A Two-Step First Difference Estimator for a Panel Data Tobit Model under Conditional Mean Independence Assumptions

Kalwij, A.S.

Publication date:
2004

CentER Discussion Paper

No. 2004–67

A TWO-STEP FIRST DIFFERENCE ESTIMATOR FOR A PANEL DATA TOBIT MODEL UNDER CONDITIONAL MEAN INDEPENDENCE ASSUMPTIONS

By A.S. Kalwij

July 2004

ISSN 0924-7815

Tilburg University

A TWO-STEP FIRST-DIFFERENCE ESTIMATOR FOR A PANEL DATA TOBIT MODEL UNDER CONDITIONAL MEAN INDEPENDENCE ASSUMPTIONS

Adriaan S. Kalwij*

Department of Economics, Tilburg University, Tilburg, The Netherlands

23 July 2004

ABSTRACT

This study develops a two-step estimator for a panel data Tobit model based on taking first-differences of the equation of interest, under conditional mean independence assumptions. The necessary correction terms are non-standard and a substantial part is therefore devoted to the formal derivation of these correction terms. The main advantage of this estimator is that it yields estimates that are far less sensitivity to misspecification of the conditional mean independence assumption than an estimation procedure set up in levels. Monte Carlo simulations are provided in support of this.

Keywords: Two-Step Estimator, Panel data, Tobit model

JEL-classification: C12, C23, C24

* *Address for correspondence*: Room B933, Department of Economics, Tilburg University,

PO Box 90153, 5000 LE Tilburg, The Netherlands. (Email: kalwij@uvt.nl)

1. INTRODUCTION

The main contribution to the literature of this study is the derivation of a two-step panel data Tobit model based on taking first-differences of the equation of interest, under conditional mean independence assumptions (Wooldridge, 1995). The main advantage of this new estimator is that it yields estimates that are far less sensitivity to misspecification of the conditional mean independence assumption than an estimation procedure set up in levels, as proposed by Wooldrigde (1995, section 4).

The literature on nonlinear panel data models, surveyed by Honoré (2002), suggests two approaches to estimating a censored regression model in which the individual specific effects are allowed to correlate with the explanatory variables. The first is a fixed effects approach (Honoré, 1992, Kyriazidou, 1997) for which consistency does not require any assumptions on the individual specific effects but does require an additional distributional assumption, for instance the stationarity assumption in Honoré (1992). The second is a random effects approach (Nijman and Verbeek, 1992, Wooldridge, 1995, Rochina-Barrachina, 1999, Kalwij, 2003) for which consistency does not require an additional distributional assumption but does require a correct specification of the correlation between the individual specific effects and the explanatory variables, the so-called conditional mean independence assumption (Wooldrigde, 1995). This latter approach is followed in this study. The main reason for preferring this random effects approach is that, unlike the fixed effects approach, it yields a fully specified model that can be used for calculating marginal effects. Consistency of the estimates is based on the assumption of correctly specified correlation between the individual specific effects and the explanatory variables and, as a consequence, parameter estimates turn out to be sensitivity to the chosen parameterization. For this reason one may wish to use a very flexible parameterization (Zabel, 1992, Wooldridge, 1995). This can, however, increase the number of parameters dramatically. An alternative solution proposed in Kalwij (2003) is to start by eliminating the individual specific effects from the equation of interest by taking first-differences. He develops a Maximum Likelihood estimator based on this idea. The disadvantage of his estimator is that it is difficult to use in practice for applied researchers since it demands programming of non-standard likelihood contributions. For this reason this study builds on the work of Kalwij (2003) and constructs a relatively easy-to-use two-step estimator for a panel data Tobit model based on taking first-differences of the equation of interest. Another reason for preferring a two-step procedure is to reduce the sensitivity of the parameter estimates with respect to the distribution assumptions. The procedure for obtaining the necessary correction terms is conceptually similar to the derivation of the inverse Mill's ratio in case of a standard Tobit model in levels. The correction terms for a Tobit model based on taking first-differences are, however, non-trivial to derive and for this purpose a large part of this study is devoted to this derivation. Section 2 derives the Moment Generating Function of a random variable that is defined as the difference of two censored normal random variables and derives the expectation of this random variable. Based on this result, section 3 constructs the correction terms and formulates the two-step first-difference estimator for a panel data Tobit model under conditional mean independence assumptions. Section 4 carries out simulations to examine how this estimator behaves, relatively to the two-step Tobit estimator of Wooldridge (1995), when the conditional mean independence assumption does not hold. Section 5 concludes.

2. THE FIRST MOMENT OF THE DIFFERENCE OF TWO CENSORED NORMAL RANDOM VARIABLES

X and *Y* are two jointly distributed normal random variables:

(1) $$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

The density function of *(X,Y)* is the bivariate normal density function and is denoted by $f_{XY}(x,y)$. In the remainder of this study $\phi(.)$ denotes the standard normal density function, $\Phi(.)$ the standard normal distribution function and $\Phi_2(.,.;\rho)$ the bivariate standard normal distribution function with a correlation coefficient ρ ($\equiv \rho_{XY}$, subscript *XY* is dropped).

