2018-03

Binary Response Panel Data Models With Sample Selection And Self-selection

Anastasia Semykina and Jeffrey M. Wooldridge

The publisher's version of record is available at https://doi.org/10.1002/jae.2592

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

April 18, 2017

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, Treatment effect

We thank Georges Bresson, Bo Honoré, 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.

Binary Response Panel Data Models with Sample Selection and Self Selection

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 a worker being employed in a job with benefits with selection occurring due to non-random 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, Wooldridge (1995) and Rochina-Barrachina (1999) propose parametric estimators of the linear panel data model under sample selection when the explanatory variables are strictly exogenous. Kyriazidou (1997) derives a semiparametric estimator for such models. Estimation of linear unobserved effects panel data models with endogenous explanatory variables and nonrandom sample selection was considered, for example, by Charlier, Melenberg, and van Soest (2001) and Semykina and Wooldridge (2010). In this paper, we discuss 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, our approach has similarities with the methodology of Meng and Schmidt (1985), who consider cross

[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.

section binary response models. To account for possible correlation between unobserved heterogeneity and explanatory variables we use the Mundlak (1978) device, which has parallels with the fixed effects estimators of linear models. Similar to the traditional fixed effects, the Mundlak method involves using values of covariates from all time periods. Therefore, the presented estimator is applicable in cases where the observability of the dependent variable is determined by a 'participation' outcome (e.g. labor force participation or decision to purchase a product), while the values of explanatory variables are always observed. We show how to combine the Mundlak device along with pooled maximum likelihood estimation to obtain simple estimators robust to general forms of dynamic misspecification.[2] Moreover, 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. This approach was initially proposed by Smith and Blundell (1986) and Rivers and Vuong (1986) to address endogeneity in the tobit and probit models, respectively. Alternative parametric techniques were developed by Terza, Basu, and Rathouz (2008) and Wooldridge (2014), although they rely on nonstandard parametric assumptions. Blundell and Powell (2004) and Rothe (2009) propose semiparametric estimators of cross section binary response models with endogenous variables.

In the present paper, we build on the methodology of Smith and Blundell (1986) and Rivers and Vuong (1986) to develop a parametric estimator of binary response models when selection variable is censored. 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 method 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.

The finite sample properties of the proposed estimators are examined using limited Monte Carlo experiments. Furthermore, we consider an empirical application, where we investigate the

[2]The consistency of the pooled maximum likelihood estimator relies on the assumption that the current-period likelihood function is correctly specified, which can be achieved even if the dynamics in the outcome variable is not fully captured (see, for example, Wooldridge, 2010, Chapter 13).

determinants of pension coverage among women.

2 General Setup

Consider a binary response model

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

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, $\mathbf{x}_{it}$ is a $1 \times M$ vector of time-varying explanatory variables, c_{i1} is the unobserved effect, and u_{it1} is the idiosyncratic error, which is independent of c_{i1} and x_{it}. In what follows, the observed covariates are assumed to be strictly exogenous conditional on c_{i1}. Specifically, for $\mathbf{x}_i \equiv (\mathbf{x}_{i1},\ \mathbf{x}_{i2},\ \ldots,\ \mathbf{x}_{it})$, assume that $y_{it}|\mathbf{x}_i, c_{i1} \sim y_{it}|\mathbf{x}_{it}, c_{i1}$. This assumption implies that after accounting for the unobserved effect, only the current-period x_{it} affects y_{it}, while past and future values do not matter. Note that this assumption does not impose restrictions on how $\mathbf{x}_i$ and c_{i1} may be related. Going back to the example mentioned in the introduction, y_{it} may be an indicator for whether the work contract includes retirement benefits (zero if no benefits), x_{it} could include the person's age and educational qualifications in period t, while c_{i1} could include innate ability and work attitudes. In the context of this example, the strict exogeneity conditional on c_{i1} means that after accounting for the individual ability and motivation, the probability of having an employer-provided retirement plan in year t only depends on the person's age and education in that same year, while it does not matter how old and well-educated the person was a year (or few years) before or after.

In addition to estimating the vector of parameters, $\boldsymbol{\beta}$, 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 either the population distribution of the unobserved heterogeneity, c_{i1}, or over the joint distribution of the covariates and heterogeneity. The discussion below covers the estimation of both parameters and APEs.

We introduce incidental truncation by modeling the selection process as

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

where $\mathbf{z}_{it} = (\mathbf{x}_{it}, \mathbf{z}_{it2})$ has dimension $1 \times L$ $(L > M)$, s_{it}^* is a latent variable, s_{it} is a selection indicator that equals one when y_{it} is observed and is zero otherwise (e.g. s_{it} may be the employment status in the benefits example), and u_{it2} is the idiosyncratic error that is independent of z_{it} and c_{i2}. The vector of covariates in the selection equation contains $\mathbf{x}_{it}$ and at least one more variable. For instance, marital status can be used in addition to age and education in the employment equation. Similar to equation (1), we assume that $s_{it}|\mathbf{z}_i, c_{i2} \sim s_{it}|\mathbf{z}_{it}, c_{i2}$, where $\mathbf{z}_i \equiv (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{it})$. Moreover, assume $y_{it}|\mathbf{x}_i, \mathbf{z}_i, c_{i1} \sim y_{it}|\mathbf{x}_{it}, c_{i1}$. A key assumption is that $\mathbf{z}_{it}$ 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. For the example mentioned above, s_{it}^* would be work hours, which are observed for a person who selects to be in the labor force. In another example, s_{it}^* could be 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

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

Partial observability of s_{it}^* makes it possible to estimate $\boldsymbol{\beta}$ 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 $\mathbf{z}_i$, leaving it in the error leads to rescaling of parameters,

but relative effects (β_j/β_k) are preserved, as are relative average partial effects of continuous covariates. The problem arises when $\mathbf{z}_i$ is not independent of c_{i1} and c_{i2}. Because independence of $\mathbf{z}_i$ and unobserved heterogeneity is rarely a realistic assumption, we employ the Mundlak (1978) device. Specifically, let

$$\begin{aligned} c_{i1} &= \eta_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}, \\ c_{i2} &= \eta_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + a_{i2}, \end{aligned} \tag{4}$$

where $\bar{\mathbf{z}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}$, a_{i1} and a_{i2} are independent of $\mathbf{z}_i$, $\text{Var}(a_{i1}) = \sigma^2_{a1}$, $\text{Var}(a_{i2}) = \sigma^2_{a2}$. Thus, $\bar{\mathbf{z}}_i$ are individual time-means of explanatory variables, such as the proportion of the time person i was married during the considered period and average income of person i's spouse.

