hat{w}_i'\hat{\theta})\right)$, $(n \times n)$ and $\hat{D}_2 = diag\left(\left\{y_i - \exp\left(\hat{w}_i'\hat{\theta}\right)\right\}^2\right)$, $(n \times n)$, are defined using only the observations in the selected equation. The $(N \times N)$ diagonal matrix $\hat{D}_3$ is defined by $\hat{D}_3 = diag\left(\frac{\phi\left(z_i'\hat{\gamma}\right)^2}{\Phi\left(z_i'\hat{\gamma}\right)\Phi\left(-z_i'\hat{\gamma}\right)}\right)$, using all the observations in the probit model. Differentiating $H(\theta, \gamma)$ with respect to γ, yields

$$\frac{1}{n}\frac{\partial H(\theta,\gamma)}{\partial \gamma'} = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \exp(w_i'\theta))\begin{pmatrix} 0 \\ \nabla\kappa_i z_i' \end{pmatrix} - \exp(w_i'\theta)(\delta'\nabla\kappa_i)w_i z_i'\right],$$

where $\nabla\kappa_i = \frac{d\kappa_i}{da}$, $(r \times 1)$, the vector of cumulant derivatives with respect to $a = z_i'\gamma$. Our arguments imply that $E\left[y_i - \exp(w_i'\theta^*)|w_i\right] = o(|\rho|^r)$ so that, consequently, $\mathcal{H}_\gamma = O\left(|\rho|\right)$ and, to $O\left(|\rho|^r\right)$, it is consistently estimated by $-\frac{1}{n}\hat{W}'\hat{D}_4 Z_1$, with $\hat{D}_4 = diag\left(\hat{\delta}'\nabla\hat{\kappa}_i \exp(\hat{w}_i'\hat{\theta})\right)$, $(n \times n)$. It follows that for $r \geq 2$ the asymptotic variance can be estimated as

$$\mathcal{V}_n\left(\hat{\theta}, \hat{\gamma}\right) = \left(\hat{W}'\hat{D}_1\hat{W}\right)^{-1}\hat{W}'\left\{\hat{D}_2 + \hat{D}_4 Z_1\left(Z'\hat{D}_3 Z\right)^{-1}Z_1'\hat{D}_4\right\}\hat{W}\left(\hat{W}'D_1\hat{W}\right)^{-1},$$

whilst for an approximation which is correct to $O\left(|\rho|\right)$, this estimator will be

$$\mathcal{V}_n\left(\hat{\theta}, \hat{\gamma}\right) = \left(\hat{W}'\hat{D}_1\hat{W}\right)^{-1}\hat{W}'\hat{D}_2\hat{W}\left(\hat{W}'D_1\hat{W}\right)^{-1},$$

the usual sandwich estimator, because the correction to the asymptotic variance for the use of constructed regressors will be $o\left(|\rho|\right)$.

For example, in the case of $r = 2$ we have $\kappa_i' = (\xi_i, 1 - \xi_i(\xi_i + z_i'\gamma))$ with

$$\begin{aligned}\nabla\kappa_i &= \begin{pmatrix} -\xi_i(\xi_i + z_i'\gamma) \\ 2\xi_i^2(\xi_i + z_i'\gamma) + \xi_i(\xi_i + z_i'\gamma)(z_i'\gamma) - \xi_i \end{pmatrix} \\ &= \begin{pmatrix} -\xi_i(\xi_i + z_i'\gamma) \\ \xi_i(\xi_i + z_i'\gamma)(2\xi_i + z_i'\gamma) - \xi_i \end{pmatrix}.\end{aligned}$$

4 Monte Carlo Experiments

A simulation experiment was undertaken in order to investigate the performance of the proposed correcting variables in alleviating the problem of sample selection bias in a standard count data regression model. The exploratory variable

design used is taken from Greene (1997), and is a random sample of a much larger database on credit worthiness analysed in Greene (1998). The orignal observed responses were the number of MDRs (Multiple Defaults Records, the number of failures to pay an account). This gave a sample size of $n = 1023$ in the count data response equation, after selection. Sample selection was indicated by ownership of a particular credit card. This gave a full sample size of $N = 1319$. The design coefficients used to generate both the count (β) and Probit (γ) equations were based upon the actual estimates from the single equation fits of the original data. The count equation contained the following variables: *Constant*, *Age*, *Income* and *Average Monthly Card Expenditure/Income*, while the selection equation contained: *Constant*, *Age*, *Income, Has a Major Credit Card, Home-owner, Number of Reported Cards* and *Number of Active Cards.*

