Binary Response Panel Data Models with Sample Selection and Self Selection

Anastasia Semykina
Department of Economics
Florida State University
Tallahassee, FL 32306-2180
asemykina@fsu.edu

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
East Lansing, MI 48824-1038
wooldri1@msu.edu

February 19, 2015

We thank Georges Bresson, Bo Honore, participants of the Latin American Meeting of the Econometric Society 2011, participants of the 19th International Panel Data Conference, and the Midwest Econometrics Group 2013 Meeting participants for helpful comments.

Abstract

We consider estimating binary response models on an unbalanced panel, where the outcome of the dependent variable may be missing due to non-random selection, or there is self selection into a treatment. In the present paper, we first consider estimation of sample selection models and treatment effects using a fully parametric approach, where the error distribution is assumed to be normal in both primary and selection equations. Arbitrary time dependence in errors is permitted. Estimation of both coefficients and partial effects, as well as tests for selection bias are discussed. Furthermore, we consider a semiparametric estimator of binary response panel data models with sample selection that is robust to a variety of error distributions. The estimator employs a control function approach to account for endogenous selection and permits consistent estimation of scaled coefficients and relative effects.

JEL Classification: C33, C34, C35, C14

Key words: Binary response models, Sample selection, Panel data, Semiparametric, Treament effect

1 Introduction

Empirical researchers have shown growing interest in estimating binary response panel data models where sample selection and self-selection issues arise. A sample selection problem is a possibility whenever a panel data set is unbalanced. For example, binary response models with unbalanced panels arise in labor economics when studying the probability of worker being employed in a job with benefits with selection occuring due to nonrandom self-selection into the labor force. In studies that focus on estimating treatment effects, complications arise if self-selection into the treatment is not random. Estimation methods that address the selection problem can be helpful to empirical researchers who do policy evaluation with binary responses.

The problem of nonrandom selection has received substantial attention in the theoretical econometrics literature. Several new methods have been proposed for estimating selection models using panel data. However, the focus of that literature has been on linear or partially linear panel data models. For example, Kyriazidou (1997) derives semiparametric estimators for the linear panel data model under sample selection when the explanatory variables are strictly exogenous. Semykina and Wooldridge (2010) show how to estimate linear unobserved effects panel data models with endogenous explanatory variables and nonrandom sample selection. In this paper, we consider estimating binary response panel data models in the presence of nonrandom selection.

We consider two types of selection rules: (i) the selection variable is binary, and (ii) the selection variable is a corner solution or censored response.[1] In the binary selection case we show how to use the Mundlak (1978) device along with pooled maximum likelihood estimation to obtain simple estimators robust to general forms of dynamic misspecifica-

[1] In most applications, the selection variable is a corner solution, where some segment of the population chooses zero. Good examples are hours worked and quantity purchased of a good. In some cases, the variable is truly censored, especially when observability of y depends on whether an event occurs before a certain duration. If the duration is censored then the selection variable is properly viewed as censored. The statistical framework is essentially the same. For brevity, we refer to this case as the censored case.

tion. The setup is easily modified to allow estimation of treatment effects with a binary treatment.

When the selection variable is censored, we derive both parametric and semiparametric estimators using a control function approach. In the parametric case, we draw on the literature that considers estimating binary response models with endogenous explanatory variables using cross-section data (Blundell and Smith, 1986, Rivers and Vuong, 1986). In particular, we use a control function approach on the selected sample. The result is an extension of Wooldridge (1995), who studied linear models, to the binary response case. Our semiparametric approach is based on the semiparametric control function methods proposed by Blundell and Powell (2004), extended here to the missing data problem.

In addition to discussing consistent estimation of selection models, we propose the Lagrange Multiplier test and simple variable addition tests for the selection bias.

2 General Setup

Consider a binary response model

$$\begin{aligned} y_{it}^{*} &= x_{it}\beta + c_{i1} + u_{it1}, \\ y_{it} &= 1[y_{it}^{*} > 0], \quad t = 1, \ldots, T, \end{aligned} \tag{1}$$

where y_{it}^{*} is a latent variable, y_{it} is the observed variable, $1[\cdot]$ is an indicator function that takes on a value of one if the expression in brackets is true and is zero otherwise, x_{it} is a $1 \times M$ vector of time-varying explanatory variables, c_{i1} is the unobserved effect, and u_{it1} is the idiosyncratic error. In what follows, the observed covariates are assumed to be strictly exogenous conditional on c_{i1}. Specifically, for $x_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iT})$, assume that $\mathrm{D}(y_{it}|x_i, c_{i1}) = \mathrm{D}(y_{it}|x_{it}, c_{i1})$, where $\mathrm{D}(\cdot)$ denotes the distribution. Note that this assumption does not impose restrictions on how x_i and c_{i1} may be related.

In addition to estimating the vector of parameters, β, one is often interested in estimating partial effects, where the partial effect is defined as a ceteris paribus effect of an increase in explanatory variable x_{itk} on the expected value of y_{it}. In panel data models, c_{i1} is an unobserved variable that affects y_{it}, but cannot be consistently estimated on a usual panel where T is fixed. Therefore, as is common in nonlinear panel data models, we consider average partial effects (APEs), where an APE is the effect of an increase in an explanatory variable on the expected value of y_{it} averaged over the population distribution of the unobserved heterogeneity, c_{i1}. The discussion below covers the estimation of both parameters and APEs.

We introduce incidental truncation by modeling the selection process as

$$\begin{aligned} s_{it}^* &= z_{it}\delta + c_{i2} + u_{it2}, \\ s_{it} &= 1[s_{it}^* > 0], \quad t = 1, \ldots, T, \end{aligned} \tag{2}$$

where $z_{it} = (x_{it}, z_{it2})$ has dimension $1 \times L$ ($L > M$), s_{it}^* is a latent variable, and s_{it} is a selection indicator that equals one when y_{it} is observed and is zero otherwise. Thus, the vector of covariates in the selection equation contains x_{it} and at least one more variable. Similar to equation (1), assume that $\mathrm{D}(s_{it}|z_i, c_{i2}) = \mathrm{D}(y_{it}|z_{it}, c_{i2})$, where $z_i \equiv (z_{i1}, z_{i2}, \ldots, z_{iT})$. Moreover, assume $\mathrm{D}(y_{it}|x_i, z_i, c_{i1}) = \mathrm{D}(y_{it}|x_i, c_{i1}) = \mathrm{D}(y_{it}|x_{it}, c_{i1})$. A key assumption is that z_{it} is observed for all i and t, even though y_{it} is observed only when $s_{it} = 1$.

In some cases, s_{it}^* may be partially observable. In particular, in addition to the sign of s_{it}^*, the value of s_{it}^* may be known when y_{it} is observed. Examples include hours of work for a person who selects to be in the labor force (y_{it} may be an indicator for whether the work contract includes retirement benefits) and the amount of medical expenses borne by an individual who requires medical treatment (y_{it} may be an indicator equal to one if the

treatment included a particular medical procedure). In such an event, we have

$$\begin{aligned} s_{it}^{*} &= z_{it}\delta + c_{i2} + u_{it2}, \\ s_{it} &= \max\{0, s_{it}^{*}\}, \quad t = 1, \ldots, T. \end{aligned} \tag{3}$$

Partial observability of s_{it}^{*} makes it possible to estimate β and the APEs under fewer assumptions, as we have more information in a range of strictly positive values for s_{it}. In this paper, we discuss two cases: (i) when the selection rule follows equation (3), and (ii) when the selection rule is binary, as specified in equation (2).

Apart from the selection problem, additional complications result from the presence of unobserved heterogeneity. Within a random effects framework, where the unobserved effect is assumed to be independent of z_i, leaving it in the error leads to rescaling of parameters, but relative effects are preserved, as are average partial effects. The problem arises when z_i is not independent of c_{i1} and c_{i2}. Because independence of z_i and unobserved heterogeneity is rarely a realistic assumption, we employ the correlated random effects approach proposed by Chamberlain (1980). Specifically, let

$$\begin{aligned} c_{i1} &= \eta_1 + \bar{z}_i\xi_1 + a_{i1}, \\ c_{i2} &= \eta_2 + \bar{z}_i\xi_2 + a_{i2}, \end{aligned} \tag{4}$$

where $\bar{z}_i = T^{-1}\sum_{t=1}^{T} z_{it}$, and a_{i1} and a_{i2} are independent of z_i. Chamberlain (1980) proposed this assumption for binary response models with normally distributed errors. The normality assumption makes the model in (4) particularly attractive, but this model may also be useful for more general error distributions.

Under (4), the primary and selection equations can be rewritten as

$$\begin{aligned} y_{it} &= 1[\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + v_{it1} > 0], \\ s_{it} &= 1[\eta_2 + z_{it}\delta + \bar{z}_i\xi_2 + v_{it2} > 0], \quad t = 1, \ldots, T, \end{aligned} \tag{5}$$

where $v_{it1} = a_{i1} + u_{it1}$ and $v_{it2} = a_{i2} + u_{it2}$. Alternatively, if selection follows a censored (or corner solution) response, the system becomes

$$\begin{aligned} y_{it} &= 1[\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + v_{it1} > 0], \\ s_{it} &= \max\{0, \eta_2 + z_{it}\delta + \bar{z}_i\xi_2 + v_{it2}], \quad t = 1, \ldots, T. \end{aligned} \tag{6}$$

By construction, z_i and v_{it1} are independent, which implies that in the case when there is no selection (y_{it} is always observed) or selection is random with respect to (a_{i1}, u_{it1}), familiar parametric and semiparametric methods can be used to estimate β and the APEs.