X is censored for values below *a* and *Y* is censored for values below *b*. The density function of the random variable $Z = X - Y$ conditional on *X*>*a* and *Y*>*b* is given by (see Kalwij, 2003):

(2) $$f_{Z|X>a,Y>b}(z \mid X>a, Y>b) = \frac{\frac{1}{\sigma_Z}\phi\left(\frac{z-\mu_Z}{\sigma_Z}\right)\Phi\left(-\frac{\left(\max\{a, b+z\}-\mu_{X|Z}\right)}{\sigma_{X|Z}}\right)}{\Phi_2\left(\left(\frac{-a-\mu_X}{\sigma_X}\right),\left(\frac{-b-\mu_Y}{\sigma_Y}\right);\rho\right)},$$

where $\mu_Z = \mu_X - \mu_Y$, $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$, $\rho_{XZ} = \frac{\sigma_X - \rho\sigma_Y}{\sigma_Z}$,

$\mu_{X|Z} = \mu_X + \rho_{XZ}\frac{\sigma_X}{\sigma_Z}(z-\mu_Z)$, and $\sigma_{X|Z}^2 = \left(1-\rho_{XZ}^2\right)\sigma_X^2$.

The Moment Generating Function of the random variable $Z = X - Y$ conditional on *X*>*a* and *Y*>*b* is given by (see Appendix A):

(3) $$e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2}\frac{\int_{-\infty}^{+\infty}\phi(s)\Phi\left(-\left(\frac{\max\{a,(b+\mu_Z+\sigma_Z(s+t\sigma_Z))\}-\mu_X}{\sigma_X\sqrt{1-\rho_{XZ}^2}} - \frac{\rho_{XZ}(s+t\sigma_Z)}{\sqrt{1-\rho_{XZ}^2}}\right)\right)\partial s}{\Phi_2\left(\left(\frac{-a-\mu_X}{\sigma_X}\right),\left(\frac{-b-\mu_Y}{\sigma_Y}\right);\rho\right)}.$$

The expectation of $Z = X - Y$ conditional on *X*>*a* and *Y*>*b* is obtained by taking the derivative of the Moment Generating Function, i.e. equation (3), with respect to *t* and evaluate this function in *t* equals 0. This yields (see Appendix B):

(4) $$\mu_{Z|X>a,Y>b} = \left[\Phi_2\left(\left(\frac{-a-\mu_X}{\sigma_X}\right),\left(\frac{-b-\mu_Y}{\sigma_Y}\right);\rho\right)\right]^{-1} \times \left[\mu_Z \int_{-\infty}^{\frac{a-b-\mu_Z}{\sigma_Z}} \phi(s)\Phi\left(-\frac{\frac{a-\mu_X}{\sigma_X} - \rho_{XZ}s}{\sqrt{1-\rho_{XZ}^2}}\right)\partial s\right.$$

$$+ \quad \mu_Z \int_{\frac{a-b-\mu_Z}{\sigma_Z}}^{+\infty} \phi(s)\Phi\left(-\frac{\frac{b-\mu_Y+\sigma_Z s}{\sigma_X}-\rho_{XZ}s}{\sqrt{1-\rho_{XZ}^2}}\right)\partial s$$

$$+ \quad (\sigma_X - \rho\sigma_Y)\phi\left(\frac{a-\mu_X}{\sigma_X}\right)\Phi\left(\frac{\rho\left(\frac{a-\mu_X}{\sigma_X}\right)-\left(\frac{b-\mu_Y}{\sigma_Y}\right)}{\sqrt{1-\rho^2}}\right)$$

$$+ \quad (\rho\sigma_X - \sigma_Y)\phi\left(\frac{b-\mu_Y}{\sigma_Y}\right)\Phi\left(\frac{\rho\left(\frac{b-\mu_Y}{\sigma_Y}\right)-\left(\frac{a-\mu_X}{\sigma_X}\right)}{\sqrt{1-\rho^2}}\right)\Bigg].$$

3. A TWO-STEP FIRST-DIFFERENCE ESTIMATOR FOR A PANEL DATA TOBIT MODEL

The model of interest is formulated as follows, assuming *T* time periods and *N* individuals:

(5) $$y_{it}^* = x_{it}\beta + \alpha_i + \varepsilon_{it} ,$$

$$y_{it} = \max(0, y_{it}^*) \qquad t = \{1,..,T\}, i = \{1,..,N\}.$$

This is a panel data Tobit model and is a straightforward extension of the standard Tobit model (Tobin, 1958) by including an additional dimension (time or individuals). The individual is indexed by i and the time period by t, x_{it} is a (1xK) vector of exogenous variables, β is a (Kx1) vector of the parameters of interest and α_i is an unobserved individual specific effect that may correlate with x_{it}. The latent dependent variable y_{it}^* is censored at zero and only y_{it} is observed. The error term ε_{it} is assumed to be a normal random variable with mean zero and variance σ_t^2, $\varepsilon_{it} \sim N(0,\sigma_t^2)$, and is allowed to be serially correlated. The panel data are characterized by having a large number of individuals over a short period of time.