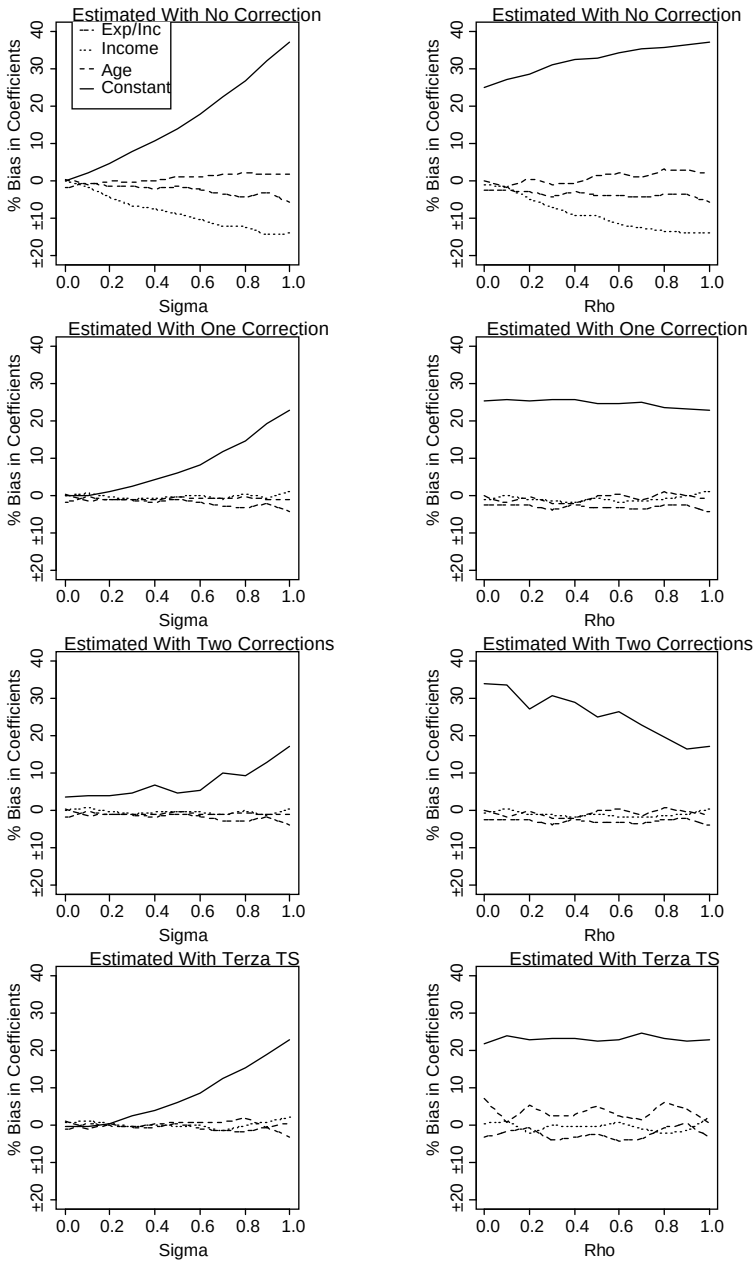
The selection equation (2) is used to generate an observed response for individual i; if the selection response is positive ($\varepsilon_i > -z_i'\gamma$) a corresponding count response is taken as the draw from a Poisson distribution with mean equal to (1), in which $u = \rho\varepsilon_i + \sqrt{1-\rho^2}\upsilon$, with υ iid $\mathcal{N}(0,1)$ independent of ε_i. In one set of experiments, $\rho = 1$, so that $u_i = \varepsilon_i$ and the mean of the observed count is simply $\exp(x_i'\beta + \sigma\varepsilon_i)$. In these experiments, σ varied from 0.0 to 1.0, in increments of 0.1, with $\sigma = 0$ corresponding to a simple Poisson model and $\sigma > 0$ corresponding to a Poisson model suffering sample selection. In the other set of experiments, the value of σ was fixed at $\sigma = 1$ and values of ρ considered were $\rho = 0, 0.1, 0.2, ..., 1.0$; when $\rho = 0$ there is no sample selection problem, but the observed counts are heterogeneous Poisson.

Each experiment involved 1000 replications with each of the following estimators calculated: (i) Poisson, (ii) Poisson with correcting variable $\hat{\kappa}_{i1} = \hat{\xi}_i$, (iii) Poisson model with correcting variables $\hat{\kappa}_{i1}$ and $\hat{\kappa}_{i2} = 1 - \hat{\xi}_i\left(\hat{\xi}_i + z_i'\hat{\gamma}\right)$, (iv) Greene's Non-Linear Least Squares (NLS), and (v) Terza's Non-Linear Weighted Least Squares (NWLS). The resulting average coefficient biases are summarised in Figure 1. Figures 2 and 3 report the ratio of average model standard errors to the simulation Root Mean Square Errors (RMSE). Various standard errors are considered as follows: (i) for the simple Poisson model, standard errors are derived from the Hessian matrix and the sandwich matrix; (ii) for the Poisson specifications that include $\hat{\kappa}_{i1}$, or both $\hat{\kappa}_{i1}$ and $\hat{\kappa}_{i2}$, they are derived from the corresponding sandwich matrix and the corrected variance matrix, $\mathcal{V}$; and, (iii) for Greene's (1995) NLS and Terza's (1998) NWLS estimators, the standard errors are calculated using the formulae provided by these authors. Tables A and B report the performance of Wald procedures when used to test for sample selection bias (testing if inclusion of the generated regressors, or correction variable for Terza's model, is significant).[4] Results are reported to 3 significant figures.[5]

[4] A likelihood ratio procedure, based on Poisson ML estimation is not appropriate (even approximately) due to the possibility of overdispersion.

[5] All computations were performed using Ox (Doornik,2001) running under RedHat Linux 5.1. The simulations used the default random number generator of Ox 2.10. ML estimation was performed using Newton-Raphson optimisation, NLS estimation used Gauss-Newton regression. Data configurations that caused problems for NLS estimation were ignored. This

Figure 1: Coefficient Biases as Sigma/Rho Increase



The *first column* of Figure 1 plots the percentage bias of the (β) coefficient estimates as σ increases, with $\rho = 1$. The coefficients can be identified from the legend on the first plot, which illustrates the simple Poisson case. Here, as expected, the bias on the constant coefficient increases as the sample selection problem becomes worse. There is also evident bias in the remaining coefficient estimators but this appears cured by including one or both of the constructed variables for the ML case (second and third plots), or by using Terza's NWLS estimator (last plot). The *second column* deals with the situation where there is a joint heterogeneity ($\sigma = 1$) and sample selection problem (with $\rho = 0$ resulting in heterogeneity only, and $\rho = 1$ sample selection only). In the standard model (first plot) pure heterogeneity is only seriously affecting the intercept coefficient estimator, however, as the sample selection effects become stronger, the remaining "slope" coefficient estimates also begin to exhibit serious bias. This, again, is largely rectified by using correcting variables in the Poisson case (second and third plots), though the biases are marginally larger as the total overdispersion in these experiments is greater than those reported in the first column. Interestingly, these results suggest that the two-step procedure works better than Terza's NWLS estimator (last plot), in this case.[6]