Before discussing the different scenarios, it is useful to obtain the APEs based on the equation (1). It is convenient to obtain the APEs using the notion of the average structural function (ASF), introduced by Blundell and Powell (2004). For the binary-response models considered in this paper we will only be able to estimate the ASF and APEs in the case of the parametric model, which assumes normality. Therefore, the discussion of the ASF below is for the model where errors are assumed to have a joint normal distribution. The "structural" equation that underlies the estimation is

$$\mathrm{P}(y_{it} = 1 | x_{it}, c_{i1}) = \Phi(x_{it}\beta + c_{i1})$$

and the ASF is a function of the argument x_t:

$$\mathrm{ASF}(x_t) = \mathrm{E}_{c_{i1}}\left[\Phi(x_t\beta + c_{i1})\right],$$

so we average out over the distribution of c_{i1}. Now, we are not directly modeling the distribution of c_{i1}, but rather the conditional distribution $\mathrm{D}(c_{i1}|z_i)$. Therefore, the following expression based on iterated expectations is useful:

$$\begin{aligned} \mathrm{ASF}(x_t) &= \mathrm{E}_{z_i}\left\{\mathrm{E}\left[\Phi(x_t\beta + c_{i1})|z_i\right]\right\} \\ &= \mathrm{E}_{z_i}\left[\Phi(\eta_{a1} + x_t\beta_a + \bar{z}_i\xi_{a1})\right], \end{aligned} \tag{7}$$

where $\beta_a = \beta/\sqrt{1+\sigma_{a_1}^2}$ and similarly for the other parameters with an a subscript. This expression is derived in Papke and Wooldridge (2008).

Equation (7) is the basis for estimating average partial effects. In particular, we can take derivatives or changes with respect to the elements of x_t and then average out the z_i. Note that the scaled parameters provide the directions of the effects and ratios of the scaled parameters are the same as ratios of the orginal parameters. Because it is the scaled parameters that appear in the ASF, those are actually more interesting for our purposes. As it turns out, the unscaled coefficients are not generally identified, anyway, unless we were to make strong serial independence assumptions and then use a much more complicated estimation method. Thus, in the next subsection we will drop the a subscript with the understanding that the coefficients have been implicitly scaled by the variance of $a_{i1} + u_{it1}$.

A major impediment in estimating β_a and the APEs is that v_{it1} and v_{it2} are likely to be correlated, which means that selection is related to unobservables affecting y_{it}. One way to solve the selection problem is to make parametric assumptions about the joint distribution of (v_{it1}, v_{it2}) and use the maximum likelihood estimation. Another possibility is to use a semiparametric estimator that imposes a linear index restriction as in (5), but remains agnostic about the specific form of the error distribution. We consider both approaches.

3 Parametric Model and Estimation

3.1 General Parametric Model

We start by assuming that (v_{it1}, v_{it2}) have a zero mean bivariate normal distribution. Because of the discussion in the previous section, we normalize the variance of v_{it1} as $\text{Var}(v_{it1}) = 1$, so we are actually estimating the scaled coefficients in (7). Generally, $\text{Var}(v_{it2}) = \sigma^2$, although when s_{it} is binary there is no loss of generality in setting $\sigma^2 = 1$ (and we could not identify σ^2, anyway). Under normality, v_{it1} and v_{it2} are linked as

$$v_{it1} = \gamma v_{it2} + e_{it1}, \quad t = 1, \ldots, T, \tag{8}$$

where $\gamma = \rho/\sigma$, $\rho = \text{Corr}(v_{it1}, v_{it2})$, and e_{it1} is independent of z_i and v_{it2} with a normal distribution. Therefore, we can write

$$\begin{aligned} y_{it} &= 1[\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \gamma v_{it2} + e_{it1} > 0], \\ e_{it1}|z_i, v_{it2} &\sim Normal(0, 1-\rho^2), \quad t = 1, \ldots, T. \end{aligned} \tag{9}$$

The equations in (9) demonstrate that conditioning on v_{it2} is irrelevant if selection is random, that is, $\rho = 0$. It is a nonzero correlation between v_{it1} and v_{it2} that makes the selection non-ignorable.

A basic but important fact is that because s_{it} is a deterministic function of z_i and v_{it2}, it follows that

$$\begin{aligned} &y_{it} = 1[\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \gamma v_{it2} + e_{it1} > 0], \\ &e_{it1}|z_i, v_{it2}, s_{it} \sim Normal(0, 1-\rho^2), \quad t = 1, \ldots, T. \end{aligned} \tag{10}$$

Therefore, by including v_{it2} in (10), we can solve the non-random selection problem. This

is an example of the "control function" approach proposed by Blundell and Smith (1986) and Rivers and Vuong (1988) for the case of endogenous explanatory variables. Here, we use the control function to account for the factors responsible for selection.

Due to normality of e_{it1}, it is also true that

$$\mathrm{P}(y_{it} = 1 | z_i, v_{it2}, s_{it}) = \mathrm{P}(y_{it} = 1 | z_i, v_{it2}) = \Phi\left(\frac{\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \gamma v_{it2}}{\sqrt{1-\rho^2}}\right), \tag{11}$$

which is a probit model with parameters rescaled by a common factor $(\sqrt{1-\rho^2})^{-1}$. Thus, if v_{it2} were known, one could estimate $\beta/\sqrt{1-\rho^2}$ rather easily. Of course, v_{it2} is never known; however, in some cases it can be estimated whenever $s_{it} = 1$, and that suffices to consistently estimate the parameters. Because estimation of (10) is performed on the selected sample, one only needs to know v_{it2} when y_{it} is observed, which can be estimated when selection follows, say, a Tobit model.

3.2 Estimation When Selection Variable Is Censored

We first consider the case where selection follows a Tobit model and all assumptions that were used for deriving (10) hold. Specifically, make the following assumption:

ASSUMPTION 3.2. (i) y_{it} is determined by equation (1), (ii) s_{it} is determined by equation (3), (iii) c_{i1} and c_{i2} follow (4), (iv) (v_{it1}, v_{it2}) are independent of z_i and have a zero mean bivariate normal distribution, where $v_{it1} = a_{i1} + u_{it1}$, $v_{it2} = a_{i2} + u_{it2}$, $\mathrm{Var}(v_{it1}) = 1$, and $\mathrm{Var}(v_{it2}) = \sigma^2$.

Under Assumption 3.2, the scaled parameters

$$\eta_{1\rho} \equiv \frac{\eta_1}{\sqrt{1-\rho^2}},\ \beta_\rho \equiv \frac{\beta}{\sqrt{1-\rho^2}},\ \xi_{1\rho} \equiv \frac{\xi_1}{\sqrt{1-\rho^2}},\ \text{and}\ \gamma_\rho \equiv \frac{\gamma}{\sqrt{1-\rho^2}}$$

can be consistently estimated in two steps:

1. Use pooled Tobit to estimate equation

$$s_{it} = \max\{0, \eta_2 + z_{it}\delta + \bar{z}_i\xi_2 + v_{it2}\}.$$

   For $s_{it} > 0$, obtain $\hat{v}_{it2} = y_{it} - \hat{\eta}_2 - z_{it}\hat{\delta} - \bar{z}_i\hat{\xi}_2$.

2. For $s_{it} > 0$, estimate (10) by pooled probit, where use $\hat{v}_{it2}$ in place of v_{it2}.

Notice that neither step one nor step two imposes restrictions on the form of serial dependence in the error terms. The estimator at each step is the partial MLE (either pooled Tobit or pooled probit), which does not require specifying the full likelihood function. Hence, the errors in each equation may be arbitrarily serially correlated, and, in fact, are expected to be serially correlated because, by construction, part of the unobserved effect remains in the error. Consequently, the estimator of the asymptotic variance of the second-step estimator should be made robust to serial dependence. Moreover, standard errors should account for the first-step estimation. A time-specific intercept is accommodated by including time indicators in the set of covariates at each step.

The two-step estimation procedure focuses on obtaining consistent estimators of $\eta_{1\rho}$, β_ρ, $\xi_{1\rho}$, and γ_ρ, rather than original parameters in the population model. The estimators of the original parameters can be obtained as

$$\hat{\rho} = \hat{\gamma}_\rho\hat{\sigma} \cdot (1 + \hat{\gamma}_\rho^2\hat{\sigma}^2)^{-1/2}, \quad \hat{\beta} = \hat{\beta}_\rho(1 - \hat{\rho}^2)^{-1/2} = \hat{\beta}_\rho(1 + \hat{\gamma}_\rho^2\hat{\sigma}^2)^{-1/2}, \tag{12}$$

and so on, where $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$, and $\hat{\sigma}^2$ is the estimated variance of v_{it2} from the Tobit regression. A consistent estimator of a relative effect for two continous covariates is easily obtained as $\hat{\beta}_{\rho,j}/\hat{\beta}_{\rho,k}$.

Given the estimates $\hat{\beta}$ in (12), with similar expressions for $\hat{\eta}_1$ and $\hat{\xi}_1$, the APEs are easily obtained. For a single value – that is, not as a function of x_t – we can average a derivative across (x_{it}, z_i).

Because z_{it} is observed for all t, this APE can be consistently estimated as

$$\widehat{\text{APE}}_k = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi(\hat{\eta}_1 + x_{it}\hat{\beta} + \bar{z}_i\hat{\xi}_1)\right]\hat{\beta}_k.$$

To obtain APEs at different values of x_t, we use

$$\widehat{\text{APE}}_k(x_t) = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi(\hat{\eta}_1 + x_t\hat{\beta} + \bar{z}_i\hat{\xi}_1)\right]\hat{\beta}_k.$$

The APE of a discrete explanatory variable, say x_{itm}, can be estimated by evaluating the response probability at the two different values, $x_{tm}^{(1)}$ and $x_{tm}^{(0)}$, and computing the average difference in probabilities:

$$\begin{aligned}
& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\Phi(\hat{\eta}_1 + x_{it}^{(1)}\hat{\beta} + \bar{z}_i\hat{\xi}_1) - \Phi(\hat{\eta}_1 + x_{it}^{(0)}\hat{\beta} + \bar{z}_i\hat{\xi}_1)\right] \\
x_{it}^{(0)} &\equiv (x_{it1}, \ldots, x_{it,m-1}, x_{tm}^{(0)}, x_{it,m+1}, \ldots, x_{itM}), \\
x_{it}^{(1)} &\equiv (x_{it1}, \ldots, x_{it,m-1}, x_{tm}^{(1)}, x_{it,m+1}, \ldots, x_{itM}).
\end{aligned}$$

In the leading case, x_{itm} is a dummy variable and $x_{tm}^{(1)} = 1$, $x_{tm}^{(0)} = 0$.