The two-step estimator proposed by Wooldridge (1995) is as follows. Based on estimates of a probit model explaining whether or not the dependent variable is positive, given x_{it}, one constructs the familiar inverse Mills ratios for each time period. Next, one estimates equation (5) on a sub-sample of observations for which the dependent variable is positive with the period-specific inverse Mills ratios as additional explanatory variables. A conditional mean independence assumption is made to deal with the possible correlations between the unobserved individual specific effects and the explanatory variables. This procedure allows for arbitrary serial correlation.

The main idea behind the Maximum Likelihood estimator of Kalwij (2003) is that one starts by eliminating the unobserved individual specific effects from equation (5) in order to make the resulting estimates less sensitivity to a conditional mean independence assumption. First-differencing equation (5) yields the following model:

$$\Delta y_{it}^* = \Delta x_{it}\beta + \eta_{it} \tag{6}$$

$$\Delta y_{it} = \begin{cases} \Delta y_{it}^* & if \quad y_{is}^* > 0 \quad and \quad y_{it}^* > 0 \\ 0 & if \quad y_{is}^* \le 0 \quad and \quad y_{it}^* \le 0 \\ y_{it}^* & if \quad y_{is}^* \le 0 \quad and \quad y_{it}^* > 0 \\ -y_{is}^* & if \quad y_{is}^* > 0 \quad and \quad y_{it}^* \le 0 \end{cases} \tag{7}$$

Where $\Delta y_i^* = y_{it}^* - y_{is}^*$, $\Delta x_{it} = x_{it} - x_{is}$ and $\eta_{it} = \varepsilon_{it} - \varepsilon_{is}$. The correlation coefficient of ε_{is} and ε_{it} is denoted by ρ_t. Given the distributional assumptions $\eta_{it} \sim N(0,\sigma_{\eta t}^2)$ with $\sigma_{\eta t}^2 = \sigma_s^2 - 2\rho_t\sigma_s\sigma_t + \sigma_t^2$, one can obtain Maximum Likelihood estimates of the parameters of interest using the density function as formulated in equation (2) (see Kalwij, 2003). An alternative is to construct a two-step estimator in the spirit of Rochina-Barrachina (1999) who develops a two-step estimator for panel data selection models. For the case of a first-difference panel data Tobit model (equations (6) and (7)) a two-step procedure is developed underneath, using the fundamental statistical results of Section 2.

Considering the model given by equation (6) and the nonzero observations on the dependent variable in both periods, the expectation of Δy_{it} conditional on $y_{is} > 0$ and $y_{it} > 0$ is given by:

(8) $$E[\Delta y_{it} \mid y_{is} > 0, y_{it} > 0] = \Delta x_{it}\beta + E[\eta_{it} \mid \varepsilon_{is} > -x_{is}\beta - \alpha_i, \varepsilon_{it} > -x_{it}\beta - \alpha_i].$$

Using equation (4), the expectation at the RHS of equation (8) is written as follows:

(9) $$E[\eta_{it} \mid \varepsilon_{is} > -x_{is}\beta - \alpha_i, \varepsilon_{it} > -x_{it}\beta - \alpha_i] = \pi_t \Lambda_t(M_{is}, M_{it}, \rho_t) - \pi_s \Lambda_s(M_{is}, M_{it}, \rho_t).$$

With

(10) $$\pi_s = (\sigma_s - \rho_t \sigma_t),$$

(11) $$\pi_t = (\sigma_t - \rho_t \sigma_s),$$

(12) $$M_{is} = (x_{is}\beta + \alpha_i)/\sigma_s,$$

(13) $$M_{it} = (x_{it}\beta + \alpha_i)/\sigma_t,$$

(14) $$\Lambda_s(M_{is}, M_{it}, \rho) = \frac{\phi(M_{is})\Phi\left((M_{it} - \rho M_{is})\big/\sqrt{1-\rho_t^2}\right)}{\Phi_2(M_{is}, M_{it}; \rho_t)},$$

(15) $$\Lambda_t(M_{is}, M_{it}, \rho) = \frac{\phi(M_{it})\Phi\left((M_{is} - \rho M_{it})\big/\sqrt{1-\rho_t^2}\right)}{\Phi_2(M_{is}, M_{it}; \rho_t)}.$$

The correction terms are now added and subtracted from equation (6):

(16) $$\Delta y_{it} = \Delta x_i \beta + \pi_t \Lambda_t(M_{is}, M_{it}, \rho_t) - \pi_s \Lambda_s(M_{is}, M_{it}, \rho_t) + \xi_{it} \qquad \text{if } y_{is} > 0 \text{ and } y_{it} > 0,$$

with

(17) $$\xi_{it} = -\pi_t \Lambda_t(M_{is}, M_{it}, \rho_t) + \pi_s \Lambda_s(M_{is}, M_{it}, \rho_t) + \eta_{it}.$$

ξ_{it} is an error term with expectation zero (conditional on $y_{is} > 0$ and $y_{it} > 0$). If the correction terms are known then equation (16) can be estimated using least squares on the sample of individuals for which the dependent variable is positive in both periods.