The standard errors also exhibit an interesting picture, with the overall effect of overdispersion in this design resulting in under estimation of coefficient estimator variability. In Figure 2, for the experiment with sample selection only ($\rho = 1$, $\sigma > 0$), the Poisson ML standard errors, based on the Hessian, are seriously biased downwards (first plot, *column 1*). At worst, the slope coefficient standard errors are approximately half the size they should be. This is partially alleviated when the robust sandwich covariance matrix estimator is used (first plot, *column 2*). Moreover, the addition of the first correcting variable in the two-step procedure also improves matters considerably, while the addition of the second correcting variable also improves the estimated variability of the intercept estimator, when using the sandwich estimator (second and third plots, *column 1*). The performance of the corrected variance matrix estimator (second and third plots, *column 2*) is very similar to the corresponding sandwich matrix when either one or two correcting variables are employed. Both Greene's NLS and Terza's NWLS estimated covariance matrix performs poorly, exhibiting behaviour similar to the raw Hessian's performance.[7]

Figure 3 reports the corresponding standard error to RMSE ratios for the experiments where there is both heterogenety and sample selection present ($\sigma = 1$, $\rho > 0$). The extra overdispersion caused by the heterogeneity has depressed the hessian-based standard errors considerably. On the other hand, when the two-step procedure is employed, in conjunction with either the sandwich or corrected matrix estimators, there is again substantial improvement. However,

occured when overdispsersion was severe (the highest number of occurences was six for Terza's estimator when ρ equalled 0.2).

[6] The results from using Greene's NLS specification are not reported here, though its biases exhibited similar behaviour to Terza's.

[7] The performance of the outer product of gradient (OPG) covariance matrix estimator is unreported. Its performance was worse than the Hessian's in all cases investigated.

Figure 2: Standard Error/RMSE, Sigma Varies

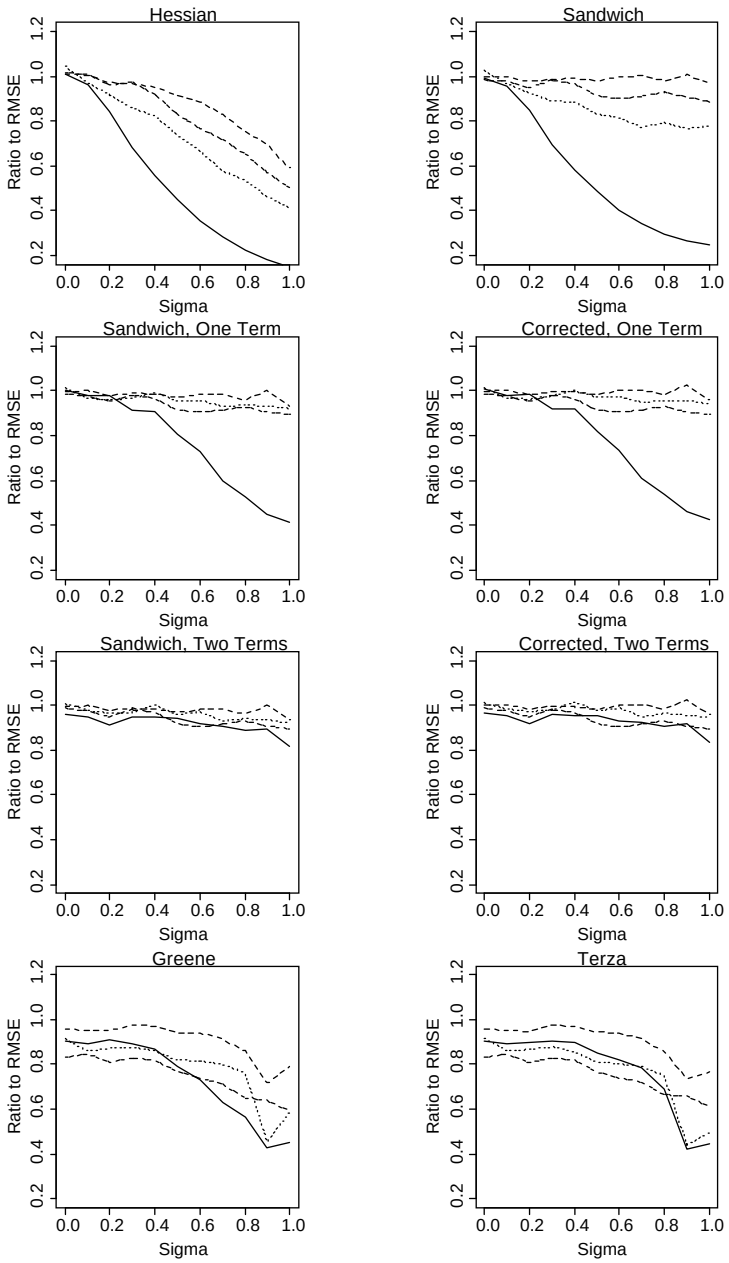
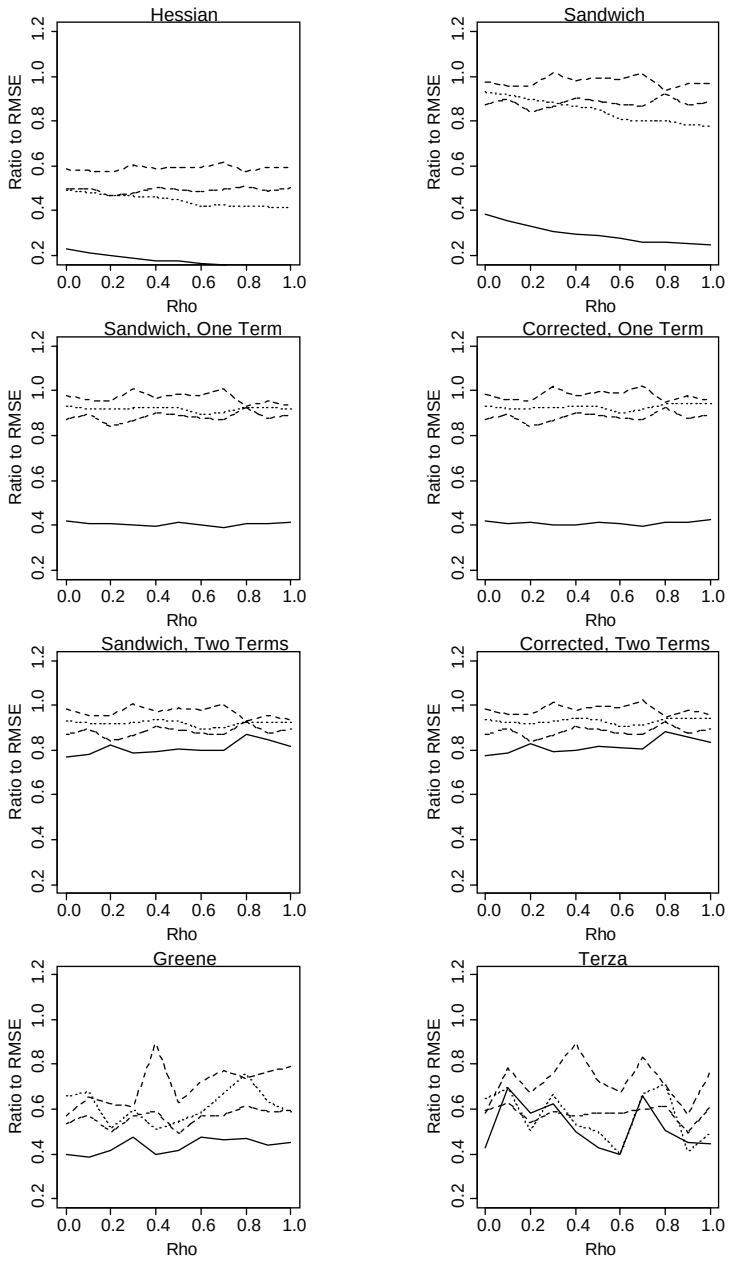