In the context of nonlinear models, the Mundlak device is also known as the correlated random effects approach, which was first introduced by Chamberlain (1980) for binary response models with normally distributed errors. The normality assumption makes the model in (4) particularly attractive because a linear combination of normally distributed random variables also has a normal distribution. However, this model may also be useful for more general error distributions. The computation of $\bar{\mathbf{z}}_i$ requires that z_{it} is observed in all t to ensure that (4) is not distorted by selection. Consequently, the method in not applicable in the case of attrition, where z_{it} is missing together with y_{it}, but it can be used if selection only affects the dependent variable.

Note that using (4) is similar in spirit to the fixed-effects estimation in the case of linear models. Indeed, modeling c_{i1} as in (4) in a linear model, augmenting the main equation by $\bar{\mathbf{x}}_i$ and estimating it by OLS would produce $\hat{\beta}$ that would be identical to fixed-effects estimates (Mundlak, 1978). In binary-response models, (4) is not equivalent to time-demeaning (fixed effects estimation). However, similar to linear models, it estimates β based on the deviations of x_{it} from its individual time mean because $\bar{\mathbf{z}}_i$ (and, hence, $\bar{\mathbf{x}}_i$) is included among controls. In other words, (4) provides a sensible analogy to the fixed effects method that is widely used for estimating linear panel data models. Among empirical applications that have utilized this approach in the past are the studies of the impact of health on wages (Jäckle and Himmler, 2010) and labor market participation (Maurer, Klein and Vella, 2011), as well as the determinants of birthweight (Abrevaya and Dahl, 2008) and student test pass rates (Papke and Wooldridge,

2008).

The discussion of (4) brings up another important point. If the original model includes time-invariant variables (for example, race, geography), they can be included in z_{it}. Naturally, the time means of such variables cannot be included in (4) because of perfect collinearity. This implies that the effects of time-constant variables are not distinguishable from the unobserved effects, and causality cannot be established, unless one assumes that these variables are conditionally independent of c_{i1} and c_{i2}.

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

$$\begin{aligned} y_{it} &= 1[\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + v_{it1} > 0], \\ s_{it} &= 1[\eta_2 + \mathbf{z}_{it}\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2} > 0], \quad t = 1, \ldots, T, \end{aligned} \tag{5}$$

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

$$\begin{aligned} y_{it} &= 1[\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + v_{it1} > 0], \\ s_{it} &= \max\{0, \eta_2 + \mathbf{z}_{it}\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}\}, \quad t = 1, \ldots, T. \end{aligned} \tag{6}$$

By construction, $\mathbf{z}_i$ 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 $\boldsymbol{\beta}$ and the APEs.

Before discussing the different scenarios, it is useful to obtain the APEs based on the equation (1). There are two widely used approaches. In both cases, we can identify the APEs only for the parametric model, and so the discussion is for the case where the heterogeneity has a conditional joint normal distribution.

The "structural" equation that underlies the estimation is

$$\mathrm{P}(y_{it} = 1|\mathbf{x}_{it}, c_{i1}) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_{i1}). \tag{7}$$

One way of measuring the partial effects of the covariates on the response probability is to obtain the average structural function (ASF), introduced by Blundell and Powell (2004). The ASF is a function of the covariates set at fixed values $\mathbf{x}_t$, and the unobserved heterogeneity is averaged out. Therefore, for the probit model,

$$\mathrm{ASF}(\mathbf{x}_t) = \mathrm{E}_{c_{i1}}\left[\Phi(\mathbf{x}_t\boldsymbol{\beta} + c_{i1})\right]. \tag{8}$$

Once we have this function we can plug in interesting values of $\mathbf{x}_t$. Now, we are not directly modeling the distribution of c_{i1}, but rather the conditional distribution $\mathrm{D}(c_{i1}|\mathbf{z}_i)$. Therefore, the following expression based on iterated expectations is useful:

$$\mathrm{ASF}(\mathbf{x}_t) = \mathrm{E}_{\mathbf{z}_i}\left\{\mathrm{E}\left[\Phi(\mathbf{x}_t\boldsymbol{\beta} + c_{i1})|\mathbf{z}_i\right]\right\} = \mathrm{E}_{\mathbf{z}_i}\left[\Phi(\eta_{a1} + \mathbf{x}_t\boldsymbol{\beta}_a + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{a1})\right], \tag{9}$$

where $\boldsymbol{\beta}_a = \boldsymbol{\beta}/\sqrt{1+\sigma_{a_1}^2}$ and similarly for the other parameters with an a subscript. This expression, which hinges on normality of c_{i1} given $\mathbf{z}_i$, is derived in Papke and Wooldridge (2008).

Note that in (9), ASF is obtained by averaging over the distribution of the entire vector $\mathbf{z}_i$. This definition is convenient for estimating average partial effects of continuous variables. In particular, we can take derivatives with respect to the continuous elements of $\mathbf{x}_t$ and then average out the $\mathbf{z}_i$. To obtain a single value, we can further average $\mathrm{ASF}(\mathbf{x}_{it})$ across the distribution of $\mathbf{x}_{it}$. To obtain average partial effects of discrete covariates, one would evaluate ASF at some fixed values, $\mathbf{x}_t$, and consider changes in $\mathrm{ASF}(\mathbf{x}_t)$.