Standard errors of $\hat{\beta}$ and APEs can be obtained using the delta method. However, because the corresponding variance formulas will be rather complicated, panel bootstrap can serve as a convenient alternative.

Rather than using a two-step estimation procedure, it is possible to estimate the parameters in one step by specifying the joint distribution of (y_{it}, s_{it}) given z_i for each t, and employing the partial maximum likelihood estimator (partial MLE). Specifically, for each t, the joint density function is

$$f(y_{it}, s_{it}|z_i) = \left\{[\Phi(r_{it})]^{y_{it}}[1 - \Phi(r_{it})]^{1-y_{it}}\frac{1}{\sigma}\,\phi\left(\frac{q_{it}}{\sigma}\right)\right\}^{1[s_{it}>0]}\left\{1 - \Phi\left(\frac{q_{it}}{\sigma}\right)\right\}^{1[s_{it}=0]}, \quad (13)$$

where

$$r_{it} \equiv \frac{\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \frac{\rho}{\sigma}(s_{it} - \eta_{2t} - z_{it}\delta_t - \bar{z}_i\xi_{2t})}{\sqrt{1-\rho^2}}, \tag{14}$$

$$q_{it} \equiv \eta_{2t} + z_{it}\delta_t + \bar{z}_i\xi_{2t}. \tag{15}$$

The MLE estimates are obtained by taking the logarithm of the conditional joint density, summing it up over all i and t, and maximizing with respect to parameters. Notice that it is not necessary to specify the full likelihood function, $f(y_{i1}, \ldots, y_{iT}, s_{i1}, \ldots, s_{iT}|z_i)$, which would be very complicated because of the serial dependence in errors. Within the partial MLE framework, it is sufficient to specify $f(y_{it}, s_{it}|z_i)$, $t = 1, \ldots, T$. When we used partial MLE the estimator of the asymptotic variance should be made robust to serial correlation in the score functions. The advantage of partial MLE over the two-step estimator is that the variance that accounts for serial dependence is correct and no further adjustments are needed to obtain valid standard errors for the parameters, and the estimated parameters would not be scaled by $(1-\hat{\rho}^2)^{-1/2}$. Nevertheless, the asymptotic variances for the APEs would still be rather complicated, and one might still want to use the panel boostrap to obtain valid standard errors.

3.3 Estimation When Selection Variable Is Binary

In this section, we consider estimation when the selection rule is binary. It is also assumed that $\text{Var}(v_{it2}) = 1$ and all assumptions used for deriving (10) hold. More formally,

ASSUMPTION 3.3. (i) y_{it} is determined by equation (1), (ii) s_{it} is determined by equation (2), (iii) c_{i1} and c_{i2} follow (4), (iv) (v_{it1}, v_{it2}) are independent of z_i and have a zero mean bivariate normal distribution, where $v_{it1} = a_{i1} + u_{it1}$, $v_{it2} = a_{i2} + u_{it2}$, $\text{Var}(v_{it1}) = \text{Var}(v_{it2}) = 1$.

Under Assumption 3.3, parameters in the model can be consistently estimated by

MLE. For each t, the joint density function of (y_{it}, s_{it}) conditional on z_i is

$$\begin{aligned} f(y_{it}, s_{it}|z_i) &= [P(y_{it}=1|s_{it}=1,z_i)P(s_{it}=1,z_i)]^{y_{it}s_{it}} \\ &\times [P(y_{it}=0|s_{it}=1,z_i)P(s_{it}=1|z_i)]^{(1-y_{it})s_{it}}[P(s_{it}=0|z_i)]^{(1-s_{it})}, \quad (16) \end{aligned}$$

where

$$\begin{aligned} P(y_{it}=1|s_{it}=1,z_i) &= \mathrm{E}(y_{it}=1|s_{it}=1,z_i) = \mathrm{E}[\mathrm{E}(y_{it}=1|v_{it2},z_i)|s_{it}=1,z_i] \\ &= \mathrm{E}[\Phi(r_{it})|s_{it}=1,z_i] = \frac{1}{\Phi(q_{it})}\int_{-\infty}^{q_{it}}\Phi(r_{it})\phi(\nu)d\nu, \quad (17) \\ P(y_{it}=0|s_{it}=1,z_i) &= \frac{1}{\Phi(q_{it})}\int_{-\infty}^{q_{it}}[1-\Phi(r_{it})]\phi(\nu)d\nu, \quad (18) \\ P(s_{it}=1|z_i) &= \Phi(q_{it}), \quad (19) \\ P(s_{it}=0|z_i) &= 1-\Phi(q_{it}), \quad (20) \end{aligned}$$

where $r_{it} = (\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \rho\nu)(1-\rho^2)^{-1/2}$ and q_{it} is defined in (15). Thus, the conditional joint likelihood function for unit i in period t is given by

$$\begin{aligned} L_{it} \equiv f(y_{it}, s_{it}|z_i) &= \left[\int_{-\infty}^{q_{it}}\Phi(r_{it})\phi(\nu)d\nu\right]^{y_{it}s_{it}} \\ &\times \left[\int_{-\infty}^{q_{it}}[1-\Phi(r_{it})]\phi(\nu)d\nu\right]^{(1-y_{it})s_{it}}[1-\Phi(q_{it}))]^{(1-s_{it})}. \quad (21) \end{aligned}$$

Similar to the Tobit case, the partial MLE estimates are obtained by taking the logarithm of the conditional joint density function, summing it up over all i and t, and maximizing the resulting sum with respect to parameters. The variance-covariance matrix should be made robust to serial correlation. Some statistical software have built-in commands that allow to easily implement this estimator in practice.[2]

Note that equation (7) still holds. Thus, the estimation of APEs discussed in Section

[2]For example, in Stata this estimation approach can be implemented by pooling the data and estimating the augmented equation that includes time means using "heckprob" command.

3.2 is directly applicable here.

The maximum likelihood estimators discussed in this and previous sections can be made robust to heteroskedasticity by appropriately modifying the joint likelihood function. This requires specifying error variances and the covariance as functions of $(x_{it}, \bar{z}_i)$. In practice, it is common to use an exponential function (see, for example, Wooldridge 2010). Accounting for heteroskedasticity makes parametric estimators more reliable when the normality assumption is violated.

3.4 Estimating Treatment Effects

The joint MLE discussed in Section 3.3 can also be used for cases when y_{it} is always observed, and s_{it} is a binary treatment indicator, so that there is no sample selection problem, but usual self-selection into the treatment is present. In this case, s_{it} appears as an additional explanatory variable:

$$\begin{aligned} y_{it} &= 1[\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \psi s_{it} + \gamma v_{it2} + e_{it1} > 0], \\ e_{it1}|z_i, v_{it2}, s_{it} &\sim Normal(0, 1-\rho^2), \quad t = 1, \ldots, T. \end{aligned} \tag{22}$$

Because s_{it} is endogenous, its individual time means should not be included in $\bar{z}_i$.

The conditional joint likelihood function for unit i in period t becomes

$$\begin{aligned} L_{it} \equiv f(y_{it}, s_{it}|z_i) &= \left[\int_{-\infty}^{q_{it}} \Phi(r_{it})\phi(\nu)d\nu\right]^{y_{it}s_{it}} \\ &\times \left[\int_{-\infty}^{q_{it}} [1-\Phi(r_{it})]\phi(\nu)d\nu\right]^{(1-y_{it})s_{it}} \\ &\times \left[1-\int_{-\infty}^{q_{it}} \Phi(r_{it})\phi(\nu)d\nu\right]^{y_{it}(1-s_{it})} \\ &\times \left[1-\int_{-\infty}^{q_{it}} [1-\Phi(r_{it})]\phi(\nu)d\nu\right]^{(1-y_{it})(1-s_{it})}, \end{aligned} \tag{23}$$

where $r_{it} = (\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \psi s_{it} + \rho\nu)(1-\rho^2)^{-1/2}$. Similar to Section 3.3, the estimator is partial MLE. Statistical inference should generally account for serial correlation in the score functions.[3] Similar to the discussion above, the estimator can be made robust to heteroskedasticity.

In most cases where s_{it} is a policy indicator, or "treatment" indicator, the main interest is in the average treatment effect. This is easily obtained once the pooled MLEs $\hat{\eta}_1$, $\hat{\beta}$, $\hat{\xi}_1$, and $\hat{\psi}$ are obtained:

$$\widehat{\text{ATE}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\Phi\left(\hat{\eta}_1 + x_{it}\hat{\beta} + \bar{z}_i\hat{\xi}_1 + \hat{\psi}\right) - \Phi\left(\hat{\eta}_1 + x_{it}\hat{\beta} + \bar{z}_i\hat{\xi}_1\right)\right]. \tag{24}$$

We can also obtain ATEs for different subpopulations by fixing x_t at different values (which means dropping the i subscript in (24)).

Many embellishments are possible. For example, the coefficient on s_{it} can be allowed to change with t in an arbitrary way (by including interactions between time period dummies and s_{it}), and then one could estimate an ATE for different time periods.

3.5 Testing for Selection Bias

Even when the model is parametric, correcting for selection bias may be somewhat challenging. As discussed in Section 3.2, the two-step estimation under the tobit-type selection mechanism involves obtaining standard errors that account for the first-stage estimation. For both censored and binary section models, when parameters are estimated by joint partial MLE, computational problems may arise. Therefore, it is useful to have a simple test for selection bias, which would help to identify cases when correction is necessary.

[3]Some statistical software packages have built-in commands that perform such estimation. For example, in Stata estimating treatment effects can be implemented by pooling the data and estimating the augmented equation (with time averages) using the "biprobit" command. Standard errors robust to serial dependence can be obtained using "cluster" option.

When the selection variable is censored, a simple test for selection bias can be performed by testing $H_0 : \gamma = 0$ after estimating equation (10) using the two-step procedure outlined in Section 3.2. An attractive feature of the test is that there is no need to correct for the first-step estimation when computing the test statistic. A standard t-statistic (Wald statistic) that uses a standard error robust to serial correlation is valid.