Figure 3: Standard Error/RMSE, Rho Varies.



both the NLS and NWLS standard errors again appear to behave more like those obtained from the raw ML Hessian, and seriously underestimate the true variability of the model coefficient estimates.

In conclusion, then, the two-step procedure analysed in this paper appears to work remarkably well relative Greene's NLS and Terza's NWLS estimators. With the addition of just a couple of the correcting variables, the sample selection bias in the slope estimators and corresponding standard errors, based on the sandwich or corrected covariance matrix estimator, is largely removed.

At this stage, only the results associated with the regressors of interest have been reported. Assessing the statistical significance of including the correcting variables, in the two-step procedure, can be simply undertaken using a Wald test. With reference to the mean specification given by (6), the hypotheses tested are: (i) $\delta_1 = 0$, given $\delta_2 = 0$; (ii) $\delta_1 = \delta_2 = 0$; and (iii) $\delta_2 = 0$, given $\delta_1 \neq 0$. The last is included as a means of assessing the adequacy of the two-step estimator which only employs the first correcting variable. The motivation, here, is that if the addition of the second correcting variable is statistically insignificant, this provides evidence in support of inferential procedures based on the two-step technique which employs just the first (Heckman) correcting variable. Table A reports the rejection rates of the various tests. Table B reports the rejection rates of Wald tests designed to check the statistical significance of, respectively, the Greene and Terza correcting variables employed in their NLS, respectively NWLS, procedures. Furthermore, we employ the following tests based upon sample rejection rates to assess how close finite sample significance levels α_n are to nominal values α, where α_n and α are both measured as percentages: Let the null hypothesis to be tested be that the actual finite sample significance level α_n satisfies $H_\alpha : \alpha_L \leq \alpha_n \leq \alpha_U$, where $\alpha_U - \alpha_L$ is $O(1)$, but small, e.g. 1%, and $\alpha_U + \alpha_L = 2\alpha$. The asymptotic (as R, the number of replications, goes to infinity) significance level is 5% and the implied decision rule is that H_α is rejected if the observed rejection frequency is outside the interval

$$\alpha_L - 1.645\sqrt{\frac{\alpha_L(100-\alpha_L)}{R}}, \quad \alpha_U + 1.645\sqrt{\frac{\alpha_U(100-\alpha_U)}{R}}.$$

For $R = 1000$ estimated rejection rates in the ranges $7.97 - 12.09$, $3.42 - 6.69$ and $0.13 - 2.13$ are deemed to be consistent with true nominal significance levels being within $\pm 0.5\%$ of the nominal 10%, 5% and 1% levels, respectively, and this is indicated by an asterisk (*) in the tables.