When one desires a single number, a different approach to APEs is usually used. For simplicity, suppose that x_{itj} is a continuous variable, and so we measure its partial effect using the derivative of the response probability. Then, we average the partial effect across the joint distribution of $(\mathbf{x}_{it}, c_{i1})$. This leads to

$$\mathrm{APE}_j = \mathrm{E}_{(\mathbf{x}_{it},c_i)}\left[\frac{\partial \mathrm{P}(y_{it}=1|\mathbf{x}_{it},c_{i1})}{\partial x_{itj}}\right] = \beta_j \mathrm{E}_{(\mathbf{x}_{it},c_i)}\left[\phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_{i1})\right], \tag{10}$$

where $\phi(\cdot)$ is the standard normal pdf. As shown in Nam and Wooldridge (2016) for a general conditional mean function, iterated expectations can be used to obtain APE_j, too:

$$\mathrm{APE}_j = \mathrm{E}_{(\mathbf{x}_{it},\bar{\mathbf{z}}_i)}\left[\beta_{aj}\phi(\eta_{a1} + \mathbf{x}_{it}\boldsymbol{\beta}_a + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{a1})\right]. \tag{11}$$

For a discrete change, the derivative is replaced by differences in the standard normal cdf.

We discuss estimation of the ASF and APEs after discussing estimation of the scaled parameters in Section 3. 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, as seen in equations (9) and (11), it is the scaled parameters that appear in the ASF and APEs, 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.[3] 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 $\boldsymbol{\beta}_a$ 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 (6), 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 (9). 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

$$v_{it1} = \gamma v_{it2} + e_{it1}, \quad t = 1, \ldots, T, \tag{12}$$

where $\gamma = \rho/\sigma$, $\rho = \text{Corr}(v_{it1}, v_{it2})$, and e_{it1} is independent of $\mathbf{z}_i$ and v_{it2} with a normal distribution.[4] Therefore, we can write

$$\begin{aligned} y_{it} &= 1[\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \gamma v_{it2} + e_{it1} > 0], \\ e_{it1}|\mathbf{z}_i, v_{it2} &\sim Normal(0, 1 - \rho^2), \quad t = 1, \ldots, T. \end{aligned} \tag{13}$$

The equations in (13) 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-

[3] See, for example, Wooldridge (2010), Chapter 15, for the discussion of such estimation.

[4] For simplicity, γ is assumed to be time constant. Generally, this assumption can be relaxed. However, this would make conditional probabilities time-specific and would substantially complicate the estimation when the selection variable is binary. In this paper, we focus on a case with $\gamma_t = \gamma$ for all t, which can be easily implemented in practice and, therefore, should be especially useful in empirical research.

ignorable.

A basic but important fact is that because s_{it} is a deterministic function of $\mathbf{z}_i$ and v_{it2}, it follows that

$$e_{it1}|\mathbf{z}_i, v_{it2}, s_{it} \sim Normal(0, 1-\rho^2), \quad t = 1, \ldots, T. \tag{14}$$

Therefore, by including v_{it2} in (13), we can solve the non-random selection problem. This is an example of the "control function" approach proposed by Smith and Blundell (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

$$\mathrm{P}(y_{it} = 1|\mathbf{z}_i, v_{it2}, s_{it}) = \mathrm{P}(y_{it} = 1|\mathbf{z}_i, v_{it2}) = \Phi\left(\frac{\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \gamma v_{it2}}{\sqrt{1-\rho^2}}\right), \tag{15}$$

which is a probit model with parameters rescaled by a common factor $\left(1-\rho^2\right)^{-1/2}$. Thus, if v_{it2} were known, one could estimate $\boldsymbol{\beta}/\sqrt{1-\rho^2}$ rather easily. Of course, v_{it2} is never known; however, in some cases it can be estimated whenever $s_{it} > 0$, and that suffices to consistently estimate the scaled parameters. Because estimation of (14) 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 (14) 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, $\eta_{1\rho} \equiv \frac{\eta_1}{\sqrt{1-\rho^2}}$, $\boldsymbol{\beta}_\rho \equiv \frac{\boldsymbol{\beta}}{\sqrt{1-\rho^2}}$, $\boldsymbol{\xi}_{1\rho} \equiv \frac{\boldsymbol{\xi}_1}{\sqrt{1-\rho^2}}$, and $\gamma_\rho \equiv \frac{\gamma}{\sqrt{1-\rho^2}}$, can be consistently estimated in two steps:

PROCEDURE 3.2.

1. Use pooled Tobit to estimate equation

$$s_{it} = \max\{0, \eta_2 + \mathbf{z}_{it}\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}\}.$$

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

2. For $s_{it} > 0$, estimate (14) by pooled probit, where use $\hat{v}_{it2}$ in place of v_{it2}.[5]

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.[6] We provide the formulas in the online Supplement. 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 $\eta_{1\rho}$, $\boldsymbol{\beta}_\rho$, $\boldsymbol{\xi}_{1\rho}$, and γ_ρ, rather than original parameters in the population model. The estimators of the original parameters can be obtained as

$$\hat{\rho} = \hat{\gamma}_\rho\hat{\sigma} \cdot (1 + \hat{\gamma}_\rho^2\hat{\sigma}^2)^{-1/2}, \quad \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\rho(1 - \hat{\rho}^2)^{-1/2} = \hat{\boldsymbol{\beta}}_\rho(1 + \hat{\gamma}_\rho^2\hat{\sigma}^2)^{1/2}, \tag{16}$$

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 $\hat{\boldsymbol{\beta}}$ in (16), with similar expressions for $\hat{\eta}_1$ and $\hat{\boldsymbol{\xi}}_1$, the APEs are easily obtained. Under the Mundlak assumption, to obtian a single APE – that is, not as a function of specific values of $\mathbf{x}_t$ – we can average derivatives across $(\mathbf{x}_{it}, \bar{\mathbf{z}}_i)$. For a continuous variable x_{itj},

$$\text{APE}_j = \text{E}_{(\mathbf{x}_{it}, \bar{\mathbf{z}}_i)}\left[\phi(\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1)\right]\beta_j. \tag{17}$$

[5] Alternatively, the second-step probit could be replaced with an MLE that would use $\hat{v}_{it2}/\hat{\sigma}$ as a covariate and incorporate ρ in the likelihood function. In that case, the original non-scaled parameters could be estimated. Such a procedure is typically not preprogrammed in the exiting statistical software and would require manual programming of the likelihood function.

