Sample Selection in the Poisson Regression Model

by

William H. Greene

May 8, 1995

Abstract

We present a correction for sample selectivity in the Poisson regression model for count data. The model is similar to that devised by Heckman for the linear regression model. Estimation by a two step method is suggested using nonlinear least squares at the second step.

The model described here was presented in Greene (1994). Terza (1995) describes an alternative approach that has a more orthodox specification of the regression function. We show in this note that Terza's approach is essentially the same as Greene's.

William H. Greene
Department of Economics
Stern School of Business
New York University
44 West 4th St.
New York, NY 10012
Phone: 212-998-0876
Internet: wgreene@stern.nyu.edu

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

1. Introduction

Heckman's (1976, 1979) model has provided the standard approach for accounting for sample selectivity in the linear regression model. The model is built around a classical regression model,

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i.$$

Data on the dependent variable, y_i, and regressors, $\mathbf{x}_i$, are only observed when an associated variable, z_i crosses a threshold. The selection equation is usually specified in the form of a binary choice model for inclusion in the sample:

$$z_i^* = \boldsymbol{\alpha}'\mathbf{w}_i + u_i$$

$$z_i = 1 \text{ iff } z_i^* > 0 \text{ and } 0 \text{ otherwise.}$$

The disturbances, ϵ_i and u_i are assumed to have bivariate normal distribution with zero means, variances σ^2 and 1, and correlation ρ, so the second equation corresponds to the familiar probit model. In the selected population,

$$E[y_i|\mathbf{x}_i, z_i = 1] = \boldsymbol{\beta}'\mathbf{x}_i + \rho\sigma M_i$$

$$M_i = \tfrac{\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(\boldsymbol{\alpha}'\mathbf{w}_i)}.$$

If $\rho \neq 0$, ordinary least squares produces inconsistent estimates of $\boldsymbol{\beta}$. Heckman's approach to estimation involves first estimating $\boldsymbol{\alpha}$ by maximum likelihood in the probit model, then estimating $(\boldsymbol{\beta}, \theta)$ $(\theta{=}\rho\sigma)$ by least squares regression of y_i on $\mathbf{x}_i$ and M_i, in which the latter is computed using the estimates computed at step 1. Heckman (1979) and Greene (1981) give details on computation of the appropriate standard errors for the estimates. The model has recently been criticized for the robustness of the normality assumption, but it remains the approach of choice in many applied studies.

This note will present extensions of Heckman's model and approach to the Poisson regression model. The Poisson regression model for count data is

$$\begin{aligned} Prob(y_i = j) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, j = 0, 1., , , \\ \lambda_i &= e^{\boldsymbol{\beta}'\mathbf{x}_i} \end{aligned}$$

The conditional mean is λ_i. Maximum likelihood estimation of the Poisson regression model is straightforward and routine with currently available software. Greene (1994) extended this model to Heckman's framework by specifying the probit model as before, then respecifying the conditional mean function as

$$E[y_i|\mathbf{x}_i, z_i = 1] = e^{\boldsymbol{\beta}'\mathbf{x}_i + \theta M_i}$$

where M_i is as before. Once again, data on y_i and $\mathbf{x}_i$ are observed only when z_i equals 1. The two step estimation procedure begins by fitting the probit model by maximum likelihood, as earlier, computing estimates of M_i for the selected observations, then fitting the Poisson model with this augmented conditional mean function by maximum likelihood. The standard errors for the latter are adjusted using the results of Murphy and Topel (1985).

Terza (1995) argues that the connection between the selection model and the Poisson conditional mean function is ill defined. He formally derives a counterpart to Heckman's model for the Poisson regression which leads to a different, albeit similar conditional mean function, and suggests a two step estimator based on the probit model followed by nonlinear least squares. In this paper, we will show how Terza's model reduces to Greene's when a linear Taylor series approximation to the conditional mean function is employed. We then reconsider the estimation techniques, and examine the results that three different estimators produce. In Section 2, the formalities of Terza's and Greene's frameworks are given. An application is presented in Section 3. Some conclusions are drawn in Section 4.

2. Sample Selection Models for Count Data

Greene's (1994) proposal for the Poisson model is essentially to mimic the conditional mean function in Heckman's framework. Thus, he proceeds directly from the selection mechanism to

$$\ln E[y_i|\mathbf{x}_i, z_i = 1] = \boldsymbol{\beta}'\mathbf{x}_i + \theta M_i = \ln \lambda_i.$$