The first column of each table deals with the experiments when $\rho = 0$ and σ varies (sample selection only), while the second column deals with the case when $\sigma = 1$ and ρ varies (heterogeneity and sample selection). The first two upper blocks of Table A are estimated rejection rates when testing for the inclusion of the variable $\hat{\kappa}_{i1}$ and, then, both $\hat{\kappa}_{i1}$ and $\hat{\kappa}_{i2}$, respectively. The third block reports rejection rates when testing for the inclusion of $\hat{\kappa}_{i2}$, given the inclusion of $\hat{\kappa}_{i1}$ as a correction for the sample selection problem in the two-step estimation procedure.

Table A

Rejection rates for the Wald tests in the two-step procedure

	$\hat{\kappa}_{i1}$ only				$\hat{\kappa}_{i1}$ only		
σ	10%	5%	1%	ρ	10%	5%	1%
0.0	10.8*	5.7*	1.5*	**0.0**	10.2*	5.1*	1.0*
0.1	17.2	10.1	3.8	**0.1**	11.2*	6.2*	1.8*
0.2	26.0	18.0	6.9	**0.2**	19.0	11.6	4.3
0.3	41.9	30.5	13.7	**0.3**	31.2	20.7	7.0
0.4	52.5	40.6	19.7	**0.4**	41.4	30.1	12.9
0.5	67.6	56.2	35.0	**0.5**	53.3	42.9	23.8
0.6	79.2	70.0	46.3	**0.6**	62.5	50.8	28.8
0.7	85.9	77.0	56.8	**0.7**	73.7	62.2	39.9
0.8	89.9	83.4	65.5	**0.8**	83.9	73.7	50.2
0.9	92.7	87.5	70.0	**0.9**	89.0	82.7	63.2
1.0	92.7	87.9	74.4	**1.0**	92.7	87.9	74.4
	$\hat{\kappa}_{i1}$ and $\hat{\kappa}_{i2}$				$\hat{\kappa}_{i1}$ and $\hat{\kappa}_{i2}$		
σ	10%	5%	1%	ρ	10%	5%	1%
0.0	11.4*	6.2*	1.1*	**0.0**	12.3	6.1*	0.6*
0.1	14.7	10.5	3.9	**0.1**	13.5	7.5	1.8*
0.2	24.7	17.3	7.0	**0.2**	19.0	10.8	3.5
0.3	35.2	23.9	12.0	**0.3**	26.6	18.7	4.7
0.4	45.8	33.5	16.3	**0.4**	37.2	26.6	10.8
0.5	61.6	49.5	27.2	**0.5**	49.5	38.0	20.3
0.6	75.9	63.6	40.1	**0.6**	57.1	43.3	23.2
0.7	79.9	71.0	49.5	**0.7**	69.1	57.0	36.5
0.8	87.2	80.2	62.1	**0.8**	77.7	66.5	47.6
0.9	89.9	83.6	66.8	**0.9**	86.5	79.4	61.5
1.0	92.5	86.4	71.6	**1.0**	92.5	86.4	71.6
	$\hat{\kappa}_{i2}$, given $\hat{\kappa}_{i1}$				$\hat{\kappa}_{i2}$, given $\hat{\kappa}_{i1}$		
σ	10%	5%	1%	ρ	10%	5%	1%
0.0	11.5*	5.9*	1.0*	**0.0**	12.2	6.8	1.1*
0.1	9.7*	5.4*	1.3*	**0.1**	11.7*	5.8*	1.2*
0.2	13.2	6.8	1.9*	**0.2**	10.7*	5.6*	1.3*
0.3	11.7*	5.1*	0.7*	**0.3**	12.0*	5.9*	0.9*
0.4	11.9*	5.7*	1.4*	**0.4**	11.9*	6.0*	1.5*
0.5	11.4*	6.8	1.1*	**0.5**	12.5	5.8*	1.0*
0.6	12.2	6.1*	1.5*	**0.6**	12.3	6.4*	1.3*
0.7	11.3*	5.9*	1.5*	**0.7**	13.2	6.9	2.1*
0.8	13.4	7.2	1.6*	**0.8**	10.9*	6.1*	1.7*
0.9	11.7*	6.3*	1.1*	**0.9**	12.8	7.4	1.7
1.0	14.2	7.8	2.4	**1.0**	14.2	7.8	2.4

When either $\sigma = 0$ or $\rho = 0$, the results indicate that the usual asymptotic null approximation to finite sample distribution of the Wald test statistic is adequate, at the 5% nominal significance level, when either one or two correcting variables are present and, in particular, that the tests are robust to the

presence of heterogeneity (as they should be asymptotically).[8] Moreover, the lower block of Table A reveals relative insensitivity to the inclusion of the extra correcting variable, indicating that inferential procedures may be taken to be approximately valid if based on the two-step procedure employing just the first correcting variable; i.e., the $O(|\rho|)$ approximation appears reasonable. This is consistent with Figures 1-3 where there was little gain from including the second correcting variable unless inference concerning the intercept coefficient was of concern. Indeed, since it was previously noted in Section 3 that the sandwich agrees with the corrected covariance matrix estimator, to $O(|\rho|)$, the fact that the Wald test indicates that the $O(|\rho|)$ is adequate provides some explanation of the observed close agreement between these two covariance matrix estimators when employing just the first correcting variable (see Figures 2 and 3).

Table B

Rejection Rates using NLS or NWLS estimation.