When the selection variable is binary, the Lagrange multiplier (score) test can be used. Let $\theta = (\eta_1, \beta', \xi_1')'$ and $w_{it} = (1, x_{it}, \bar{z}_i)$. Let $\tilde{r}_{it}$ be r_{it} evaluated at $\rho = 0$ and parameter estimates $\tilde{\theta}$, which are obtained from the restricted model. The restricted model is simply a Chamberlain pooled probit estimation using the unbalanced panel. Let $\hat{q}_{it}$ be q_{it} evaluated at the parameters in the probit estimation at time t, $(\hat{\eta}_{2t}, \hat{\delta}_t, \hat{\xi}_{2t})$, where q_{it} is given in (15). Using the likelihood function in equation (21) as a starting point, the Lagrange multiplier (LM) statistic for testing $H_0 : \rho = 0$ is given by[4]

$$LM = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{S}_{it,\rho} \right)' \tilde{A}^{22} \left[\tilde{V}_{22} \right]^{-1} \tilde{A}^{22} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{S}_{it,\rho} \right) / N, \tag{25}$$

[4]See Wooldridge (2010), Section 12.6.2 for the detailed derivation of equation (25).

where

$$\tilde{S}_{it,\rho} \equiv \frac{\partial \ln L_{it}}{\partial \rho}|_{\theta=\tilde{\theta},\rho=0} = s_{it}\frac{y_{it}-\Phi(\tilde{r}_{it})}{\Phi(\tilde{r}_{it})[1-\Phi(\tilde{r}_{it})]}\phi(\tilde{r}_{it})\hat{\lambda}_{it}, \tag{26}$$

$$\tilde{A} = -\frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left(\frac{\partial \ln L_{it}}{\partial\theta\partial\theta'}|s_{it},z_i\right)|_{\theta=\tilde{\theta},\rho=0} & \sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left(\frac{\partial \ln L_{it}}{\partial\rho\partial\theta}|s_{it},z_i\right)|_{\theta=\tilde{\theta},\rho=0} \\ \sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left(\frac{\partial \ln L_{it}}{\partial\theta\partial\rho}|s_{it},z_i\right)|_{\theta=\tilde{\theta},\rho=0} & \sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left(\frac{\partial \ln L_{it}}{\partial\rho\partial\rho}|s_{it},z_i\right)|_{\theta=\tilde{\theta},\rho=0} \end{pmatrix},$$

$$= \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\tilde{r}_{it})^2}{\Phi(\tilde{r}_{it})[1-\Phi(\tilde{r}_{it})]}\omega_{it}'\omega_{it} & \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\tilde{r}_{it})^2}{\Phi(\tilde{r}_{it})[1-\Phi(\tilde{r}_{it})]}\omega_{it}'\hat{\lambda}_{it} \\ \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\tilde{r}_{it})^2}{\Phi(\tilde{r}_{it})[1-\Phi(\tilde{r}_{it})]}\hat{\lambda}_{it}\omega_{it} & \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\tilde{r}_{it})^2}{\Phi(\tilde{r}_{it})[1-\Phi(\tilde{r}_{it})]}\hat{\lambda}_{it}^2 \end{pmatrix},$$

$$\tilde{A}^{-1} = \begin{pmatrix} \tilde{A}^{11} & \tilde{A}^{12} \\ \tilde{A}^{21} & \tilde{A}^{22} \end{pmatrix}, \tag{27}$$

$$\tilde{V} = \tilde{A}^{-1}\tilde{B}\tilde{A}^{-1} = \begin{pmatrix} \tilde{V}_{11} & \tilde{V}_{12} \\ \tilde{V}_{21} & \tilde{V}_{22} \end{pmatrix}, \tag{28}$$

$$\tilde{B} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\tilde{S}_{it,\rho}\sum_{t=1}^{T}\tilde{S}_{it,\rho}'\right), \tag{29}$$

$$\hat{\lambda}_{it} \equiv \frac{\phi(\hat{q}_{it})}{\Phi(\hat{q}_{it})}. \tag{30}$$

Matrix $\tilde{A}$ above is an estimator of the expected value of the negative Hessian matrix that uses the expected Hessian form. Alternatively, the outer product of scores or usual Hessian form of the matrix could be used.

Another simple test, which is asymptotically equivalent to the LM test, is a variable addition test. The test can be performed as follows:

(i) Use probit to estimate the selection equation for each t. For each i and t, compute the inverse Mills ratio, $\hat{\lambda}_{it}$. Alternatively, one can use pooled probit to estimate one set of parameters (although separate time intercepts is usually a must).

(ii) For $s_{it} = 1$, augment the primary probit equation by $\hat{\lambda}_{it}$ and estimate by pooled probit. Use the t-test (robust to serial correlation) to test statistical significance of

$\hat{\lambda}_{it}$.

Under the null hypothesis the coefficient on $\hat{\lambda}_{it}$ is zero, and so estimation of the parameters in $\hat{\lambda}_{it}$ does not affect the $\sqrt{N}$-asymptotic distribution of the test statistic. In other words, there is no need to account for the first-step estimation when performing the test, but there is a need to account for serial correlation.

To show that the variable addition test is asymptotically equivalent to the LM test, first write the second-step likelihood function for unit i in period t as

$$L_{it} = s_{it}\Phi(\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \gamma\lambda_{it})^{y_{it}}[1 - \Phi(\eta_1 + x_{it}\beta + \bar{z}_i\xi_1 + \gamma\lambda_{it}]^{(1-y_{it})}, \qquad (31)$$

where, to simplify notation, we ignore the fact that λ_{it} is estimated at the first step. As mentioned above, replacing λ_{it} with its consistent estimator will not affect the asymptotic distribution of the test statistic when the null is true.

Based on (31), the score vector is

$$S_{it} = s_{it}\frac{y_{it} - \Phi(w_{it}\theta + \gamma\lambda_{it})}{\Phi(w_{it}\theta + \gamma\lambda_{it})[1 - \Phi(w_{it}\theta + \gamma\lambda_{it})]}\phi(w_{it}\theta + \gamma\lambda_{it})\begin{pmatrix} w_{it} \\ \lambda_{it} \end{pmatrix}. \qquad (32)$$

Summing the score vector over all i and t and using a mean-value expansion about the true parameter vector gives

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{S}_{it} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}S_{it} - A\sqrt{N}\begin{pmatrix} \hat{\theta} - \theta \\ \hat{\gamma} - \gamma \end{pmatrix} + o_p(1), \qquad (33)$$

where $\hat{S}_{it}$ is the score vector evaluated at the estimated parameter values, $(\hat{\theta}', \hat{\gamma})'$, and A is the expected value of the negative Hessian matrix.

From (33), it follows that

$$\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta\\ \hat{\gamma}-\gamma\end{pmatrix} = -A^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{S}_{it}-S_{it})+o_p(1) = A^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}S_{it}+o_p(1). \quad (34)$$

When testing $\mathrm{H}_0 : \gamma = 0$, the robust Wald test statistic, is given by

$$W = (\hat{\gamma}-\gamma)'(\hat{V}_{22}/N)^{-1}(\hat{\gamma}-\gamma) = \sqrt{N}(\hat{\gamma}-\gamma)'\hat{V}_{22}^{-1}\sqrt{N}(\hat{\gamma}-\gamma), \quad (35)$$

where

$$\hat{V} = \hat{A}^{-1}\hat{B}\hat{A}^{-1} = \begin{pmatrix}\hat{V}_{11} & \hat{V}_{12}\\ \hat{V}_{21} & \hat{V}_{22}\end{pmatrix}, \quad (36)$$

$$\hat{B} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{S}_{it,\gamma}\sum_{t=1}^{T}\hat{S}'_{it,\gamma}\right), \quad (37)$$

$$\hat{A} = \frac{1}{N}\begin{pmatrix}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\hat{p}_{it})^2}{\Phi(\hat{p}_{it})[1-\Phi(\hat{p}_{it})]}\omega'_{it}\omega_{it} & \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\hat{p}_{it})^2}{\Phi(\hat{p}_{it})[1-\Phi(\hat{p}_{it})]}\omega'_{it}\lambda(\hat{q}_{it})\\ \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\hat{p}_{it})^2}{\Phi(\hat{p}_{it})[1-\Phi(\hat{p}_{it})]}\lambda(\hat{q}_{it})\omega_{it} & \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\frac{\phi(\hat{p}_{it})^2}{\Phi(\hat{p}_{it})[1-\Phi(\hat{p}_{it})]}\lambda(\hat{q}_{it})^2\end{pmatrix},$$

$$\hat{p}_{it} = w_{it}\hat{\theta}+\hat{\gamma}\lambda(\hat{q}_{it}), \quad (38)$$

$$\hat{A}^{-1} \xrightarrow{p} A^{-1} = \begin{pmatrix}A^{11} & A^{12}\\ A^{21} & A^{22}\end{pmatrix}. \quad (39)$$

From (34), we can also write the Wald statistic as

$$W = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}S_{it,\gamma}\right)' A^{22}\hat{V}_{22}^{-1}A^{22}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}S_{it,\gamma}\right)/N, \quad (40)$$

which is asymptotically distributed as χ^2_1. It is easily seen that under the null of no selection bias ($\rho = 0, \gamma = 0$), the scores and Hessian matrices used in (25) and (40) are the same when evaluated at true parameter values. Moreover, when the null is true, $\hat{\gamma} \xrightarrow{p} 0$, and $\sqrt{N}(\tilde{\theta}-\theta)$ and $\sqrt{N}(\hat{\theta}-\theta)$ converge in distribution. Therefore, $LM - W \xrightarrow{p} 0$, so

that the tests are asymptotically equivalent.

4 Semiparametric Estimation When Selection Variable Is Censored

In this section, we consider a semiparametric binary dependent variable model with non-random selection. The discussion is limited to the case when the selection variable is censored, as stated in equation (3), so that $s_{it} = s_{it}^*$ whenever y_{it} is observed. The unobserved effect is still modeled using the Chamberlain's device, i.e. (4) is assumed to hold. A key distinction between the approach of this section and the estimators discussed in the previous section is that the assumption of joint normality of the error terms in the selection and primary equations is dropped. Instead, we employ the control function approach of Blundell and Powell (2004) and derive a consistent estimator of parameters under relatively weak distributional assumptions.