The functional form is suggested as an ad hoc method of incorporating the selection mechanism into the Poisson model. He then suggests the same two step approach as Heckman, a probit model in step 1 followed, in this case, by maximum likelihood estimation of $(\boldsymbol{\beta}, \theta)$ in step 2. If this conditional distribution is correctly specified, the inverted Hessian of the log-likelihood would an appropriate estimator for the asymptotic covariance matrix of the MLE. Since M_i has been estimated using parameters estimated at an earlier step, the estimated asymptotic covariance matrix must be adjusted. Murphy and Topel's (1985) results are used to obtain the corrected covariance matrix: Let $\mathbf{X}$ and $\mathbf{W}$ denote the data matrices whose N rows are $\mathbf{x}_i'$ and $\mathbf{w}_i'$ and let $\mathbf{\Sigma}$ be the asymptotic covariance matrix of the probit maximum likelihood estimator of $\boldsymbol{\alpha}$. Then, let

$$v_i = y_i - \lambda_i,$$

$$\mathbf{V} = diag[v_i],$$

$$\delta_i = (\boldsymbol{\alpha}'\mathbf{w}_i)M_i + M_i^2$$

$$\mathbf{\Delta} = diag[\delta_i].$$

The asymptotic covariance matrix for this MLE is estimable with the sample estimate of

$$\mathbf{Q}_p = [\mathbf{X}'\mathbf{\Lambda}\mathbf{X}]^{-1} + \theta[\mathbf{X}'\mathbf{\Lambda}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{V}^2\mathbf{\Delta}\mathbf{W}]\mathbf{\Sigma}[\mathbf{W}'\mathbf{V}^2\mathbf{\Delta}\mathbf{X}][\mathbf{X}'\mathbf{\Lambda}\mathbf{X}]^{-1}.$$

(The third and fourth terms in the expression reported in Greene (equation 3.6) vanish asymptotically and may thus be omitted.)

There is no direct connection drawn between the selection equation and the respecified conditional mean function in the preceding. As noted, the appendage of θM$_i$ to E[y$_i$|$\mathbf{x}_i$] is strictly ad hoc. Terza (1995) begins with

$$\ln E[y_i|\mathbf{x}_i, u_i] = \boldsymbol{\beta}'\mathbf{x}_i + u_i.$$

With joint normality of u$_i$ and ϵ_i, it follows that

$$E[y_i|\mathbf{x}_i, z_i = 1] = E[y_i|\mathbf{x}_i, u_i > -\boldsymbol{\alpha}'\mathbf{w}_i]$$

$$= e^{\boldsymbol{\beta}'\mathbf{x}_i + \sigma^2/2}\left[\frac{\Phi(\theta+\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(\boldsymbol{\alpha}'\mathbf{w}_i)}\right]$$

$$= e^{\boldsymbol{\beta}^{*\prime}\mathbf{x}_i}[\Psi(\theta, \tau_i)]$$

$$= \lambda_i\Psi_i,$$

where $\tau_i = \boldsymbol{\alpha}'\mathbf{w}_i$.

(A full derivation of this result is given in Terza (1995).) The last result provides a nonlinear conditional mean function. This gives the functional form for nonlinear least squares estimation of $(\boldsymbol{\beta}^*, \theta)$, where $\boldsymbol{\beta}^*$ equals $\boldsymbol{\beta}$ save for the constant term, which is offset by $\sigma^2/2$. To construct the expression for the asymptotic covariance matrix , let

$$\boldsymbol{\Lambda} = diag(\lambda_i)$$

$$v_i = y_i - E[y_i|\mathbf{x}_i, z_i = 1]$$

$$\mathbf{V} = diag(v_i)$$

$$p_i = \lambda_i \tfrac{\partial \Psi_i}{\partial \tau_i}$$

$$\mathbf{P} = diag(p_i)$$

Then, we use

$$\mathbf{Q}_T = [\mathbf{X}'\boldsymbol{\Lambda}^2\mathbf{X}]^{-1}\{\mathbf{X}'\boldsymbol{\Lambda}^2\mathbf{V}^2\mathbf{X} + (\mathbf{X}'\boldsymbol{\Lambda}^2\mathbf{P}\mathbf{W})\boldsymbol{\Sigma}(\mathbf{W}'\mathbf{P}\boldsymbol{\Lambda}^2\mathbf{X})\}[\mathbf{X}'\boldsymbol{\Lambda}^2\mathbf{X}]^{-1}$$

It is tempting at this point to proceed to full information maximum likelihood estimation of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \theta)$ based on the Poisson distribution. But, y_i is not distributed as Poisson in the selected propulation if it is in the full population. The conditional (on the selection) distribution in Terza's model remains to be derived. (It is assumed a priori in Greene's.)