[6] See, for example, Murphy and Topel (1985) and Wooldridge (2010) for a detailed discussion of inference in parametric two-step models.

Replacing the expectation with a sample average, and inserting the estimates, an APE that also averages across t is consistently estimated as

$$\widehat{\text{APE}}_j = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi(\hat{\eta}_1 + \mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_1)\right]\hat{\beta}_j. \tag{18}$$

This average partial effect is typically computed as the default in commonly used econometric software packages such as Stata.

Although APEs in (17) are usually of primary interest, one may also be interested in evaluating APEs at particular values of explanatory variables ($\mathbf{x}_t$). For example, one may be interested in knowing partial effects for an "average person" in the sample, or for the median person, or a young well-educated person. In such cases we can use the derivative of the ASF with $\mathbf{x}_t$ plugged in:

$$\widehat{\text{APE}}_j(\mathbf{x}_t) = \left[N^{-1}\sum_{i=1}^{N}\phi(\hat{\eta}_1 + \mathbf{x}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_1)\right]\hat{\beta}_j. \tag{19}$$

Similarly, 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:

$$(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\Phi(\hat{\eta}_1 + \mathbf{x}_{it}^{(1)}\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_1) - \Phi(\hat{\eta}_1 + \mathbf{x}_{it}^{(0)}\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_1)\right] \tag{20}$$

$$\mathbf{x}_{it}^{(0)} \equiv (x_{it1}, \ldots, x_{it,m-1}, x_{tm}^{(0)}, x_{it,m+1}, \ldots, x_{itM}),$$

$$\mathbf{x}_{it}^{(1)} \equiv (x_{it1}, \ldots, x_{it,m-1}, x_{tm}^{(1)}, x_{it,m+1}, \ldots, x_{itM}).$$

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

Standard errors of $\hat{\boldsymbol{\beta}}$ 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 $\mathbf{z}_i$ for each t, and employing the partial maximum likelihood estimator (partial MLE). Specifically, for each t, the joint density function is

$$f(y_{it}, s_{it}|\mathbf{z}_i) = \left\{[\Phi(r_{it})]^{y_{it}}[1-\Phi(r_{it})]^{1-y_{it}}\frac{1}{\sigma}\phi\left(\frac{s_{it}-q_{it}}{\sigma}\right)\right\}^{1[s_{it}>0]}\left\{1-\Phi\left(\frac{q_{it}}{\sigma}\right)\right\}^{1[s_{it}=0]}, \tag{21}$$

where

$$r_{it} \equiv \frac{\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \frac{\rho}{\sigma}(s_{it} - q_{it})}{\sqrt{1-\rho^2}}, \tag{22}$$

$$q_{it} \equiv \eta_2 + \mathbf{z}_{it}\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2. \tag{23}$$

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, $f(y_{i1}, \ldots, y_{iT}, s_{i1}, \ldots, s_{iT}|\mathbf{z}_i)$, which would be very complicated because of the serial dependence in errors. Within the partial MLE framework, it is sufficient to specify $f(y_{it}, s_{it}|\mathbf{z}_i)$, $t = 1, \ldots, T$; however, the estimator of the asymptotic variance should be made robust to serial correlation in the score functions.[7] 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 (14) 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 $\mathbf{z}_i$ 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 $\mathbf{z}_i$ is

$$\begin{aligned} f(y_{it}, s_{it}|\mathbf{z}_i) &= [P(y_{it}=1|s_{it}=1,\mathbf{z}_i)P(s_{it}=1|\mathbf{z}_i)]^{y_{it}s_{it}} \\ &\times [P(y_{it}=0|s_{it}=1,\mathbf{z}_i)P(s_{it}=1|\mathbf{z}_i)]^{(1-y_{it})s_{it}}[P(s_{it}=0|\mathbf{z}_i)]^{(1-s_{it})}, \end{aligned} \tag{24}$$

[7] For example, in Stata this can be done by using the 'cluster' option.

where

$$P(y_{it}=1|s_{it}=1,\mathbf{z}_i) = \mathrm{E}[\Phi(r_{it})|s_{it}=1,\mathbf{z}_i] = \frac{1}{\Phi(q_{it})}\int_{-\infty}^{q_{it}}\Phi(r_{it})\phi(\nu)d\nu, \quad (25)$$

$$P(y_{it}=0|s_{it}=1,\mathbf{z}_i) = \frac{1}{\Phi(q_{it})}\int_{-\infty}^{q_{it}}[1-\Phi(r_{it})]\phi(\nu)d\nu, \quad (26)$$

$$P(s_{it}=1|\mathbf{z}_i) = \Phi(q_{it}), \quad (27)$$

$$P(s_{it}=0|\mathbf{z}_i) = 1-\Phi(q_{it}), \quad (28)$$

where $r_{it} = (\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \rho\nu)\left(1-\rho^2\right)^{-1/2}$ and q_{it} is defined in (23). Thus, the conditional joint likelihood function for unit i in period t is given by

$$\begin{aligned} L_{it} \equiv f(y_{it},s_{it}|\mathbf{z}_i) &= \left[\int_{-\infty}^{q_{it}}\Phi(r_{it})\phi(\nu)d\nu\right]^{y_{it}s_{it}} \\ &\times \left[\int_{-\infty}^{q_{it}}[1-\Phi(r_{it})]\phi(\nu)d\nu\right]^{(1-y_{it})s_{it}}[1-\Phi(q_{it})]^{(1-s_{it})}. \end{aligned} \quad (29)$$

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.[8]

Note that equation (9) still holds. Thus, the estimation of APEs discussed in Section 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 $(\mathbf{x}_{it}, \bar{\mathbf{z}}_i)$. 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 constant variance assumption is violated.

The joint MLE discussed in this Section can also be used in cases where y_{it} is always observed, and s_{it} is a binary treatment indicator, so that there is no sample selection problem, but a usual self-selection into the treatment is present. Estimation of the treatment effects is discussed in more detail in the Appendix.