Assume that the following condition holds:

$$v_{it1}|z_i, s_{it} \sim v_{it1}|z_i, v_{it2} \sim v_{it1}|v_{it2}, \quad t = 1, \ldots, T. \tag{41}$$

That is, for each given t, the conditional distribution of v_{it1} given the exogenous and selection variables is completely described by error v_{it2}. Additionally, we need to change the notation from the previous sections. Let $w_{it} = (x_{it}, \bar{z}_i)$ and where we now drop the time effects in defining w_{it}. Now, $\theta = (\beta', \xi_1')'$. Then, under condition (41) the conditional expectation of y_{it} is given by

$$\mathrm{E}(y_{it}|z_i, s_{it}) = \mathrm{P}(y_{it} = 1|z_i, v_{it2}) = \mathrm{P}(-v_{it1} \leq w_{it}\theta|z_i, v_{it2}) = F(w_{it}\theta, v_{it2}). \tag{42}$$

where $F(\cdot, v_{it2})$ is the cumulative distribution function of $-v_{it1}$ conditional on v_{it2}.

Similar to the parametric case, one can use a (semiparametric) estimator to obtain $\hat{v}_{it2}$ and use it to estimate θ. We will return to this issue when discussing the estimation procedure. For now, to simplify the presentation, assume that v_{it2} is known.

Assuming that function $F(w_{it}\theta, v_{it2})$ is continuous and monotonic in its first argument, it can be inverted with respect to its first argument. Denote the inverse function $\psi(\cdot, v) \equiv F^{-1}(\cdot, v)$. Then, define $r_{it} = (w_{it}, v_{it2})$, $g(r_{it}) \equiv \mathrm{E}(y_{it}|r_{it})$, can write

$$\begin{aligned} & \psi[g(r_{it}), v_{it2}] = w_{it}\theta \\ \text{or,} \qquad & \psi[g(r_{it}), v_{it2}] - w_{it}\theta = 0, \end{aligned} \tag{43}$$

with probability approaching one. The result in (43) implies that for any two observations, i and j, in a given period t, if $\mathrm{E}(y_{it}|r_{it}) = \mathrm{E}(y_{jt}|r_{jt})$ and $v_{it2} = v_{jt2}$, it should be the case that $w_{it}\theta = w_{jt}\theta$ with probability approaching one. As discussed in Blundell and Powell (2004), this property permits constructing a matching estimator, where any two observations with the same (or, in practice, 'similar') conditional expectations for the binary dependent variable in the primary equation and the same error terms in the selection equation are matched and used to recover the vector of parameters θ, which satisfies the equality of the indices $w_{it}\theta$ and $w_{jt}\theta$.

Formally, for a non-negative weighting function $\omega_{ijt} \equiv \omega(r_{it}, r_{jt})$, for each t can write

$$\mathrm{E}\left[\omega_{ijt} \cdot ((w_{it} - w_{jt})\theta)^2 | g(r_{it}) = g(r_{jt}), v_{it2} = v_{jt2}\right] = 0, \tag{44}$$

so that

$$\begin{aligned} & \sum_{t=1}^{T} \mathrm{E}\left[\omega_{ijt} \cdot ((w_{it} - w_{jt})\theta)^2 | g(r_{it}) = g(r_{jt}), v_{it2} = v_{jt2}\right] \\ & \qquad \equiv \theta' \Sigma_{\omega} \theta = 0, \end{aligned} \tag{45}$$

where $\Sigma_\omega \equiv \sum_{t=1}^{T} \Sigma_\omega^t$, $\Sigma_\omega^t \equiv \mathrm{E}[\omega_{ijt} \cdot (w_{it} - w_{jt})'(w_{it} - w_{jt}) | g(r_{it}) = g(r_{jt}), v_{it2} = v_{jt2}]$. Assuming that in the population θ is not zero, Σ_ω must be singular. Moreover, assuming that θ is a unique solution to the population condition (45), Σ_ω has only one zero eigenvalue. A consistent estimator of θ can then be obtained by constructing a sample analog of matrix Σ_ω and finding its eigenvalue that is closest to zero. The estimator of θ is the eigenvector that corresponds to the smallest eigenvalue. This approach was also used by Ahn, Ichimura and Powell (2004) in application to general single-index models with exogenous regressors.

Based on discussion above, estimation of θ is performed in two steps:

1. For each t and $s_{it} > 0$, obtain consistent estimators of the parameters in v_{it2} and the function $g(\cdot)$.

2. For $s_{it} > 0$, find the eigenvector of the sample analog of matrix Σ_ω that corresponds to the eigenvalue that is closest to zero.

At step one, v_{it2} needs to be estimated first. Similar to the Tobit case, because v_{it2} is a true structural error that has to be independent of exogenous variables, it is crucial that the selection equation is correctly specified. It is also more appropriate to use a general version of Chamberlain's correlated random coefficients model of the form:

$$s_{it} = \max\{0, \eta_2 + z_{it}\delta + z_{i1}\xi_{21} + \cdots + z_{iT}\xi_{2T} + v_{it2}\}, \quad t = 1, \ldots, T. \tag{46}$$

If error v_{it2} is continuously distributed with median zero, and its density function is positive at zero, then parameters in equation (46) can be consistently estimated by the censored least absolute deviations estimator proposed by Powell (1984). If the error distribution is also symmetric, then Powell's symmetrically trimmed least squares estimator (Powell, 1986) can be used. Under appropriate regularity conditions these estimators are consistent and $\sqrt{N}$-asymptotically normal for a variety of error distributions. They

are also robust to heteroskedasticity. Moreover, because selection equation is estimated separately for each t, $\{v_{it2}\}_{t=1}^{T}$ may be arbitrarily serially related and can have different variances.

Alternatively, a nonparametric estimator proposed by Lewbel and Linton (2002) could be used to estimate the conditional mean of s_{it}^{*} for each t, which then could be subtracted from s_{it} (for $s_{it} > 0$) to obtain $\hat{v}_{it2}$. This approach involves obtaining nonparametric estimators of $\mathrm{E}(s_{it}|z_i)$ and $\mathrm{E}\{1[s_{it} > 0]|\mathrm{E}(s_{it}|z_i)\}$, followed by integration of a function of the latter estimator. An important advantage of this estimator is that the conditional mean of s_{it}^{*} does not have to be linear in parameters. However, estimation is relatively complicated and is subject to the "curse of dimensionality" problem. Moreover, the estimator has a relatively slow rate of convergence. Therefore, using simpler $\sqrt{N}$-consistent Powell's estimators may be preferred.

While modeling the unobserved effect as a linear function of exogenous variables in all time periods – as in equation (46) above – is somewhat restrictive, this approach has important advantages over other existing estimators of unobserved effects censored regression models. For example, estimators considered by Honore (1992) and Honore, Kyriazidou and Powell (2000) require that $\{u_{it2}\}_{t=1}^{T}$ in equation (3) are either i.i.d. or strictly stationary conditional on (z_i, c_{i2}). Importantly, because these estimators use differencing to remove c_{i2}, it is only possible to estimate $c_{i2} + u_{it2}$, which are generally correlated with z_i, so that condition (41) necessarily fails.

Once residuals $\hat{v}_{it2}$ are obtained, the conditional mean of y_{it} for observations with $s_{it} > 0$ can be estimated for each t using the Nadaraya-Watson kernel regression estimator:

$$\hat{g}_{it} \equiv \hat{g}(r_{it}) = \frac{\sum_{j=1}^{N} K\left(\frac{r_{jt}-r_{it}}{h_g}\right) y_{jt}}{\sum_{j=1}^{N} K\left(\frac{r_{jt}-r_{it}}{h_g}\right)}, \quad t = 1, \ldots, T, \tag{47}$$

for kernel function $K(\cdot)$ and bandwidth h_g, such that $h_g \to 0$ and $N h_g^{M+L+1} \to \infty$ as

$N \to \infty$.

The above estimators of v_{it} and g_{it} can be used for obtaining a sample analog of matrix Σ_ω:

$$\begin{aligned} \hat{S} &\equiv \sum_{t=1}^{T} \hat{S}^t, \\ \hat{S}^t &\equiv \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum\nolimits_{i<j} \hat{\omega}_{ijt} \cdot (w_{it} - w_{jt})'(w_{it} - w_{jt}), \\ \hat{\omega}_{ijt} &\equiv \frac{1}{h_\omega^2} \kappa_g \left(\frac{\hat{g}_{it} - \hat{g}_{jt}}{h_\omega} \right) \kappa_v \left(\frac{\hat{v}_{it2} - \hat{v}_{jt2}}{h_\omega} \right) d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, \end{aligned} \tag{48}$$

where $d_{it} = 1[s_{it} > 0]$, τ_{it} and τ_{jt} are trimming terms that are set to zero for observations where $\hat{g}_{it}$ and/or $\hat{v}_{it2}$ are imprecise, and $h_\omega \to 0$, $Nh_\omega^2 \to \infty$ as $N \to \infty$.

Under appropriate regularity conditions, it can be shown that $\hat{S}$ is a consistent estimator of Σ_0, which is matrix Σ_ω that uses a particular weighting function,

$$\omega_{ijt} = (f_{it} f_{jt})^{1/2} \cdot d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt} = f_{it} \cdot d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, \quad t = 1, \ldots, T, \tag{49}$$

where $f_{it} \equiv f(g_{it}, v_{it2})$ is the conditional joint density of g_{it} and v_{it2} for a given t.

Let ζ denote the eigenvalue of $\hat{S}$ that is closest to zero. Then $\hat{\theta}$ is the eigenvector that corresponds to eigenvalue ζ and can be obtained by solving

$$(\hat{S} - \zeta \mathrm{I}_{M+L})\hat{\theta} = 0, \tag{50}$$

where I_{M+L} is the identity matrix of dimension $M + L$.