3.1 A two-step estimation procedure

A two-step estimation procedure for estimating equation (16) is constructed as follows. In the first step we estimate the following bivariate probit model for every time-pair, say two subsequent periods *s* and *t*:

(18) $$y_{is}^{*} = x_{is}\beta + \alpha_i + \varepsilon_{is},$$

(19) $$y_{it}^{*} = x_{it}\beta + \alpha_i + \varepsilon_{it},$$

(20) $$I_{is} = \begin{cases} 0 & if \quad y_{is}^{*} \leq 0 \\ 1 & if \quad y_{is}^{*} > 0 \end{cases},$$

(21) $$I_{it} = \begin{cases} 0 & if \quad y_{it}^{*} \leq 0 \\ 1 & if \quad y_{it}^{*} > 0 \end{cases}.$$

We use the conditional mean independence assumptions as discussed in Wooldridge (1995) to deal with correlations between the unobserved individual specific effects and the explanatory variables. This approach essentially models the correlation between the unobserved individual specific effect and the explanatory variables as a pre-specified function of the explanatory variables and a random individual specific error term, for instance:

(22) $$\alpha_i = h(x_i)\gamma + \mu_i,$$

with $\mu_i \mid x_i \sim N(0, \sigma_\mu^2)$ and $x_i \equiv (x_{i1},..,x_{iT})$. The parameter γ is an additional parameter of the model to be estimated. A popular choice for $h(x_i)$ in empirical research is to take the average over time. Substituting the parameterisation of the individual specific effects in equations (18) and (19) yields:

(23) $$y_{is}^{*} = x_{is}\beta + h(x_i)\gamma + u_{is},$$

(24) $$y_{it}^{*} = x_{it}\beta + h(x_i)\gamma + u_{it},$$

with $u_{is} = \mu_i + \varepsilon_{is}$ and $u_{it} = \mu_i + \varepsilon_{it}$, and variances $\sigma_{u,s}^2 = \sigma_\mu^2 + \sigma_s^2$ and $\sigma_{u,t}^2 = \sigma_\mu^2 + \sigma_t^2$, respectively. The correlation between u_{is} and u_{it} is denoted by ρ_{st}. Using Maximum Likelihood for estimating the bivariate probit model yields estimates for $\beta_s = \dfrac{\beta}{\sigma_{u,s}}, \gamma_s = \dfrac{\gamma}{\sigma_{u,s}}, \beta_t = \dfrac{\beta}{\sigma_{u,t}}, \gamma_t = \dfrac{\gamma}{\sigma_{u,t}}$ and ρ_{st}.

In the second step we substitute $\hat{M}_{is} = x_{is}\hat{\beta}_s + h(x_i)\hat{\gamma}_s$, $\hat{M}_{it} = x_{it}\hat{\beta}_t + h(x_i)\hat{\gamma}_t$ and $\hat{\rho}_{st}$ in equation (16):

(25) $\Delta y_{it} = \Delta x_{it}\beta + \pi_t^* \Lambda_t \left(\hat{M}_{is}, \hat{M}_{it}, \hat{\rho}_{st}\right) - \pi_s^* \Lambda_1 \left(\hat{M}_{is}, \hat{M}_{it}, \hat{\rho}_{st}\right) + \xi_{it}$ if $y_{is} > 0$ and $y_{it} > 0$.

This equation is estimated using Least Squares on a sample of individuals for which the dependent variable is positive in both periods ($y_{is} > 0$ and $y_{it} > 0$). The auxiliary parameters π_s^* and π_t^* are defined as $\pi_s^* = \left(\sigma_{u,s} - \rho_{st}\sigma_{u,t}\right)$ and $\pi_t^* = \left(\sigma_{u,t} - \rho_{st}\sigma_{u,s}\right)$.

4. A MONTE CARLO STUDY

The main idea behind constructing a new two-step estimator for a panel data Tobit model based on taking first-differences of the equation of interest is to reduce the bias that may arise if the conditional mean independence assumption does not hold, i.e. a misspecification of the individual specific effects (equation (22)). To examine this a Monte Carlo study is carried out to compare the two-step first-difference Tobit estimator of Section 3 with the 'standard' two-step panel data Tobit estimator as proposed by Wooldrigde (1995), both under conditional mean independence assumptions.