Greene's NLS correction				Greene's NLS correction			
σ	10%	5%	1%	ρ	10%	5%	1%
0.0	11.8*	6.4*	2.1*	**0.0**	14.6	8.9	3.6
0.1	17.4	10.7	3.8	**0.1**	15.5	9.7	3.0
0.2	25.8	16.6	7.3	**0.2**	21.8	13.8	5.5
0.3	37.2	27.5	12.9	**0.3**	29.4	18.8	7.3
0.4	48.2	36.9	19.5	**0.4**	36.3	26.6	12.2
0.5	61.2	49.0	27.3	**0.5**	47.3	37.5	20.3
0.6	69.1	57.2	35.8	**0.6**	53.3	43.0	24.4
0.7	77.4	67.7	47.5	**0.7**	62.3	50.7	30.4
0.8	81.8	75.3	57.0	**0.8**	69.4	60.8	42.4
0.9	82.0	75.5	56.3	**0.9**	79.1	70.9	54.2
1.0	84.4	78.7	63.6	**1.0**	84.4	78.7	63.6
Terza's NWLS correction.				Terza's NWLS correction.			
σ	10%	5%	1%	ρ	10%	5%	1%
0.0	10.7*	6.4*	3.1	**0.0**	13.5	9.0	5.1
0.1	12.6	6.5	1.8	**0.1**	11.8	8.1	4.4
0.2	16.6	9.1	2.6	**0.2**	14.8	8.6	4.5
0.3	26.9	14.4	4.3	**0.3**	14.6	7.4	3.5
0.4	36.3	20.7	6.7	**0.4**	20.7	11.1	5.0
0.5	48.7	29.2	8.5	**0.5**	27.6	15.7	6.2
0.6	54.2	34.9	13.2	**0.6**	32.6	19.0	6.3
0.7	64.7	46.1	18.5	**0.7**	41.4	24.7	10.2
0.8	68.2	50.0	21.5	**0.8**	47.7	31.8	13.6
0.9	66.7	46.3	20.6	**0.9**	57.8	41.3	17.2
1.0	63.0	46.7	22.0	**1.0**	63.0	46.7	22.0

For the Greene NLS and Terza NWLS procedures (Table B), the finite sample

[8] The χ^2 forms of the test have been used here. This is just the empirical t-test squared when one variable is tested, but $\widehat{\delta}.V[\widehat{\delta}]^{-1}.\widehat{\delta}'$ when two variables are tested together. $V[\widehat{\delta}]$ is the sub-block of the *corrected* variance covariance matrix associated with the estimated coefficients, $\widehat{\delta}$, of the correction variables.

behaviour of the Wald tests does not compares well with two-step ML estimator. These results indicate that the tests are mildly over-sized under the null ($\rho = 0$, $\sigma = 1$) and, in terms of detecting the statistical significance of the sample selection correcting variable, the power of these tests are dominated by the Wald tests based on the two-step ML procedure (when either $\rho > 0$ or $\sigma > 0$). Also note the drop in power associated with Terza's estimator which appears to be because, when there is a serious sample selection problem, the correction factor can become large with a larger variance and smaller test statistic value.[9]

5 Conclusion

This article has presented an extension to, and theoretical justification of, the empirical practice of using Heckman-type two-step estimators to correct for the sample selection and related problems in count data models. The performance of the extended method was investigated in a simulation study and found to perform very well in correcting for sample selection bias in a model's coefficient estimates, being much preferable to two non-linear least squares estimators previously suggested in the literature (especially when there is extra non-sample selection overdispersion present in the data).

On the basis of the evidence suggested here, the suggested empirical practice would be to use the two correction variables $\hat{\kappa}_{i1}$ and $\hat{\kappa}_{i2}$ to compensate for the presence of a sample selection problem, and to test for it using a robust Wald test. The test has its best performance when calculated using the corrected covariance matrix form, $\mathcal{V}_n\left(\hat{\theta}, \hat{\gamma}\right)$, although inference about the variates of interest could be adequately taken using the sandwich matrix. The two-step ML estimator and associated inferential procedures advocated in this paper can be implemented in a variety of different econometric programs and avoid any computational difficulties associated with other methods (FIML or NLS/NWLS). Moreover, Wald procedures to assess the significance of $\hat{\kappa}_{i2}$ given $\hat{\kappa}_{i1}$, could also provide valuable information as to the appropriateness of simply employing the first correcting variable, $\hat{\kappa}_{i1}$, as originally suggested by Greene (1994).

Further research might usefully examine whether or not the method is adequate in more complex count data situations (e.g. zero-inflated count data models).

[9] Terza estimates η in $\Phi\left(z_i'\gamma + \eta\right)/\Phi\left(z_i'\gamma\right)$. When η is large with respect to $z_i'\gamma$, the correction tends to either 0 or $1/\Phi\left(z_i'\gamma\right)$, and the derivative of the mean function with respect to η becomes close to zero. This can cause problems with the calculation of the covariance matrix used by Terza(1998), and with elements of the optimisation.

Appendix 1: Consistency of $\hat{\theta}$

In order to investigate the consistency properties of the two-step estimator, define the *true* parameter value $\theta_0^{\dagger\prime} = \left(\beta_0^{\dagger\prime}, \delta_0'\right)$, where $\delta_0' = \left(\eta_0, \frac{\eta_0^2}{2!}, ..., \frac{\eta_0^r}{r!}\right)$, and $\eta_0 = \sigma_0\rho_0$. Dealing specifically with the sample selection case, we now show that the two-step estimator, $\hat{\theta}$, satisfies $p\lim\hat{\theta} = \theta_0^{\dagger} + o(|\rho_0|^r)$, where ρ_0 is the true value of ρ (similar analysis will apply for the dummy endogenous regressor case).