Because any multiple of the true parameter vector θ will satisfy equation (50), it is convenient to set the first parameter in θ to unity, so that $\theta = (1, \alpha')'$. Correspondingly,

matrix $\hat{S}$ can be partitioned as

$$\hat{S} = \begin{bmatrix} \hat{S}_{11} & \hat{S}_{12} \\ \hat{S}_{21} & \hat{S}_{22} \end{bmatrix}. \tag{51}$$

Then, using the normalization mentioned above and solving (50) for α gives

$$\hat{\alpha} = -[\hat{S}_{22} - \zeta \mathrm{I}_{M+L-1}]^{-1} \hat{S}_{21}. \tag{52}$$

It can be shown that $\hat{\theta} = (1, \hat{\alpha}')'$ is consistent for θ and $\sqrt{N}$-asymptotically normal. The formal consistency argument and derivation of the asymptotic variance are provided in the Appendix. As an alternative to using the analytical formula, the asymptotic variance of $\hat{\alpha}$ can be estimated using panel bootstrap.

Several points are worth mentioning. In (45), and therefore in (48), matching is performed within a given period, so that time-specific shocks that are common to all cross-section units are permitted (although the time-specific intercept cannot be estimated). Also, errors may be arbitrarily serially related. An alternative approach would be to match observations for the same cross-section unit i in any two periods, t and s, where $g(r_{it}) = g(r_{is})$ and $v_{it2} = v_{is2}$, as was proposed by Kyriazidou (1997) in application to linear panel data models with selection. Such an approach would be robust to an arbitrary form of dependence between exogenous variables and unobserved effect. However, an important disadvantage of such method is that it requires a strong form of stationarity and implies that there are no common time-specific shocks to y_{it}, which rarely holds in practice. Moreover, for observations where $g(r_{it})$ and $g(r_{is})$ are similar, it would often be the case that w_{it} and w_{is} would also be similar, which would cause identification problems, especially in short panels.

A general shortcoming of a semiparametric approach is that it does not permit estimating average partial effects. Because v_{it2} is not known for the part of the population

with $s_{it} = 0$, it is not possible to "integrate out" v_{it2} across its entire distribution. Therefore, the ASF and APEs cannot be estimated. In fact, it appears that in the sample selection context, partial effects can be identified only for parametric models. However, the semiparametric approach can be used to estimate relative effects of continuous variables. Specifically, for continuous explanatory variables

$$\frac{\text{APE}_k}{\text{APE}_j} = \frac{\beta_k}{\beta_j},$$

and we have consistent estimators for the β_j up to a common scale factor. Unfortunately, relative effects of discrete variables cannot be estimated.

5 Monte Carlo Simulations

This section presents results from limited Monte Carlo experiments that have been conducted to examine the finite-sample properties of proposed estimators. The focus is on the censored selection variable case where both parametric and semiparametric methods apply. For the same reason we do not simulate average partial effects. Because parameters can only be estimated up to scale, relative effects are reported.

Data are generated using equation (6), with $x_{it} = (x_{1it}, x_{2it})$ and $z_{it} = (1, x_{1it}, x_{2it}, x_{3it})$. Model parameters are set at $\beta = (1, 0.6)'$, $\delta = (1, 0.5, 0.8, 1.2)'$, $\xi_1 = (-0.3, -0.3, -0.3)'$, $\xi_2 = (0.3, 0.3, 0.3)'$. Unobserved effects, a_{i1} and a_{i2}, are independent across i and distributed as $Normal(0, \sigma_a^2)$ with $\text{Corr}(a_{i1}, a_{i2}) = 0.25$. Idiosyncratic errors, u_{it1} and u_{it2}, are independent across i and t and distributed as $Normal(0, \sigma_u^2)$; ρ is either 0.5 or 0. The total variance of the composite errors is set to unity, whereas σ_a^2/σ_u^2 is either 0.3 or 0.

Exogenous variables are generated according to the model:

$$x_{itj} = b_{ij} + \epsilon_{itj}, \quad j = 1, 2, 3, \tag{53}$$

where b_{ij} are independent across i and distributed as $Normal(0, \sigma_b^2)$; ϵ_{ij} are independent across i and t and distributed as $Normal(0, \sigma_\epsilon^2)$; $\sigma_b^2 + \sigma_\epsilon^2 = 1$ with $\sigma_b^2/\sigma_\epsilon^2 = 0.3$; $\text{Corr}(b_{ij}, b_{ih}) = \text{Corr}(b_{ij}, a_{ik}) = 0.25$ for $j = 1, 2, 3$, $h \neq j$, $k = 1, 2$. The employed data generating process results in about 33% of the sample having missing values for y_{it} in a given t.

In the semiparametric estimation, the cross-validation criterion was used when selecting the optimal bandwidth for the conditional expectation function g_{it} and weighting (joint density) function ω_{ijt} (see Li and Racine, 2007, for example). We follow the common practice and set trimming terms equal to one for all observations.

In addition to comparing performance of the parametric and semiparametric estimators discussed in sections 3 and 4, we consider two commonly used parametric methods that do not account for selection. Specifically, model (1) is estimated by probit, so that both selection and unobserved heterogeneity are ignored. We also report results obtained from a probit regression that includes the time means of exogenous variables, but does not account for selection. Simulations were performed for $N = 500$ and 1000, $T = 3$, using 1000 replications.

Results for the estimated relative effect, $\hat{\beta}_2/\hat{\beta}_1$, are reported in Table 1. Under no unobserved heterogeneity and random selection ($\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$), the computed bias is small for all estimators, while the root mean squared error is the smallest for the usual probit estimator. When adding the correlated unobserved effect ($\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$), the bias in the probit estimator noticeably increases, while there are only minor changes in the biases of the other four estimators; the root mean squared errors increase for all estimators. Finally, when both the correlated unobserved heterogeneity and nonrandom selection are present, the standard probit estimator has the largest computed bias, and the probit estimator of a model that includes time means has the second largest bias.

Both parametric estimators that implement selection correction (two-step and full

Table 1: Simulation results for $\hat{\beta}_2/\hat{\beta}_1$ ($\beta_2/\beta_1 = 0.6$), $u_{it1} \sim Normal(0, \sigma_u^2)$

		No correction Probit (1)	No correction Probit, time means (2)	Censored selection two-step MLE (3)	Binary selection full MLE (4)	Censored selection Semiparametric (5)
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=500	Bias	0.0022	0.0025	0.0012	0.0017	-0.0045
	RMSE	0.0571	0.0679	0.0722	0.0706	0.0921
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=500	Bias	-0.0342	-0.0005	0.0022	0.0025	-0.0011
	RMSE	0.0737	0.0721	0.0764	0.0736	0.1144
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=500	Bias	-0.0736	-0.0338	0.0014	0.0011	-0.0156
	RMSE	0.0990	0.0812	0.0752	0.0732	0.1362
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=1000	Bias	-0.0002	0.0006	0.0009	0.0014	-0.0081
	RMSE	0.0394	0.0475	0.0500	0.0488	0.0564
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=1000	Bias	-0.0375	-0.0027	0.0012	0.0018	-0.0099
	RMSE	0.0587	0.0545	0.0569	0.0550	0.0717
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=1000	Bias	-0.0733	-0.0335	0.0027	0.0024	-0.0300
	RMSE	0.0863	0.0631	0.0547	0.0531	0.0840

MLE) have small biases under all scenarios. However, the computed bias of the semiparametric estimator is higher than that of the parametric correction methods when $\rho = 0.5$. Increasing the sample size from 500 to 1000 decreases root mean squared errors for all estimators, but does not necessarily help to reduce the bias. Perhaps not surprisingly, the mean squared errors are larger for the semiparametric estimator. However, if compared to the other correction procedures, the precision of the semiparametric estimator improves more substantially when the sample size grows.

To check the properties of the estimators when the error distribution is not normal, we consider an alternative specification, where u_{it1} has chi-square distribution with three degrees of freedom. The distribution was transformed to have zero mean and variance equal to $\text{Var}(u_{it1})$ in the normal distribution case. Results from that specification are presented in Table 2.

As seen in Table 2, results do not change much. Similar to the case where $u_{it1} \sim$

Table 2: Simulation results for $\hat{\beta}_2/\hat{\beta}_1$ ($\beta_2/\beta_1 = 0.6$), u_{it1} has chi-square distribution

		No correction Probit (1)	No correction Probit, time means (2)	Censored selection two-step MLE (3)	Binary selection full MLE (4)	Censored selection Semiparametric (5)
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=500	Bias	0.0005	0.0004	-0.0011	0.0036	0.0002
	RMSE	0.0464	0.0578	0.0606	0.0589	0.0924
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=500	Bias	-0.0305	-0.0006	-0.0014	0.0090	-0.0051
	RMSE	0.0694	0.0696	0.0721	0.0706	0.1081
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=500	Bias	-0.0687	-0.0320	0.0011	0.0099	-0.0224
	RMSE	0.0939	0.0777	0.0730	0.0718	0.1325
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=1000	Bias	0.0012	0.0017	0.0009	0.0054	-0.0023
	RMSE	0.0319	0.0398	0.0417	0.0409	0.0536
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=1000	Bias	-0.0319	-0.0022	-0.0021	0.0068	-0.0065
	RMSE	0.0534	0.0473	0.0501	0.0485	0.0630
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=1000	Bias	-0.0692	-0.0344	-0.0011	0.0074	-0.0253
	RMSE	0.0825	0.0598	0.0511	0.0492	0.0788

$Normal(0, \sigma_u^2)$, both the usual probit estimator and probit estimator of an augmented equation that includes time means have sizable biases when $\rho = 0.5$. The two-step and full MLE estimators that account for nonrandom selection perform well under all scenarios: both have only small biases and small root mean squared errors. The semiparametric estimator tends to have larger biases and root mean squared errors than parametric correction methods, but performs better than the estimators that ignore non-random selection. Similar to the trends observed in Table 1, mean squared errors of all estimators decrease when N increases.