[8]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.

3.4 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 selection models, when parameters are estimated by joint partial MLE, computational problems may arise.[9] Therefore, it is useful to have a simple test for selection bias, which would help to identify cases when correction is necessary.

When the selection variable is censored, a simple test for selection bias can be performed by testing $\mathrm{H}_0 : \gamma = 0$ after estimating equation (13) using the two-step procedure outlined in Section 3.2. The test is similar to those derived by Smith and Blundell (1986) and Vella (1992). 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, one can consider the correlation between the generalized residuals in the main and selection equations to develop a test similar to the one in Vella (1992). Here, we use our knowledge of the error distribution to derive the Lagrange multiplier (score) test.

Let $\boldsymbol{\theta} = (\eta_1, \boldsymbol{\beta}', \boldsymbol{\xi}_1')'$ and $\mathbf{w}_{it} = (1, \mathbf{x}_{it}, \bar{\mathbf{z}}_i)$. Let $\tilde{r}_{it}$ be r_{it} evaluated at $\rho = 0$ and parameter estimates $\tilde{\boldsymbol{\theta}}$, which are obtained from the restricted model. The restricted model is simply a Chamberlain pooled probit estimator that uses the unbalanced panel. Let $\hat{q}_{it}$ be q_{it} evaluated at the time t probit estimates, $(\hat{\eta}_{2t}, \hat{\boldsymbol{\delta}}_t, \hat{\boldsymbol{\xi}}_{2t})$, where q_{it} is given in (23). Using the likelihood function in equation (29) as a starting point, the Lagrange multiplier (LM) statistic for testing $\mathrm{H}_0 : \rho = 0$ is given by[10]

$$LM = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{\mathbf{S}}_{it,\rho} \right)' \tilde{\mathbf{A}}^{22} \left[\tilde{\mathbf{V}}_{22} \right]^{-1} \tilde{\mathbf{A}}^{22} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{\mathbf{S}}_{it,\rho} \right) / N, \tag{30}$$

[9] For example, an optimization routine may fail to identify a unique maximum if the likelihood function is flat.

[10] See, for example, Wooldridge (2010), Section 12.6.2 for the detailed derivation of equation (30).

where

$$\tilde{\mathbf{S}}_{it,\rho} \equiv \frac{\partial \ln L_{it}}{\partial \rho}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}},\rho=0} = s_{it}\frac{y_{it}-\Phi(\tilde{r}_{it})}{\Phi(\tilde{r}_{it})[1-\Phi(\tilde{r}_{it})]}\phi(\tilde{r}_{it})\hat{\lambda}_{it}, \tag{31}$$

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

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

$$\tilde{\mathbf{A}}^{-1} = \begin{pmatrix} \tilde{\mathbf{A}}^{11} & \tilde{\mathbf{A}}^{12} \\ \tilde{\mathbf{A}}^{21} & \tilde{\mathbf{A}}^{22} \end{pmatrix}^{-1}, \tag{32}$$

$$\tilde{\mathbf{V}} = \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{A}}^{-1} = \begin{pmatrix} \tilde{\mathbf{V}}_{11} & \tilde{\mathbf{V}}_{12} \\ \tilde{\mathbf{V}}_{21} & \tilde{\mathbf{V}}_{22} \end{pmatrix}, \tag{33}$$

$$\tilde{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\tilde{\mathbf{S}}_{it}\sum_{t=1}^{T}\tilde{\mathbf{S}}'_{it}\right), \tag{34}$$

$$\tilde{\mathbf{S}}_{it} \equiv \left(\frac{\partial \ln L_{it}}{\partial\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}},\rho=0}\ ,\ \frac{\partial \ln L_{it}}{\partial\rho}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}},\rho=0}\right)', \tag{35}$$

$$\hat{\lambda}_{it} \equiv \frac{\phi(\hat{q}_{it})}{\Phi(\hat{q}_{it})}. \tag{36}$$

Matrix $\tilde{\mathbf{A}}$ 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:

PROCEDURE 3.4.

(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 the statistical significance of $\hat{\lambda}_{it}$.

Under the null hypothesis the coefficient on $\hat{\lambda}_{it}$ is zero, and so the estimation of the param-

eters 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

$$L_{it} = s_{it}\Phi(\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \gamma\lambda_{it})^{y_{it}}[1 - \Phi(\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \gamma\lambda_{it})]^{(1-y_{it})}, \tag{37}$$

where, to simplify the 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 (37), the score vector is

$$\mathbf{S}_{it} = s_{it}\frac{y_{it} - \Phi(\mathbf{w}_{it}\boldsymbol{\theta} + \gamma\lambda_{it})}{\Phi(\mathbf{w}_{it}\boldsymbol{\theta} + \gamma\lambda_{it})[1 - \Phi(\mathbf{w}_{it}\boldsymbol{\theta} + \gamma\lambda_{it})]}\phi(\mathbf{w}_{it}\boldsymbol{\theta} + \gamma\lambda_{it})\begin{pmatrix}\mathbf{w}_{it}\\ \lambda_{it}\end{pmatrix}. \tag{38}$$

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

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{S}}_{it} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{S}_{it} - \mathbf{A}\sqrt{N}\begin{pmatrix}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\\ \hat{\gamma} - \gamma\end{pmatrix} + o_p(1), \tag{39}$$

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

From (39), it follows that

$$\sqrt{N}\begin{pmatrix}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\\ \hat{\gamma} - \gamma\end{pmatrix} = -\mathbf{A}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\mathbf{S}}_{it} - \mathbf{S}_{it}) + o_p(1) = \mathbf{A}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{S}_{it} + o_p(1). \tag{40}$$

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

$$W = (\hat{\gamma} - \gamma)'(\hat{\mathbf{V}}_{22}/N)^{-1}(\hat{\gamma} - \gamma) = \sqrt{N}(\hat{\gamma} - \gamma)'\hat{\mathbf{V}}_{22}^{-1}\sqrt{N}(\hat{\gamma} - \gamma), \tag{41}$$

where

$$\hat{\mathbf{V}} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1} = \begin{pmatrix} \hat{\mathbf{V}}_{11} & \hat{\mathbf{V}}_{12} \\ \hat{\mathbf{V}}_{21} & \hat{\mathbf{V}}_{22} \end{pmatrix}, \tag{42}$$

$$\hat{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{\mathbf{S}}_{it}\sum_{t=1}^{T}\hat{\mathbf{S}}_{it}'\right), \tag{43}$$