The design of the Monte Carlo simulations is as follows. The data is generated by: $y_{it} = \max(0, 0.2 + x_{it} + \alpha_i + \varepsilon_{it})$, $x_{it} = 0.8x_{it-1} + \xi_{it}$, $\varepsilon_{it} = 0.4\varepsilon_{it-1} + \zeta_{it}$, $x_{i1} = \xi_{i1}$, $\varepsilon_{i1} = \zeta_{i1}$ and the three error terms $\varepsilon_{it}, \xi_{it}$ and ζ_{it} are $N(0,1)$ distributed. So the true value of β is 1.The individual specific effects are generated using a nonlinear function of the time-averages of the explanatory variables: $\alpha_i = \bar{x}_i \,|\, \bar{x}_i \,| + \mu_i$, $\mu_i \sim N(0,1)$. The simulations are based on 100 replications and the values chosen for

N and T are, respectively, {500, 1000} and {2, 4, 8, 12}. These are plausible values in empirical research using panel data. We have experimented with several different designs and the main conclusion of this study remained unchanged. Nevertheless, the results below have to be interpreted with caution since they may depend on the particular design chosen.

We estimate β using three different models and using the two-step estimators outlined in Section 3 and in Wooldridge (1995). The three models differ with respect to the conditional mean independence assumption chosen, i.e. the choice of *h(.)* in equation (22). The first model is a correctly specified model using $h(x_i) = \bar{x}_i \mid \bar{x}_i \mid$. The second model is a random effects panel data Tobit model assuming no correlation between the individual specific effect and the explanatory variable: $h(x_i) = 0$. The third model uses an in empirical research often made conditional mean independence assumption that the correlation between the unobserved individual specific effect and the explanatory variable is a linear function of the averages over time of the explanatory variable: $h(x_i) = \bar{x}_i$. Hence, the second and third models are misspecified. When estimated in levels the second model is commonly estimated when using a standard statistical package such as STATA (*xttobit* command).

The simulation results are reported in table 1. The simulation results of model 1 show that both estimators yield, of course, unbiased estimates of the parameter of interest (β=1). The root mean squared error (RMSE) shows that the first-difference Tobit estimator is often less efficient but the differences with Wooldridge's Tobit estimator are relatively small. Nevertheless, using the first-difference Tobit estimator instead of Wooldridge's Tobit estimator is likely to result in a loss of efficiency. The simulation results when using models 2 and 3 provide some measure of the relative performance of the two estimators when the conditional mean independence assumption, as specified in equation (22), does not hold. When estimating model 2, both estimators yield biased estimates but the magnitude of the biases of the two estimators differ considerably. The estimator based on first-differences yields an estimate that is far less sensitive to misspecification of the individual specific effects, relatively to Wooldridge's Tobit estimator. Finally we estimate the model with a commonly chosen conditional mean independence assumption, i.e.

model 3. Here we see substantial bias reductions for both estimators but the relative bias when using Wooldridge's estimator remains high.

We experimented with alternative simulation designs but they all yielded the same conclusion: using the two-step first-difference Tobit estimator substantially reduces the sensitivity of the estimates of the parameters of interest with respect to the conditional mean independence assumption.

5. CONCLUSIONS

This study developed a new two-step first-difference estimator for a panel data Tobit model under conditional mean independence assumptions. In particular, since the necessary correction terms are non-standard, a substantial part of this study has been devoted to the formal derivation of these correction terms. This new estimator yields estimates that are far less sensitivity to a misspecification of the conditional mean independence assumption than an estimation procedure set up in levels. The results of the Monte Carlo study in Section 4 support this.

We conclude that the two-step first-difference Tobit estimator proposed in this study can be regarded as a very attractive alternative to the fixed effects Tobit estimator of Honoré (1992), since it allows one to calculate marginal effects, and as a bias reduction technique when applied to a random effects panel data Tobit model under conditional mean independence assumptions (Wooldridge, 1995). We successfully applied this new two-step first-difference Tobit estimator in Kalwij and Gregory (2005) to an analysis of paid overtime work in Britain, hereby comparing the empirical results in detail with alternative approaches commonly used in the literature.

ACKNOWLEDGEMENTS

I would like to thank Mary Gregory, Maria Rochina-Barrachina, Jeff Wooldridge and seminar participants at the 9th International Conference on Panel Data for very helpful comments.

REFERENCES

Honoré, B.E., 1992, Trimmed LAD and Least Squares estimation of truncated and censored regression models with fixed effects, Econometrica, 60, 533-565

Honoré, B.E., 2002, Nonlinear models with panel data, Portuguese Economic Journal, 1 (2), 163-179.

Kalwij, A.S., 2003, A Maximum Likelihood estimator based on first differences for a panel data Tobit model with individual specific effects, Economics Letters, Volume 81, Issue 2, 19-26.

Kalwij, A.S., and M. Gregory, 2005, A panel data analysis of the effects of wages, standard hours and unionisation on paid overtime work in Britain, Journal of the Royal Statistical Society, Series A, 168, forthcoming.

Kyriazidou, E., 1997, Estimation of a panel data sample selection model, Econometrica, 65 (6), 1335-1364.

Nijman, T. and M. Verbeek, 1992, Nonresponse in panel data: the impact on estimates of a life-cycle consumption function, Journal of Applied Econometrics, 7, 243-257.

Tobin, J., 1958, Estimation of relationships for limited dependent variables, Econometrica, 26, 24-36.