With the exception of the constant term, the parameters in Terza's and Greene's model are the same. Consider an alternative formulation of Terza's result:

$$\ln E[y|\mathbf{x}_i, z_i = 1] = \boldsymbol{\beta}^{*\prime}\mathbf{x} + \ln \Psi(\theta + \tau_i).$$

Now, expand this expression in a linear Taylor series around $\theta = 0$. The central result is

$$\ln \Psi(\theta, \tau_i) \simeq \ln \Psi(0, \tau_i) + \theta \tfrac{\partial \ln \Psi(\theta, \tau_i)}{\partial \theta}_{|\theta=0}$$

$$= 0 + \theta \tfrac{\partial \ln \Phi(\theta + \tau_i)}{\partial \theta}_{|\theta=0}$$

$$= \theta \tfrac{\phi(\theta+\tau_i)}{\Phi(\theta+\tau_i)}_{|\theta=0}$$

$$= \theta M_i.$$

Thus, the expansion produces the conditional mean function specified in Greene (1994). (The constant term must be reinterpreted.) As such, the formulation in Greene can be interpreted as the linear approximation to what is obtained by a formal derivation of the conditional mean function under selection. The parameters can be estimated, as before, by nonlinear least squares. The estimator of the asymptotic covariance matrix for $(\boldsymbol{\beta}^*, \theta)$ is the sample estimate of

$$\mathbf{Q}_G = [\mathbf{X}'\mathbf{\Lambda}^2\mathbf{X}]^{-1}\{\mathbf{X}'\mathbf{\Lambda}^2\mathbf{V}^2\mathbf{X} + \theta(\mathbf{X}'\mathbf{\Lambda}^2\mathbf{\Delta}\mathbf{W})\mathbf{\Sigma}(\mathbf{W}'\mathbf{\Lambda}^2\mathbf{\Delta}\mathbf{X})\}[\mathbf{X}'\mathbf{\Lambda}^2\mathbf{X}]^{-1},$$

where the individual parts were defined earlier.

3. Application to Credit Scoring

Greene (1994) presents an application of the Poisson model with selectivity. The illustration involves an aspect of consumer credit behavior, the number of major derogatory reports (MDRs) in the recent credit history of a sample of credit card applicants. (An MDR is defined as a reported delinquency of sixty days or more on a credit account.) For most people and most accounts, the number is zero. But, as shown in Table 1,

INSERT TABLE 1 HERE

there is a large amount of variation across applicants and, in this sample, a large number of nonzero entries.[1] We have data on applications for a major credit card and whether the application was granted or not. The figures in Table 1 show clearly that the distribution of MDRs differs substantially between those accepted and those rejected. (This is to be expected, as this is one of the major criteria for acceptance.) Table 2 lists descriptive statistics for the covariates in the Poisson regression for MDRs and probit model for cardholder acceptance.

INSERT TABLE 2 HERE

Table 3 gives estimates for the probit model and the Poisson regression model computed by the three methods outlined earlier. All three estimators

INSERT TABLE 3 HERE

[1] The data used here are a random 10 percent sample from the data used in Greene (1992). We have used 1319 observations in total. The selected sample of accepted applications is 1023 observations. The data are available from the author upon request.

produce quite similar results. Greene's estimator yields a slightly smaller sum of squared residuals. Although the estimated slopes are similar, the two nonlinear least squares estimates of θ are quite different, and the extremely large standard error for the one computed by Terza's method is surprising.The estimated models are actually a bit closer even than suggested by the similar slope estimates. Table 4 lists the estimated marginal effects, $\partial E[y_i|\mathbf{x}_i, z_i = 1]/\partial \mathbf{x}_i$ for the four sets of estimates. For the four cases,

$$\begin{aligned}
\frac{\partial E[y_i|x_i, z_i = 1]}{\partial x_{ik}} &= E[y_i|x_i, z_i = 1] \left\{\beta_k + a\alpha_k\right\} \\
\alpha_k &= 0 \text{ if } x_{ik} \text{ does not appear in } w_i, \\
a &= 1 \text{ for the uncorrected model,} \\
a &= \theta\delta_i \text{ in Greene's formulations,} \\
a &= M_{\theta i} - M_i \text{ in Terza's formulation.}
\end{aligned}$$

In the last of these, $M_{\theta i}$ is M_i evaluated at $\theta + \tau_i$. (A linear Taylor series approximation to this function around $\theta = 0$ produces $\theta\delta_i$ as might be expected.) The marginal effects for the four variables in the regression model are given in Table 4. As noted, these are quite similar for the two models estimated by nonlinear least squares.

INSERT TABLE 4 HERE

4. Conclusions

The preceding has shown two methods of accommodating sample selection in the Poisson regression model. (Extensions to the negative binomial regression model and explicit treatment of heteroskedasticity are examined in Terza's study.) The parameter estimates are similar, and neither is appreciably more difficult to estimate than the other. Terza's approach produces a slightly poorer fit to the data, but this appears not to be systematic. The very large standard error produced his aproach compared to Greene's is surprising, however.