Under fairly general conditions, such as (x_i, z_i) being *iid*, $\theta^* = p\lim\hat{\theta}$ will be the unique solution to $E[h(\theta^*, \gamma)] = 0$, observing that $p\lim\hat{\gamma} = \gamma$, and where expectations are conditional on $s = 1$. The solution is unique, since $\dfrac{\partial E[h(\theta,\gamma)]}{\partial\theta'} = -E[\exp(w'\theta)ww']$ is negative definite for all θ and γ, assuming that there are no linear dependencies among the elements of $w' = (x', \kappa')$. Furthermore, $E[\mu ww']$ will be positive definite for any positive random variable, μ. To proceed, re-parameterise the problem from θ to $\psi = \theta - \theta_0^{\dagger}$ so that $\psi^* = p\lim\hat{\psi}$ equals zero *iff* $\theta^* = \theta_0^{\dagger}$. From (3), ψ^* is the unique solution to

$$E\left[\lambda_0^{\dagger}\left\{\exp\left(k\left(\eta_0\right)\right) - \exp\left(w'\psi^*\right)\exp\left(\kappa'\delta_0\right)\right\}w\right] = 0, \tag{9}$$

where $\lambda_0^{\dagger} = \exp\left(x'\beta_0^{\dagger}\right) > 0$. Since non-zero η_0 (through non-zero ρ_0) will introduce inconsistency in the parameter estimator, $\hat{\theta}$, we regard ψ^* as a function of η_0; i.e., $\psi^* = \psi^*\left(\eta_0\right)$, such that $\psi^*\left(0\right) = 0$. Then, expanding ψ^* about $\eta_0 = 0$ we may write

$$\psi^*\left(\eta_0\right) = \sum_{j=1}^{r}\frac{\partial^j\psi^*\left(0\right)}{\partial\eta_0^j}\delta_{0j} + O\left(|\eta_0|^{r+1}\right)$$

where $\delta_{0j} = \eta_0^j/j!$. To obtain expressions for $\frac{\partial^j\psi^*(0)}{\partial\eta_0^j}$, repeatedly differentiate (9) with respect to η_0 and evaluate the result at $\eta_0 = 0$. This yields,

$$E\left[\lambda_0^{\dagger}\left\{\frac{\partial^j\exp\left(k\left(\eta_0\right)\right)}{\partial\eta_0^j} - \frac{\partial^j\exp\left(\kappa'\delta_0\right)}{\partial\eta_0^j}\exp\left(w'\psi^*\right) - \frac{\partial^j\exp\left(w'\psi^*\right)}{\partial\eta_0^j}\exp\left(\kappa'\delta_0\right)\right\}w\right] = 0, \tag{10}$$

where the following recursion holds

$$\frac{\partial^j\exp\left(w'\psi^*\right)}{\partial\eta_0^j} = \sum_{m=0}^{j-1}\binom{j-1}{m}\frac{\partial^m\exp\left(w'\psi^*\right)}{\partial\eta_0^m}\frac{\partial^{j-m}\left(w'\psi^*\right)}{\partial\eta_0^{j-m}}. \tag{11}$$

Assuming that $\frac{\partial^m\psi^*(0)}{\partial\eta_0^m} = 0$ for all $m = 1, ..., j-1$, implies

$$\frac{\partial^j\exp\left(w'\psi^*\right)}{\partial\eta_0^j}|_{\eta_0=0} = w'\frac{\partial^j\psi^*\left(0\right)}{\partial\eta_0^j}. \tag{12}$$

Then, because $k(\eta) = \kappa'\delta + o(|\eta|^r)$ implies $\frac{\partial^j \exp(k(\eta_0))}{\partial \eta_0^j}|_{\eta_0=0} = \frac{\partial^j \exp\left(\kappa'\delta_0\right)}{\partial \eta_0^j}|_{\eta_0=0}$ for all $j \leq r$, and $\psi^*(0) = 0$, the j^{th} partial derivative of (9) evaluated at $\eta_0 = 0$ satisfies

$$-E\left[\lambda_0^\dagger w w'\right] \frac{\partial^j \psi^*(0)}{\partial \eta_0^j} = 0, \qquad j \leq r. \tag{13}$$

Therefore $\frac{\partial^j \psi^*(0)}{\partial \eta_0^j} = 0$, $j \leq r$, because $E\left[\lambda_0^\dagger w w'\right]$ is positive definite, and $\psi^*(\eta_0) = O\left(|\eta_0|^{r+1}\right)$.

By assuming that $\frac{\partial^m \psi^*(0)}{\partial \eta_0^m} = 0$, for all $m = 1, ..., j-1$, it was possible to show that $\frac{\partial^j \psi^*(0)}{\partial \eta_0^j} = 0$ and, in particular, if $\frac{\partial \psi^*(0)}{\partial \eta_0} = 0$, then $\frac{\partial^2 \psi^*(0)}{\partial \eta_0^2} = 0$, implying $\frac{\partial^j \psi^*(0)}{\partial \eta_0^j} = 0$, for all $j = 1, ..., r$. Thus, to prove that $\frac{\partial \psi^*(0)}{\partial \eta_0} = 0$, evaluate (10) at $\eta_0 = 0$ and $j = 1$. Thus yields

$$E\left[-\lambda_0^\dagger w w'\right] \frac{\partial \psi^*(0)}{\partial \eta_0} = 0,$$

from which it follows that $\frac{\partial \psi^*(0)}{\partial \eta_0} = 0$.

We conclude that $\theta^* = \theta^\dagger + o(|\eta_0|^r)$, in which the terms denoted $o(|\eta_0|^r)$ are also $o(|\rho_0|^r)$ and become negligible as r increases.

Appendix 2: Asymptotic Distribution of $\hat{\theta}$

In this appendix we derive the asymptotic distribution of the general two-step estimator defined in Section 3. The arguments involved follow Newey and McFadden (1994). As in the main text, let θ^* denote $p\lim\hat{\theta}$, in which $\hat{\theta}$ is the two-step estimator defined as the solution to (7) and let $\hat{\gamma}$ be the probit MLE from the selection equation which is consistent for γ; see Section 3.1.