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

$$\hat{p}_{it} = \mathbf{w}_{it}\hat{\boldsymbol{\theta}} + \hat{\gamma}\lambda(\hat{q}_{it}), \tag{44}$$

$$\hat{\mathbf{A}}^{-1} \xrightarrow{p} \mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix}^{-1}. \tag{45}$$

From (40), we can also write the Wald statistic as

$$W = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{S}}_{it,\gamma}\right)'\hat{\mathbf{A}}^{22}\hat{\mathbf{V}}_{22}^{-1}\hat{\mathbf{A}}^{22}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{S}}_{it,\gamma}\right)/N, \tag{46}$$

which is asymptotically distributed as χ_1^2. It is easily seen that under the null of no selection bias ($\rho = 0$, $\gamma = 0$), the scores and Hessian matrices used in (30) and (46) are the same when evaluated at true parameter values. Moreover, when the null is true, $\hat{\gamma} \xrightarrow{p} 0$, and $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$ and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converge in distribution. Therefore, $LM - W \xrightarrow{p} 0$, so that the tests are asymptotically equivalent.

It is important to note that for the tests presented in this Section, the rejection of the null indicates the existence of a selection bias only if Assumption 3.3 holds. If this assumption fails, the rejection may indicate that either the correlation between covariates and unobserved effects is not properly captured or error distribution is not normal, or both.

4 Semiparametric Estimation When Selection Variable Is Censored

4.1 The Model and Derivation of the Estimator

In this section, we consider a semiparametric binary dependent variable model with non-random selection. The main outcome, y_{it}, is still defined as in equation (1), but 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 modeled using the Chamberlain's device, and the estimating equations are given in (6). 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:

$$v_{it1}|\mathbf{z}_i, s_{it} \sim v_{it1}|\mathbf{z}_i, v_{it2} \sim v_{it1}|v_{it2}, \quad t = 1, \ldots, T. \tag{47}$$

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 $\mathbf{w}_{it} = (\mathbf{x}_{it}, \bar{\mathbf{z}}_i)$, where we now drop the time effects in defining $\mathbf{w}_{it}$. Now, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\xi}_1')'$. Then, under condition (47) the conditional expectation of y_{it} is given by

$$\mathrm{E}(y_{it}|\mathbf{z}_i, s_{it}) = \mathrm{P}(y_{it} = 1|\mathbf{z}_i, v_{it2}) = \mathrm{P}(-v_{it1} \leq \mathbf{w}_{it}\boldsymbol{\theta}|\mathbf{z}_i, v_{it2}) = F(\mathbf{w}_{it}\boldsymbol{\theta}, v_{it2}). \tag{48}$$

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 $\boldsymbol{\theta}$. 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 $F(\mathbf{w}_{it}\boldsymbol{\theta}, v_{it2})$ is continuous and monotonic in its first argument, it can be inverted with respect to its first argument. Denote the inverse function $\psi(\cdot, v) \equiv F^{-1}(\cdot, v)$. Also, to shorten the notation, define $\mathbf{r}_{it} = (\mathbf{w}_{it}, v_{it2})$ and $g(\mathbf{r}_{it}) \equiv \mathrm{E}(y_{it}|\mathbf{r}_{it}) = F(\mathbf{w}_{it}\boldsymbol{\theta}, v_{it2})$. Then, we can write

$$\psi[g(\mathbf{r}_{it}), v_{it2}] = \mathbf{w}_{it}\boldsymbol{\theta}$$

$$\text{or,} \qquad \psi[g(\mathbf{r}_{it}), v_{it2}] - \mathbf{w}_{it}\boldsymbol{\theta} = 0, \tag{49}$$

where the result holds a.s. Equation (49) implies that for any two observations, i and j, in a given period t, if $\mathrm{E}(y_{it}|\mathbf{r}_{it}) = \mathrm{E}(y_{jt}|\mathbf{r}_{jt})$ and $v_{it2} = v_{jt2}$, it should be the case that $\mathbf{w}_{it}\boldsymbol{\theta} = \mathbf{w}_{jt}\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$, which satisfies the equality of the indices $\mathbf{w}_{it}\boldsymbol{\theta}$ and $\mathbf{w}_{jt}\boldsymbol{\theta}$.

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

$$\mathrm{E}\left[\omega_{ijt} \cdot ((\mathbf{w}_{it} - \mathbf{w}_{jt})\boldsymbol{\theta})^2 | g(\mathbf{r}_{it}) = g(\mathbf{r}_{jt}), v_{it2} = v_{jt2}\right] = 0, \tag{50}$$

so that

$$\sum_{t=1}^{T} \mathrm{E}\left[\omega_{ijt} \cdot ((\mathbf{w}_{it} - \mathbf{w}_{jt})\boldsymbol{\theta})^2 | g(\mathbf{r}_{it}) = g(\mathbf{r}_{jt}), v_{it2} = v_{jt2}\right]$$
$$\equiv \boldsymbol{\theta}'\boldsymbol{\Sigma}_{\omega}\boldsymbol{\theta} = 0, \tag{51}$$

where $\boldsymbol{\Sigma}_{\omega} \equiv \sum_{t=1}^{T} \boldsymbol{\Sigma}_{\omega}^{t}$, $\boldsymbol{\Sigma}_{\omega}^{t} \equiv \mathrm{E}[\omega_{ijt} \cdot (\mathbf{w}_{it} - \mathbf{w}_{jt})'(\mathbf{w}_{it} - \mathbf{w}_{jt}) | g(\mathbf{r}_{it}) = g(\mathbf{r}_{jt}), v_{it2} = v_{jt2}]$. Assuming that in the population $\boldsymbol{\theta}$ is not zero, $\boldsymbol{\Sigma}_{\omega}$ must be singular. Moreover, assuming that $\boldsymbol{\Sigma}_{\omega}$ has rank $(M + L - 1)$, it has only one zero eigenvalue. Then, in the population, $\boldsymbol{\theta}$ is the eigenvector that corresponds to the zero eigenvalue of matrix $\boldsymbol{\Sigma}_{\omega}$.