6 Conclusion

This paper considers estimation of binary-response panel data models in the presence of non-random sample selection and self-selection. Parametric estimators proposed in the paper can be used when the selection variable is either censored or binary. The discussed

approach permits estimating both coefficients and partial effects, as well as treatment effects. The considered parametric methods are simple in implementation and perform well in simulations even when the underlying distributional assumptions do not hold. Moreover, we discuss tests that provide a simple way of detecting a selection bias.

The paper also proposes a semiparametric estimator that does not impose distributional assumptions, but can only be used when the selection variable is censored. In Monte Carlo experiments, this estimator performs reasonably well, although it is less precise and has a larger computed bias than parametric estimators. The relatively large bias of the semiparametric estimator may be due to our inability to fully optimize the choice of bandwidths. In simulations, the optimal bandwidths were selected for the two nonparametric components (conditional expectation function, g_{it}, and weighting function, ω_{ijt}) separately. Future research could focus on the choice of the optimal bandwidths jointly.

Appendix

In this appendix we discuss asymptotic properties of the semiparametric estimator proposed in Section 4. The argument below is very similar to the one in Blundell and Powell (2004).

To demonstrate the consistency of the semiparametric estimator, first show that $\hat{S}^t$ is consistent for Σ_0^t, $t = 1, ..., T$, where Σ_0^t is a particular form of matrix Σ_ω^t that uses the weighting matrix specified in equation (49). Using the first-order mean-value expansion,

for each t we can write:

$$\begin{aligned}
\hat{S}^t &= S_0^t + S_1^t, \quad \text{where} && (54)\\
S_l^t &\equiv \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum\nolimits_{i<j} \omega_{ijt}^l (w_{it} - w_{jt})'(w_{it} - w_{jt}),\ l = 0, 1, && (55)\\
\omega_{ijt}^0 &\equiv \frac{1}{h_\omega^2} \kappa_g \left(\frac{g_{it} - g_{jt}}{h_\omega} \right) \kappa_v \left(\frac{v_{it2} - v_{jt2}}{h_\omega} \right) d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, && (56)\\
\omega_{ijt}^1 &\equiv \frac{1}{h_\omega^3} \left\{ \kappa_g^{(1)} \left(\frac{g_{ijt}^*}{h_\omega} \right) \kappa_v \left(\frac{v_{ijt2}^*}{h_\omega} \right) (\hat{g}_{it} - g_{it} - \hat{g}_{jt} + g_{jt}) \right.\\
&- \kappa_g^{(1)} \left(\frac{g_{ijt}^*}{h_\omega} \right) \kappa_v \left(\frac{v_{ijt2}^*}{h_\omega} \right) g_v^{(1)}(w_{it}, v_{it2}^*) \cdot q_{it}(\hat{\pi}_t - \pi_t)\\
&+ \kappa_g^{(1)} \left(\frac{g_{ijt}^*}{h_\omega} \right) \kappa_v \left(\frac{v_{ijt2}^*}{h_\omega} \right) g_v^{(1)}(w_{jt}, v_{it2}^*) \cdot q_{jt}(\hat{\pi}_t - \pi_t)\\
&- \left. \kappa_g \left(\frac{g_{ijt}^*}{h_\omega} \right) \kappa_v^{(1)} \left(\frac{v_{ijt2}^*}{h_\omega} \right) (q_{it} - q_{jt})(\hat{\pi}_t - \pi_t) \right\} d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, && (57)
\end{aligned}$$

where $\kappa_g^{(1)}(\cdot)$ and $\kappa_v^{(1)}(\cdot)$ are vectors of first derivatives of functions $\kappa_g(\cdot)$ and $\kappa_v(\cdot)$, respectively, $g_v^{(1)}(\cdot)$ is the first derivative of function $g(\cdot)$ with respect to v_{it2}, $q_{it} = (1, z_{i1}, \ldots, z_{iT})$, $\pi_t = (\eta_{2t}, \xi_{21}, \ldots, \delta_t + \xi_{2t}, \ldots, \xi_{2T})'$, and $\hat{\pi}_t$ is the first-step CLAD estimator of π_t.

Similar to Blundell and Bond (2004), the summand in (**??**) is of order $\frac{1}{h_\omega^2}$ when the first four moments of r_{it} and s_{it} are finite, and $\kappa_g(\cdot)$, $\kappa_v(\cdot)$, τ_{it} are bounded. Therefore, when $h_\omega \to 0$, $h_\omega^2 N \to \infty$, it is true that $\hat{S}^t = \Sigma_0^t + o_p(1)$, $t = 1, ..., T$.

To show that S_1^t converges in probability to zero, $t = 1, \ldots, T$, assume that functions $\kappa_g(\cdot)$, $\kappa_v(\cdot)$, $\kappa_g^{(1)}(\cdot)$, $\kappa_v^{(1)}(\cdot)$, $g_v^{(1)}(\cdot)$ are uniformly bounded, and the first two moments of q_{it} exist. Furthermore, when using Powell's censored least absolute deviations estimator (Powell, 1984) or symmetrically trimmed censored least squares estimator (Powell, 1986) to estimate π, assume that the appropriate regularity conditions hold, and let $h_\omega^6 N \to \infty$ as $N \to \infty$. This ensures that $h_\omega^{-3}(\hat{\pi} - \pi) = o_p(1)$. Moreover, assume that regularity conditions provided in Ahn and Powell (1993) hold. These include smoothness assumptions for conditional expectation and density functions, the use of higher-order kernel functions,

and restrictions on the speed with which h_g and h_ω converge to zero as $N \to \infty$. Then, $h_\omega^{-3}(\hat{g}_{it} - g_{it})$ uniformly converges to zero.

From above, it follows that under the specified conditions,

$$\hat{S} \equiv \sum_{t=1}^{T} \hat{S}^t = \sum_{t=1}^{T} \Sigma_0^t + o_p(1) \equiv \Sigma_0 + o_p(1). \tag{58}$$

Moreover, using the law of iterated expectations:

$$\begin{aligned} \Sigma_0^t &\equiv \mathrm{E}[f_{it} \cdot d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt} \cdot (w_{it} - w_{jt})'(w_{it} - w_{jt}) | g_{it} = g, v_{it2} = v] \\ &= \mathrm{E}\left\{2 f_{it} \cdot (\varrho_{it} \mu_{ww,it} - \mu_{w,it}' \mu_{w,it})\right\}, \quad t = 1, \ldots, T, \end{aligned} \tag{59}$$

$$\begin{aligned} \varrho_{it} &\equiv \mathrm{E}[d_{it} \cdot \tau_{it} | g_{it} = g, v_{it2} = v], \\ \mu_{w,it} &\equiv \mathrm{E}[d_{it} \cdot \tau_{it} \cdot w_{it} | g_{it} = d, v_{it2} = v], \\ \mu_{ww,it} &\equiv \mathrm{E}[d_{it} \cdot \tau_{it} \cdot w_{it}' r_{it} | g_{it} = g, v_{it2} = v]. \end{aligned} \tag{60}$$

Furthermore, $\Sigma_0 \theta = 0$ because

$$\begin{aligned} \sum_{t=1}^{T} [\varrho_{it} \mu_{ww,it} - \mu_{w,it}' \mu_{w,it}] \theta &= \sum_{t=1}^{T} \left[\varrho_{it} \mathrm{E}(w_{it}' w_{it} \theta | g_{it}, v_{it2}) - \mu_{w,it}' \mathrm{E}(w_{it} \theta | g_{it}, v_{it2})\right] \\ &= \sum_{t=1}^{T} (\varrho_{it} \mu_{w,it}' g_{it} - \varrho_{it} \mu_{w,it}' g_{it}) = 0, \end{aligned} \tag{61}$$

where we use the fact that $w_{it}\theta = g_{it}$, $t = 1, \ldots, T$.

Finally, we need to specify the identification condition. Regarding the first-step estimation, necessary identification conditions for the censored least absolute deviations estimator and symmetrically trimmed least squares estimator are provided in Powell (1984) and Powell (1986), respectively. The second part of the identification condition is that in the population, θ is a unique nontrivial solution to $\Sigma_0 \theta = 0$ after the normalization

$\theta = (1, \alpha')'$ is imposed. Specifically, assume that matrix Σ_0^{22}, which is the lower-right $(M + L - 1) \times (M + L - 1)$ sub-matrix of matrix Σ_0, has full rank. This completes the consistency argument.

In order to establish $\sqrt{N}$-asymptotic normality, first use the second order mean value expansion to write

$$\hat{S} = S_0 + S_1 + S_2 \equiv \sum_{t=1}^{T} S_0^t + \sum_{t=1}^{T} S_1^t + \sum_{t=1}^{T} S_2^t, \tag{62}$$

where

$$S_l^t \equiv \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum\nolimits_{i<j} \omega_{ijt}^l (w_{it} - w_{jt})'(w_{it} - w_{jt}),\ l = 0, 1, \tag{63}$$

$$\omega_{ijt}^0 \equiv \frac{1}{h_\omega^2} \kappa_g \left(\frac{g_{it} - g_{jt}}{h_\omega} \right) \kappa_v \left(\frac{v_{it2} - v_{jt2}}{h_\omega} \right) d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, \tag{64}$$

$$\begin{aligned} \omega_{ijt}^1 \equiv & \frac{1}{h_\omega^3} \left\{ \kappa_g^{(1)} \left(\frac{g_{it} - g_{jt}}{h_\omega} \right) \kappa_v \left(\frac{v_{it2} - v_{jt2}}{h_\omega} \right) (\hat{g}_{it} - g_{it} - \hat{g}_{jt} + g_{jt}) \right. \\ & - \kappa_g^{(1)} \left(\frac{g_{it} - g_{jt}}{h_\omega} \right) \kappa_v \left(\frac{v_{it2} - v_{jt2}}{h_\omega} \right) g_v^{(1)}(r_{it}) \cdot q_{it}(\hat{\pi}_t - \pi_t) \\ & + \kappa_g^{(1)} \left(\frac{g_{it} - g_{jt}}{h_\omega} \right) \kappa_v \left(\frac{v_{it2} - v_{jt2}}{h_\omega} \right) g_v^{(1)}(r_{jt}) \cdot q_{jt}(\hat{\pi}_t - \pi_t) \\ & \left. - \kappa_g \left(\frac{g_{it} - g_{jt}}{h_\omega} \right) \kappa_v^{(1)} \left(\frac{v_{it2} - v_{jt2}}{h_\omega} \right) (q_{it} - q_{jt})(\hat{\pi}_t - \pi_t) \right\} d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, \end{aligned} \tag{65}$$