Expanding $H(\hat{\theta},\hat{\gamma})=0$ about $\hat{\theta}=\theta^*$ and $\hat{\gamma}=\gamma$, yields

$$0=\frac{1}{\sqrt{n}}H\left(\theta^*,\gamma\right)+\mathcal{H}_\theta\sqrt{n}\left(\hat{\theta}-\theta^*\right)+\sqrt{c_N}\mathcal{H}_\gamma\sqrt{N}\left(\hat{\gamma}-\gamma\right)+o_p(1)$$

where $c_N=n/N\to c$. Re-arranging and substituting the linear expansion

$$\sqrt{N}\left(\hat{\gamma}-\gamma\right)=\left(\mathcal{I}^p_{\gamma\gamma}\right)^{-1}\frac{1}{\sqrt{N}}L^p_\gamma\left(\gamma\right)+o_p(1)$$

for the Probit maximum likelihood estimator, where $L^p_\gamma\left(\gamma\right)=\dfrac{\partial L^p\left(\gamma\right)}{\partial\gamma}$, gives

$$\sqrt{n}\left(\hat{\theta}-\theta^*\right)=-\mathcal{H}_\theta^{-1}\left[\frac{1}{\sqrt{n}}H\left(\theta^*,\gamma\right)+\sqrt{c}\mathcal{H}_\gamma\left(\mathcal{I}^p_{\gamma\gamma}\right)^{-1}\frac{1}{\sqrt{N}}L^p_\gamma\left(\gamma\right)\right]+o_p(1). \tag{14}$$

Now, note that $L^p_\gamma\left(\gamma\right)=\sum_{i=1}^N e_iz_i$, where $e_i=\left(2s_i-1\right)\xi_i^*$, the first order Probit generalised error, and $H\left(\theta^*,\gamma\right)=\sum_{i=1}^N s_ih_i(\theta^*,\gamma)$. Then, due to independence of observations,

$$p\lim\left\{n^{-1/2}H\left(\theta^*,\gamma\right)\times N^{-1/2}L^p_\gamma\left(\gamma\right)'\right\}=\frac{1}{\sqrt{c}}\,p\lim N^{-1}\sum_{i=1}^N\left\{s_ih_i(\theta^*,\gamma)e_iz_i'\right\}.$$

But each term in the sum on the right hand side is zero when $s_i=0$ and, to the order of approximation employed in this paper, has expectation zero when $s_i=1$ (because, $E\left[y-\exp(w'\theta^*)|w\right]=o(|\rho|^r)$). Thus, $N^{-1}\sum_{i=1}^N\left\{s_ih_i(\theta^*,\gamma)e_iz_i'\right\}\xrightarrow{p}0$. Therefore assuming a suitable central limit theorem can be applied to $\frac{1}{\sqrt{n}}H\left(\theta^*,\gamma\right)$ and $\frac{1}{\sqrt{N}}L^p_\gamma\left(\gamma\right)$, both of which have zero mean, we obtain

$$\sqrt{n}\left(\hat{\theta}-\theta^*\right)\xrightarrow{d}\mathcal{N}\left(0,\mathcal{V}\left(\theta^*,\gamma\right)\right) \tag{15}$$

where

$$\mathcal{V}\left(\theta^*,\gamma\right)=\mathcal{H}_\theta^{-1}\mathcal{G}\mathcal{H}_\theta^{-1}+c\mathcal{H}_\theta^{-1}\mathcal{H}_\gamma\left(\mathcal{I}^p_{\gamma\gamma}\right)^{-1}\mathcal{H}_\gamma'\mathcal{H}_\theta^{-1}.$$

and we have used the standard result that $\frac{1}{\sqrt{N}}L^p_\gamma\left(\gamma\right)\xrightarrow{d}\mathcal{N}\left(0,\mathcal{I}^p_{\gamma\gamma}\right)$.

References

[1] Cameron, C. and Trivedi, P. (1997). *Models for Count Data.* Oxford University Press.

[2] Doornik, J.A. (2001). Ox: An Object-Orrientated Matrix Programming Language, London: Timberlake Consultants Press.

[3] Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo-maximum likelihood methods: applications to Poisson models, *Econometrica*, 52, 701-720.

[4] Greene, W.H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomila Regression Models. *Working Paper No. EC-94-10,* Department of Economics, Stern School of Business, New York University.

[5] Greene, W.H. (1995). Sample Selection in the Poisson Regression Model. *Working Paper No. EC-95-6,* Department of Economics, Stern School of Business, New York University.

[6] Greene, W.H. (1997). FIML Estimation of Sample Selection Models for Count Data. *Unpublished mimeo.* New York University.

[7] Greene,W.H. (1998). Sample Selection in Credit Scoring Models, *Japan and the World Economy*, 10, 299-316.

[8] Newey, W.K. and D.L. McFadden (1994). *Large Sample estimations and Hypothesis Testing*, in *Handbook of Econometrics, Volume 4,* edited by Engle, R.F., and D.L. McFadden. Amstersdam: Elsevier Science B.V.

[9] Orme, C.D. and S.A. Peters (2000). Linear Approximations to maximum Likelihood Models with Selectivity. *Unpublished mimeo.* University of Manchester.

[10] Terza, J.V. (1998). Estimating count Data Models With Endogenous Switching: Sample Selection and Endogenous Treatment Effects. *J. Econometrics*, 84, 129-154.

[11] White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50, 1-25.

[12] Winkelman, R. (1997). *Econometric Analysis of Count Data.* 2nd Ed. Heidelberg, Germany: Springer-Verlag.

[13] Winkelman, R. (1998). Count Data Models With Selectivity. *Econometric Reviews*, 17, 339-360.