To summarize, the following assumption was used to derive equation (51) and obtain the semiparametric estimator:

ASSUMPTION 4. (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) $v_{it1}|z_i, s_{it} \sim v_{it1}|v_{it2}$, (v) cumulative distribution function, $F(\mathbf{w}_{it}\boldsymbol{\theta}, v_{it2})$, is continuous and monotonic in its first argument, (vi) matrix $\boldsymbol{\Sigma}_{\omega}$ has rank $(M + L - 1)$, (vii) a consistent estimator of v_{it2} is available.

The rank condition 4(vi) implies that there is variation in $(\mathbf{w}_{it} - \mathbf{w}_{jt})$ conditional on $g(\mathbf{r}_{it}) = g(\mathbf{r}_{jt})$ and $v_{it2} = v_{jt2}$, i.e. there are observations for which the values of explanatory variables are not the same even though their values of $g(\mathbf{r}_{it})$ and v_{it2} match. This means that the intercept cannot be included, and neither can any factors that are common to all i in a given t, such as macroeconomic indicators. Also, there should be no perfect collinearity among "differenced" variables conditional on matching. The assumption ensures that θ is a unique solution to equation (51) in the population.

4.2 The Estimator

If Assumption 4 holds, a consistent estimator of $\boldsymbol{\theta}$ can then be obtained by constructing a sample analog of matrix $\boldsymbol{\Sigma}_{\omega}$ and finding its eigenvalue that is closest to zero. The estimator of $\boldsymbol{\theta}$ 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. Specifically, the estimation can be performed in two steps:

PROCEDURE 4.

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 $\mathbf{\Sigma}_\omega$ 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:

$$s_{it} = \max\{0, \eta_2 + \mathbf{z}_{it}\boldsymbol{\delta} + \mathbf{z}_{i1}\boldsymbol{\xi}_{21} + \cdots + \mathbf{z}_{iT}\boldsymbol{\xi}_{2T} + v_{it2}\}, \quad t = 1, \ldots, T. \tag{52}$$

If error v_{it2} is continuously distributed with median zero, and its density function is positive at zero, then the parameters in equation (52) 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 the 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 $\mathrm{E}(s_{it}|\mathbf{z}_i)$ and $\mathrm{E}\{1[s_{it} > 0]|\mathrm{E}(s_{it}|\mathbf{z}_i)\}$, followed by the 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 (52) 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 $(\mathbf{z}_i, c_{i2})$. 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 $\mathbf{z}_i$, so that condition (47) 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:

$$\hat{g}_{it} \equiv \hat{g}(\mathbf{r}_{it}) = \frac{\sum_{j=1}^{N} K\left(\frac{\mathbf{r}_{jt}-\mathbf{r}_{it}}{h_g}\right) y_{jt}}{\sum_{j=1}^{N} K\left(\frac{\mathbf{r}_{jt}-\mathbf{r}_{it}}{h_g}\right)}, \quad t = 1, \ldots, T, \tag{53}$$

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 $\mathbf{\Sigma}_\omega$:

$$\begin{aligned} \hat{\mathbf{S}} &\equiv \sum_{t=1}^{T} \hat{\mathbf{S}}^t, \qquad (54) \\ \hat{\mathbf{S}}^t &\equiv \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum\nolimits_{i<j} \hat{\omega}_{ijt} \cdot (\mathbf{w}_{it} - \mathbf{w}_{jt})'(\mathbf{w}_{it} - \mathbf{w}_{jt}), \\ \hat{\omega}_{ijt} &\equiv \frac{1}{h_\omega^2} \kappa_g\left(\frac{\hat{g}_{it} - \hat{g}_{jt}}{h_\omega}\right) \kappa_v\left(\frac{\hat{v}_{it2} - \hat{v}_{jt2}}{h_\omega}\right) d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, \end{aligned}$$

where $d_{it} = 1[s_{it} > 0]$, τ_{it} and τ_{jt} are trimming terms that are set to zero for observations where $\hat{g}_{it}$ is imprecise, and $h_\omega \to 0$, $N h_\omega^2 \to \infty$ as $N \to \infty$. Trimming helps to avoid convergence problems that arise when the density of $\mathbf{r}_{it}$ (and, hence, the denominator in $\hat{g}_{it}$) is close to zero. Unfortunately, there is no theory that would guide the choice of the trimming terms in practice. Therefore, it is rather common to set $\tau_{it} = 1$, $\forall$ i, t. Generally, it is recommended that the results obtained using any kind of trimming are compared to those without trimming to evaluate robustness.

Under appropriate regularity conditions, it can be shown that $\hat{\mathbf{S}}$ is a consistent estimator of

$\mathbf{\Sigma}_0$, which is matrix $\mathbf{\Sigma}_\omega$ that uses a particular weighting function,

$$\omega_{ijt} = (f_{it}f_{jt})^{1/2} \cdot d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt} = f_{it} \cdot d_{it} \cdot d_{jt} \cdot \tau_{it} \cdot \tau_{jt}, \quad t = 1, \ldots, T, \tag{55}$$

where $f_{it} \equiv f(g_{it}, v_{it2})$ is the conditional joint density of g_{it} and v_{it2} for a given t. The regularity conditions include the requirement that function g_{it} has a continuous distribution, and the conditional expectation functions of explanatory variables and conditional density function f_{it} are continuously differentiable. Both first-step and second-step kernels are assumed to be higher-order (bias-reducing) kernels, while the bandwidths h_g and h_w are assumed to converge to zero at specific rates as $N \to \infty$, which helps to reduce the bias and achieve consistency.

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

$$(\hat{\mathbf{S}} - \zeta \mathbf{I}_{M+L})\hat{\boldsymbol{\theta}} = 0, \tag{56}$$

where $\mathbf{I}_{M+L}$ is the identity matrix of dimension $M + L$.