Rochina-Barrachina, M.E., 1999, A new estimator for panel data sample selection models, Annales d'Économie et de Statistique, 55/56, 153-181.

Wooldridge, J.M., 1995, Selection corrections for panel data models under conditional mean independence assumptions, Journal of Econometrics, 68, 115-132.

Zabel, J., 1992, Estimating fixed and random effects with selectivity, Economics Letters, 40, 269-272.

APPENDIX A: THE MOMENT GENERATING FUNCTION

The Moment Generating Function of the random variable Z given $X>a$ and $Y>b$ is defined as:

(A1) $$m(t) = \int_{-\infty}^{+\infty} e^{tz} \frac{\int_{\max\{a,b+z\}}^{\infty} f_{XZ}(x,z)\partial x}{\int_a^{\infty}\int_b^{\infty} f_{XY}(x,y)\partial y \partial x} \partial z$$

(A2) $$m(t) = \frac{\int_{-\infty}^{+\infty} e^{tz}\left[\int_{\max\{a,b+z\}}^{\infty} f_{XZ}(x,z)\partial x\right]\partial z}{\int_a^{\infty}\int_b^{\infty} f_{XY}(x,y)\partial y \partial x}$$

The denominator of equation (A2) can be written as:

$$\int_a^{\infty}\int_b^{\infty} f_{XY}(x,y)\partial y \partial x = \Phi_2\left(\frac{-(a-\mu_X)}{\sigma_X}, \frac{-(b-\mu_Y)}{\sigma_Y}, \rho\right)$$

The numerator of equation (A2) can be written as:

(A3) $$\int_{-\infty}^{+\infty}\left[\int_{\max\{a,b+z\}}^{\infty} e^{tz} f_{XZ}(x,z)\partial x\right]\partial z$$

To simplify the integral we substitute

(A4) $$u = \frac{x-\mu_X}{\sigma_X} \text{ and } v = \frac{z-\mu_Z}{\sigma_Z}.$$

Furthermore, to adjust the borders of integration we define:

(A5) $$c_1(v) = \frac{a-\mu_X}{\sigma_X} \text{ and } c_2(v) = \frac{b+(\mu_Z+\sigma_Z v)-\mu_X}{\sigma_X}.$$

Now (A3) becomes:

(A6) $$\int_{-\infty}^{+\infty}\left[\int_{\max\{c_1(v),c_2(v)\}}^{\infty} e^{t(\mu_Z+\sigma_Z v)}\phi_2(u,v;\rho_{XZ})\partial u\right]\partial v,$$

where $\phi_2(.,.;.)$ denoted the bivariate standard normal distribution. Equation (A6) is rewritten as follows:

$$\text{(A7)} \quad e^{t\mu_Z} \int_{-\infty}^{+\infty} \left[\int_{\max\{c_1(v),c_2(v)\}}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_{XZ}^2}} \right.$$

$$\left. \times \exp\left\{ -\frac{1}{2(1-\rho_{XZ}^2)} \left[u^2 - 2\rho_{XZ}uv + v^2 - 2(1-\rho_{XZ}^2)t\sigma_Z v \right] \right\} \partial u \right] \partial v$$

To simplify this integral we substitute:

$$\text{(A8)} \quad w = \frac{u - \rho_{XZ} v}{\sqrt{1-\rho_{XZ}^2}} \qquad \text{and} \qquad s = v - t\sigma_Z .$$

The new borders of integration become:

$$\text{(A9)} \quad c_3(s,t) = \frac{\frac{a-\mu_X}{\sigma_X} - \rho_{XZ}(s + t\sigma_Z)}{\sqrt{1-\rho_{XZ}^2}} \quad \text{and}$$

$$c_4(s,t) = \frac{\frac{b + (\mu_Z + \sigma_Z(s + t\sigma_Z)) - \mu_X}{\sigma_X} - \rho_{XZ}(s + t\sigma_Z)}{\sqrt{1-\rho_{XZ}^2}}$$

Equation (A6) now becomes:

$$\text{(A10)} \quad e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{-\infty}^{+\infty} \phi(s) \left[\int_{\max\{c_3(s,t),c_4(s,t)\}}^{\infty} \phi(w)\partial w \right] \partial s$$

with

$$\text{(A11)} \quad \max\{c_3(s,t), c_4(s,t)\} = \begin{cases} c_3(s,t) & if \quad s < \dfrac{a - b - \mu_Z - t\sigma_Z^2}{\sigma_Z} \\ c_4(s,t) & if \quad s \geq \dfrac{a - b - \mu_Z - t\sigma_Z^2}{\sigma_Z} \end{cases}$$

APPENDIX B

$m_1(t)$ denotes the numerator of the Moment Generating Function of equation (3), i.e. equation (A10):

(B1) $$m_1(t) = e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{-\infty}^{+\infty} \phi(s) \left[\int_{\max\{c_3(s,t),c_4(s,t)\}}^{\infty} \phi(w) \partial w \right] \partial s \, .$$

The borders $c_3(s,t)$ and $c_4(s,t)$ are defined as in Appendix A (equation (A9)). First we need the derivative of $m_1(t)$ with respect to t:

(B2) $$\frac{\partial m_1(t)}{\partial t} = \left(\mu_Z + \sigma_Z^2 t\right) e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{-\infty}^{+\infty} \left[\int_{\max\{c_3(s,t),c_4(s,t)\}}^{\infty} \phi_2(w, s; 0) \partial w \right] \partial s$$

$$+ e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{-\infty}^{+\infty} \left[-\frac{\partial \max\{c_3(s,t), c_4(s,t)\}}{\partial t} \times \phi_2\left(\max\{c_3(s,t), c_4(s,t)\}, s; 0\right) \right] \partial s \, ,$$

(B3) $$\frac{\partial m_1(t)}{\partial t} =$$

$$\left(\mu_Z + \sigma_Z^2 t\right) e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{-\infty}^{c(t)} \left[\int_{c_3(s,t)}^{\infty} \frac{1}{2\pi} \times \exp\left\{ -\frac{1}{2}\left[w^2 + s^2\right] \right\} \partial w \right] \partial s$$

$$+ \left(\mu_Z + \sigma_Z^2 t\right) e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{c(t)}^{+\infty} \left[\int_{c_4(s,t)}^{\infty} \frac{1}{2\pi} \times \exp\left\{ -\frac{1}{2}\left[w^2 + s^2\right] \right\} \partial w \right] \partial s$$

$$+ e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{-\infty}^{c(t)} \left[-\frac{\partial c_3(s,t)}{\partial t} \frac{1}{2\pi} \times \exp\left\{ -\frac{1}{2}\left[\left(c_3(s,t)\right)^2 + s^2\right] \right\} \right] \partial s$$

$$+ e^{t\mu_Z + \frac{1}{2}\sigma_Z^2 t^2} \int_{c(t)}^{+\infty} \left[-\frac{\partial c_4(s,t)}{\partial t} \frac{1}{2\pi} \times \exp\left\{ -\frac{1}{2}\left[\left(c_4(s,t)\right)^2 + s^2\right] \right\} \right] \partial s$$

with

(B4) $$\frac{\partial c_3(s,t)}{\partial t} = \frac{-\rho_{XZ}\sigma_Z}{\sqrt{1-\rho_{XZ}^2}} \text{, and } \frac{\partial c_4(s,t)}{\partial t} = \frac{\frac{\sigma_Z^2}{\sigma_X} - \rho_{XZ}\sigma_Z}{\sqrt{1-\rho_{XZ}^2}} \, ,$$

(B5) $$\left(c_3(s,t)\right)^2 + s^2\Big|_{t=0} = \left(\frac{s - \frac{a-\mu_X}{\sigma_X}\rho_{XZ}}{\sqrt{1-\rho_{XZ}^2}}\right)^2 + \left(\frac{a-\mu_X}{\sigma_X}\right)^2,$$

and

(B6) $$\left(c_4(s,t)\right)^2 + s^2\Big|_{t=0} = \left(\frac{s - \frac{b-\mu_Y}{\sigma_Y}\times\frac{\rho_{XZ}\sigma_X - \sigma_Z}{\sigma_Y}}{\frac{\sigma_X}{\sigma_Y}\sqrt{1-\rho_{XZ}^2}}\right)^2 + \left(\frac{b-\mu_Y}{\sigma_Y}\right)^2.$$

Next we evaluate equation (B3) in t=0:

(B7) $$\left.\frac{\partial m_1(t)}{\partial t}\right|_{t=0} = (\mu_Z)\int_{-\infty}^{c(0)}\phi(s)\Phi(-c_3(s,0))\partial s + (\mu_Z)\int_{c(0)}^{+\infty}\phi(s)\Phi(-c_4(s,0))\partial s$$

$$+\rho_{XZ}\sigma_Z\phi\left(\frac{a-\mu_X}{\sigma_X}\right)\Phi\left(\frac{c(0)-\frac{a-\mu_X}{\sigma_X}\rho_{XY}}{\sqrt{1-\rho_{XZ}^2}}\right)$$

$$+\frac{\sigma_X}{\sigma_Y}\left(\rho_{XZ}\sigma_Z - \frac{\sigma_Z^2}{\sigma_X}\right)\phi\left(\frac{b-\mu_Y}{\sigma_Y}\right)\Phi\left(-\frac{c(0)-\frac{b+\mu_Z-\mu_X}{\sigma_Y}\times\frac{\rho_{XZ}\sigma_X-\sigma_Z}{\sigma_Y}}{\frac{\sigma_X}{\sigma_Y}\sqrt{1-\rho_{XZ}^2}}\right),$$

with $c(t) = \frac{a-b-\mu_Z - t\sigma_Z^2}{\sigma_Z}$. Substituting $c(0) = \frac{a-b-\mu_Z}{\sigma_Z}$ in equation (B7) yields:

(B8) $$\left.\frac{\partial m_1(t)}{\partial t}\right|_{t=0} = (\mu_Z)\int_{-\infty}^{c(0)}\phi(s)\Phi(-c_3(s,0))\partial s + (\mu_Z)\int_{c(0)}^{+\infty}\phi(s)\Phi(-c_4(s,0))\partial s$$

$$+(\sigma_X - \rho\sigma_Y)\phi\left(\frac{a-\mu_X}{\sigma_X}\right)\Phi\left(\frac{\rho\left(\frac{a-\mu_X}{\sigma_X}\right)-\left(\frac{b-\mu_Y}{\sigma_Y}\right)}{\sqrt{1-\rho^2}}\right)$$

$$+\left(\rho\sigma_X - \sigma_Y\right)\phi\left(\frac{b-\mu_Y}{\sigma_Y}\right)\Phi\left(\frac{\rho\left(\frac{b-\mu_Y}{\sigma_Y}\right)-\left(\frac{a-\mu_X}{\sigma_X}\right)}{\sqrt{1-\rho^2}}\right).$$

Combining equation (B8) with the denominator of equation (3), which does not depend on *t*, yields equation (4), i.e. the derivative of equation (3) (the Moment Generating Function) with respect to *t* and evaluated in *t*=0.

Table 1: Simulation results. Three model specifications are estimated with the two-step estimator of Wooldridge (1995) and the two-step first-difference Tobit estimator of Section 3. The specifications differ with respect to the conditional mean independence assumption (equation (22)). RMSE is the root mean squared error and MAD is the median absolute deviation. The true value of *ß* is 1.

Model 1: correct specification $h(x_i) = \bar{x}_i \mid \bar{x}_i \mid$

Two-step Tobit (Wooldridge, 1995)							*Two-step first-difference Tobit (Section 3)*					
N	T	Mean bias	RMSE	Median bias	MAD		N	T	Mean bias	RMSE	Median bias	MAD
500	2	-0.001	0.111	0.008	0.078		500	2	0.019	0.132	0.002	0.093
500	4	-0.014	0.058	-0.011	0.035		500	4	-0.007	0.078	-0.010	0.052
500	8	-0.007	0.051	-0.001	0.032		500	8	-0.006	0.051	-0.010	0.036
500	12	-0.006	0.046	-0.005	0.030		500	12	-0.003	0.038	-0.003	0.027
1000	2	-0.008	0.091	-0.008	0.068		1000	2	-0.009	0.108	-0.004	0.071
1000	4	0.001	0.043	-0.003	0.026		1000	4	-0.002	0.048	-0.008	0.033
1000	8	-0.007	0.033	0.002	0.024		1000	8	-0.006	0.030	-0.003	0.021
1000	12	-0.008	0.029	-0.014	0.021		1000	12	-0.001	0.030	-0.002	0.018

Model 2, incorrect specification: $h(x_i) = 0$, a standard random effects panel data model

Two-step Tobit (Wooldridge, 1995)							*Two-step first-difference Tobit (Section 3)*					
N	T	Mean bias	RMSE	Median bias	MAD		N	T	Mean bias	RMSE	Median bias	MAD
500	2	2.264	0.309	2.247	2.247		500	2	0.077	0.167	0.065	0.118
500	4	1.907	0.286	1.929	1.929		500	4	0.038	0.094	0.037	0.065
500	8	1.512	0.278	1.468	1.468		500	8	0.024	0.064	0.016	0.039
500	12	1.183	0.190	1.136	1.136		500	12	0.023	0.046	0.021	0.034
1000	2	2.282	0.265	2.267	2.267		1000	2	0.040	0.131	0.045	0.084
1000	4	1.882	0.203	1.862	1.862		1000	4	0.037	0.060	0.033	0.043
1000	8	1.508	0.188	1.508	1.508		1000	8	0.019	0.038	0.023	0.028
1000	12	1.186	0.140	1.200	1.200		1000	12	0.022	0.036	0.019	0.029

Model 3, incorrect specification: $h(x_i) = \bar{x}_i$, an often chosen empirical specification

Two-step Tobit (Wooldridge, 1995)							*Two-step first-difference Tobit (Section 3)*					
N	T	Mean bias	RMSE	Median bias	MAD		N	T	Mean bias	RMSE	Median bias	MAD
500	2	0.709	0.150	0.687	0.687		500	2	0.023	0.133	0.004	0.092
500	4	0.591	0.117	0.577	0.577		500	4	-0.003	0.079	-0.007	0.049
500	8	0.510	0.112	0.492	0.492		500	8	-0.005	0.051	-0.007	0.037
500	12	0.440	0.089	0.429	0.429		500	12	0.000	0.038	0.000	0.025
1000	2	0.705	0.138	0.722	0.722		1000	2	-0.005	0.109	-0.003	0.069
1000	4	0.616	0.092	0.606	0.606		1000	4	0.002	0.050	-0.002	0.031
1000	8	0.516	0.066	0.510	0.510		1000	8	-0.004	0.031	-0.002	0.022
1000	12	0.440	0.056	0.439	0.439		1000	12	0.001	0.030	0.001	0.018