The literature on sample selection in this context is relatively thin, and there is only limited existing theory to draw on. There is a loose end in the preceding, as well as in Terza's formulation. In the original design of the sample selection models, the phenomenon under study is that in the selected population, the observed response variable will tend to be above or below the conditional mean (depending on the sign of the correlation between the disturbances) in the selected population. Here, the selectivity has been modelled, instead, as a form of heterogeneity which shifts the conditional mean function in the selected population. Thus, it is not quite the same phenomenon as Heckman addressed in his early work on the subject. We leave pursuit of that issue for continuing work on this subject.

5. References

Greene, W., "Sample Selection as a Specification Error: Comment," *Econometrica*, 49, 1981, pp. 795-798.

Greene, W.,"A Statistical Model for Credit Scoring," Working Paper number 92-10, Department of Economics, Stern School of Business, New York University, 1992.

Greene, W., "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models," Working Paper number 94-20, Department of Economics, Stern School of Business, New York University, 1994.

Heckman, J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 1976, p. 475-492.

Heckman, J., "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 1979, p. 153-161.

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

Terza, J., "Estimating Count Data Models with Endogeous Switching and Sample Selection," Working Paper, Department of Economics, Pennsylvania State University, State College, PA, March, 1995.

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										

Table 2. Descriptive statistics for independent variables.

Variable		Mean	Standard deviation	Minimum	Maximum
Income	Income, in 10,000s				
Full sample		3.365	1.694	0.21	13.50
Cardholders		3.451	1.707	0.21	13.50
Age	Age, in years				
Full sample		33.21	10.14	17.00	83.50
Cardholders		33.22	10.22	17.00	83.50
Cur._Add.	Number of months residing at current address				
Full sample		55.268	66.272	0.000	540.000
Cardholders		55.258	64.710	0.000	540.000
Exp._Inc.	Average monthly expenditure divided by yearly income				
Full sample		0.0687	0.0946	0.0001	0.9063
Cardholders		0.0885	0.0991	0.0001	0.9063
Avg._Exp.	Average monthly credit card expenditure				
Full sample		185.06	272.22	0.00	3100.0
Cardholders		238.61	287.71	0.00	3100.0
Own_Rent	Binary variable indicating home ownership				
Full sample		0.440	0.497	0.00	1.00
Cardholders		0.062	0.500	0.00	1.00
Self_Empl.	Binary variable, one for self employed				
Full sample		0.070	0.253	0.00	1.00
Cardholders		0.062	0.241	0.00	1.00
Depndt.	Number of dependents, not including the individual				
Full sample		0.994	1.248	0.00	6.00
Cardholders		0.970	1.243	0.00	6.00
Inc._Per	Income per dependent				
Full sample		2.156	1.363	0.070	11.00
Cardholders		2/219	1.351	0.070	11.00
Major	Binary variable for whether the individual holds a major credit card				
Full sample		0.817	0.387	0.00	1.00
Cardholders		0.840	0.367	0.00	1.00
Open	Number of open, current trade accounts				
Full sample		6.360	6.053	0.00	37.00
Cardholders		7.049	6.026	0.00	30.00
Active	Number of active credit card accounts				
Full sample		6.997	6.306	0.000	46.000
Cardholders		7.049	6.026	0.000	30.000

Table 3. Estimated Parameters. Estimated standard errors are in parentheses.

	Cardholder	Number of Major Derogatory Reports			
			Selection Corrected		
	Probit	Poisson	Poisson	Greene	Terza
Constant	0.542 (0.184)	-3.616 (0.422)	-4.594 (0.521)	-5.345 (0.740)	-4.068 (0.596)
Age	-0.00857 (0.00498)	0.0188 (0.00872)	0.0162 (0.00996)	0.0128 (0.0110)	0.0142 (0.0106)
Income	0.0920 (0.0532)	0.134 (0.0543)	0.183 (0.0613)	0.191 (0.0596)	0.136 (0.0586)
Exp._Inc.		1.986 (1.265)	1.878 (1.296)	1.775 (0.943)	1.734 (1.075)
Avg._Exp.		0.0000483 (0.000395)	-0.0000236 (0.000419)	-0.0000268 (0.000308)	-0.0000362 (0.000405)
Major	0.212 (0.103)	0.242 (0.268)	0.572 (0.316)	1.376 (0.590)	0.811 (0.491)
Mills Ratio			1.788 (0.431)	1.989 (0.296)	3.465 (30.689)
Sum of Squared Deviations				165.319	168.262
Own_Rent	0.349 (0.101)				
Depndt.	-0.131 (0.069)				
Inc._Per	-0.0150 (0.0714)				
Self_Empl.	-0.201 (0.163)				
Open	-0.286 (0.0245)				
Cur._Add.	-0.000409 (0.000700)				
Active	-0.230 (0.0214)				
Log-Likelihood		-407.944	-394.157		

Table 4. Estimated Marginal Effects

Variable	Uncorrected Poisson	Corrected Poisson MLE	Greene Nonlinear LS	Terza Nonlinear LS