$$
\begin{aligned}
\omega_{ijt}^2 \equiv\ & \frac{1}{2h_\omega^4}\left\{\kappa_g^{(2)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v\left(\frac{v_{ijt2}^*}{h_\omega}\right)(\hat{g}_{it}-g_{it}-\hat{g}_{jt}+g_{jt})^2\right. \\
-\ & 2\kappa_g^{(2)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v\left(\frac{v_{ijt2}^*}{h_\omega}\right)g_v^{(1)}(r_{it}^*)\cdot q_{it}(\hat{\pi}_t-\pi_t)(\hat{g}_{it}-g_{it}) \\
+\ & 2\kappa_g^{(2)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v\left(\frac{v_{ijt2}^*}{h_\omega}\right)g_v^{(1)}(r_{jt}^*)\cdot q_{jt}(\hat{\pi}_t-\pi_t)(\hat{g}_{jt}-g_{jt}) \\
-\ & 2\kappa_g^{(1)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v^{(1)}\left(\frac{v_{ijt2}^*}{h_\omega}\right)\cdot(q_{it}-q_{jt})(\hat{\pi}_t-\pi_t)(\hat{g}_{it}-g_{it}-\hat{g}_{jt}+g_{jt}) \\
+\ & 2\kappa_g^{(1)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v^{(1)}\left(\frac{v_{ijt2}^*}{h_\omega}\right)\left[g_v^{(1)}(r_{it}^*)q_{it}-g_v^{(1)}(r_{jt}^*)q_{jt}\right](\hat{\pi}_t-\pi_t)(\hat{\pi}_t-\pi_t)'(q_{it}-q_{jt})' \\
+\ & h_w\kappa_g^{(1)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v\left(\frac{v_{ijt2}^*}{h_\omega}\right)\left[g_v^{(2)}(w_{it}^*)q_{it}-g_v^{(2)}(w_{jt}^*)q_{jt}\right](\hat{\pi}_t-\pi_t)(\hat{\pi}_t-\pi_t)'(q_{it}-q_{jt})' \\
+\ & \kappa_g^{(2)}\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v\left(\frac{v_{ijt2}^*}{h_\omega}\right)\left(\left[g_v^{(1)}(w_{it}^*)\right]^2 q_{it}-\left[g_v^{(1)}(w_{jt}^*)\right]^2 q_{jt}\right)(\hat{\pi}_t-\pi_t)(\hat{\pi}_t-\pi_t)'(q_{it}-q_{jt})' \\
+\ & \left.\kappa_g\left(\frac{g_{ijt}^*}{h_\omega}\right)\kappa_v^{(2)}\left(\frac{v_{ijt2}^*}{h_\omega}\right)(q_{it}-q_{jt})(\hat{\pi}_t-\pi_t)(\hat{\pi}_t-\pi_t)'(q_{it}-q_{jt})'\right\}d_{it}\cdot d_{jt}\cdot\tau_{it}\cdot\tau_{jt} \qquad (66)
\end{aligned}
$$

Under assumptions stated in Ahn and Powell (1993), using $\sqrt{N}$-consistency of the first-step estimator $\hat{\pi}$, and following the same argument as in Blundell and Powell (2004), it should be the case that

$$\sqrt{N}S_0\theta = o_p(1), \quad \sqrt{N}S_2\theta = o_p(1). \qquad (67)$$

Furthermore, when the selection equation is estimated using either Powell's censored least absolute deviations estimator or symmetrically trimmed censored least squares estimator, $\hat{\pi}$ satisfies

$$\sqrt{N}(\hat{\pi}-\pi) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_i + o_p(1),$$

where $\mathrm{E}(m_i)=0$, and $\mathrm{E}(m_i m_i')$ exists and is nonsingular.

Then, can show

$$\sqrt{N}\hat{S}\theta = \sqrt{N}S_1\theta + o_p(1) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(e_{i1}+e_{i2}) + o_p(1), \qquad (68)$$

where

$$
\begin{aligned}
e_{i1} &\equiv \sum_{t=1}^{T} 2 f_{it} \varrho_{it} (\varrho_{it} w_{it} - \mu_{w,it})' \cdot \frac{\partial \psi(g_{it}, v_{it2})}{\partial\, g_{it}} \cdot [y_{it} - g(w_{it})], \\
e_{i2} &\equiv -F m_i(\pi), \\
F &\equiv \mathrm{E}\left[\sum_{t=1}^{T} 2 f_{it} \varrho_{it} (\varrho_{it} w_{it} - \mu_{w,it})' \left(\frac{\partial \psi(g_{it}, v_{it2})}{\partial\, g_{it}} \cdot \frac{\partial g_{it}}{\partial\, v_{it2}} + \frac{\partial \psi(g_{it}, v_{it2})}{\partial\, v_{it2}} \right) q_{it} \right]. \quad (69)
\end{aligned}
$$

If the censored least absolute deviations estimator (Powell, 1984) is used as a first-step estimator of π, and estimation is performed separately for each t, then

$$
\begin{aligned}
m_i(\pi) &= \begin{pmatrix} m_{i1}(\pi_1) \\ \dots \\ m_{iT}(\pi_T) \end{pmatrix}, \\
m_{it}(\pi_t) &= [\mathrm{f}_t(0) \cdot J_t]^{-1} \cdot 1[q_{it}\pi_t > 0] \cdot q_{it}' \left(\frac{1}{2} - 1[v_{it2} > 0] \right), \\
J_t &\equiv \mathrm{E}\left[1[q_{it}\pi_t > 0] \cdot q_{it}' q_{it} \right], \quad t = 1, \dots, T, \quad (70)
\end{aligned}
$$

where $\mathrm{f}_t(\cdot)$ is the density function of error v_{it2} in period t.

If π_t, $t = 1, \dots, T$, is estimated using the symmetrically trimmed lease squares estimator (Powell, 1986), then

$$
\begin{aligned}
m_{it}(\pi_t) &= C_t^{-1} \cdot 1[q_{it}\pi_t > 0] \cdot q_{it}' \cdot (\min\{s_{it}, 2q_{it}\pi_t\} - q_{it}\pi_t), \\
C_t &\equiv \mathrm{E}\left\{ 1[-q_{it}\pi_t < v_{it2} < q_{it}\pi_t] \cdot q_{it}' q_{it} \right\}, \quad t = 1, \dots, T. \quad (71)
\end{aligned}
$$

From (61) and (68) it follows that

$$
\sqrt{N}\theta' \hat{S} \theta = o_p(1), \quad (72)
$$

so that for the subvector $\hat{\alpha}$ of $\hat{\theta} = (1, \hat{\alpha}')'$, we obtain

$$\sqrt{N}(\hat{\alpha} - \alpha) \xrightarrow{d} Normal(0, \Sigma_{22}^{-1} V_{22} \Sigma_{22}^{-1}), \tag{73}$$

where Σ_{22} is the lower $(M + L - 1) \times (M + L - 1)$ diagonal submatrix of Σ_0, and V_{22} is the lower $(M + L - 1) \times (M + L - 1)$ diagonal submatrix of V,

$$V \equiv \text{Var}(e_{i1} + e_{i2}) = \text{E}[(e_{i1} + e_{i2})(e_{i1} + e_{i2})']. \tag{74}$$

Note that this is a robust form of the variance that accounts for serial dependence in the errors.

References

Ahn, H., Ichimura, H., and Powell, J. L., 2004, Simple Estimators for Monotone Index Models, manuscript, Department of Economics, U.C. Berkley.

Ahn, H. and Powell, J.L., 1993, Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism. Journal of Econometrics 58, 3-29.

Blundell, R.W. and Powell, J. L. , 2004, Endogeneity in Semiparametric Binary Response Models. Review of Economic Studies 71, 655-679.

Chamberlain, G., 1980, Analysis with Qualitative Data, Review of Economic Studies 47, 225-238.

Chamberlain, G., 2010, Binary Response Models for Panel Data: Identification and Information. Econometrica 78, 159-168.

Honore, B. E., 1992, Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects. Econometrica 60, 533-565.

Honore, B. E., Kyriazidou, E., and Powell, J.L., 2000, Estimation of Tobit-Type Models with Individual Specific Effects. Econometric Reviews 19, 341-366.

Kyriazidou, E., 1997, Estimation of a Panel Data Sample Selection Model. Econometrica 65, 1335-1364.

Lewbel, A. and Linton, O., 2002, Nonparametric Censored and Truncated Regression. Econometrica 70, 765-779.

Li, Q. and Racine, J. S., 2007, Nonparametric Econometrics: Theory and Practice, Princeton and Oxford: Princeton University Press

Mundlak, Y., 1978, On the Pooling of Time Series and Cross Section Data, Econometrica 46, 69-85.

Papke, L.E. and J.M. Wooldridge, 2008, Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates. Journal of Econometrics 145, 121–133.

Powell, J.L., 1984, Least Absolute Deviations Estimation for the Censored Regression Model. Journal of Econometrics 25, 303-325.

Powell, J.L., 1986, Symmetrically Trimmed Least Squares Estimation for Tobit Models. Econometrica 54, 1435-1460.

Rivers, D. and Vuong, Q.H., 1988, Limited Information Estimators and Exogeneity Tests for Simultaneous Probit. Journal of Econometrics 39, 347-366.

Semykina, A. and J.M. Wooldridge, 2010, Estimating Panel Data Models in the Presence of Endogeneity and Selection. Journal of Econometrics 157, 375–380.

Smith, R.J. and Blundell, R.W., 1986, An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply. Econometrica 54, 679-685.

Wooldridge, J.M., 1995, Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions. Journal of Econometrics 68, 115–132.

Wooldridge, J.M., 2010, Econometric Analysis of Cross Section and Panel Data. MIT Press: Cambridge, MA.