Because any multiple of the true parameter vector $\boldsymbol{\theta}$ will satisfy equation (56), it is convenient to set the first parameter in $\boldsymbol{\theta}$ to unity, so that $\boldsymbol{\theta} = (1, \alpha')'$. Correspondingly, matrix $\hat{\mathbf{S}}$ can be partitioned as

$$\hat{\mathbf{S}} = \begin{bmatrix} \hat{\mathbf{S}}_{11} & \hat{\mathbf{S}}_{12} \\ \hat{\mathbf{S}}_{21} & \hat{\mathbf{S}}_{22} \end{bmatrix}. \tag{57}$$

Then, using the normalization mentioned above and solving (56) for α gives

$$\hat{\alpha} = -[\hat{\mathbf{S}}_{22} - \zeta \mathbf{I}_{M+L-1}]^{-1}\hat{\mathbf{S}}_{21}. \tag{58}$$

It can be shown that under appropriate regularity conditions $\hat{\boldsymbol{\theta}} = (1, \hat{\alpha}')'$ is consistent for $\boldsymbol{\theta}$ and $\sqrt{N}$-asymptotically normal. The formal consistency argument and derivation of the asymptotic variance are provided in the online Supplement. Because the analytical formula for the variance is rather complicated, and its estimation involves advanced programming, a simpler alternative would be to estimate the variance using panel bootstrap.

Several points are worth mentioning. In (51), and therefore in (54), 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 $g(\mathbf{r}_{it}) = g(\mathbf{r}_{is})$ 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 a 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 $g(\mathbf{r}_{it})$ and $g(\mathbf{r}_{is})$ are similar, it would often be the case that $\mathbf{w}_{it}$ and $\mathbf{w}_{is}$ would also be similar, which would cause identification problems, especially in short panels.

A general shortcoming of the 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 any two continuous explanatory variables

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

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. Because the semiparametric method is not applicable when selection is binary and can not be used to estimate average partial effects, we focus on the censored selection variable case and estimate relative population parameters to facilitate comparisons.

Data are generated using equation (6), with $\mathbf{x}_{it} = (x_{it1}, x_{it2})$, $\mathbf{z}_{it} = (x_{it1}, x_{it2}, x_{it3})$, and population parameters set at $\boldsymbol{\beta} = (1, 0.6)'$, $\boldsymbol{\delta} = (0.5, 0.8, 1.2)'$, $\eta_1 = 0$, $\eta_2 = 1$, $\boldsymbol{\xi}_1 = (-0.3, -0.3, -0.3)'$, $\boldsymbol{\xi}_2 = (0.3, 0.3, 0.3)'$. Unobserved effects, a_{i1} and a_{i2}, are independent across i and distributed as $Normal(0, \sigma_a^2)$ with $\text{Corr}(a_{i1}, a_{i2}) = \rho$, where ρ is either 0 or 0.5. Idiosyncratic errors, u_{it1} and

u_{it2}, are independent across i and t and distributed as $Normal(0, \sigma_u^2)$, $\text{Corr}(u_{it1}, u_{it2}) = \rho$. The total variance of the composite errors is set to unity, whereas σ_a^2/σ_u^2 is either 0 or 0.3.

Exogenous variables are generated according to the model:

$$x_{itj} = b_{ij} + \epsilon_{itj}, \quad j = 1, 2, 3, \tag{59}$$

where b_{ij} are independent across i and distributed as $Normal(0, \sigma_b^2)$; ϵ_{itj} are independent across i and t and distributed as $Normal(0, \sigma_\epsilon^2)$; $\sigma_b^2 + \sigma_\epsilon^2 = 1$ with $\sigma_b^2/\sigma_\epsilon^2 = 0.3$; $\text{Corr}(b_{ij}, b_{ih}) = 0.25$ for $j = 1, 2, 3$, $h \neq j$. 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, we obtained $\hat{v}_{it2}$ using the symmetrically trimmed censored least squares estimator. At the second step, the covariates were orthogonalized prior to performing the estimation.[11] 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.[12]

In addition to comparing the 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 $T = 3$, $N = 500, 1000, 2500$, using 1000 replications.

Results for the estimated relative effect, $\hat{\beta}_2/\hat{\beta}_1$, are presented in Table 1. In addition to the computed biases we report root-mean squared errors (RMSE) and average analytical standard errors. For the semiparametric estimator, the average bootstrap standard errors (with 50 replications) are also reported. Because of the substantial computational intensity, we perform bootstrap only for $N = 500$ and 1000.

Under no unobserved heterogeneity and random selection ($\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$), the

[11] Specifically, we use $\tilde{\mathbf{w}}_{it} = \mathbf{w}_{it}\hat{\Omega}^{-1/2}$, where $\hat{\Omega}$ is the sample variance of $\mathbf{w}_{it}$. This produces a vector of estimated parameters $\tilde{\theta}$. The estimator of the original parameters is then recovered as $\hat{\theta} = \hat{\Omega}^{-1/2}\tilde{\theta}$.

[12] Stata do-files that were used to perform simulations, as well as the data and do-files used for implementing the empirical application, are available in the *Journal of Applied Econometrics* data archive. These are also available from the authors upon request.

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

		Probit (1)	Probit, time means (2)	s_{it} censored, 2-step MLE (3)	s_{it} binary, full MLE (4)	s_{it} censored, Semiparametric (5)
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=500	Bias	0.0041	0.0059	0.0051	0.0056	0.0055
	RMSE	0.0542	0.0685	0.0714	0.0706	0.0752
	Average se	0.0542	0.0663	0.0695	0.0683	0.0728
	Bootstrap se					0.0910
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=500	Bias	-0.0551	-0.0002	-0.0019	-0.0009	0.0067
	RMSE	0.0881	0.0743	0.0774	0.0757	0.0863
	Average se	0.0680	0.0729	0.0771	0.0745	0.0897
	Bootstrap se					0.1021
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=500	Bias	-0.1145	-0.0448	0.0006	0.0015	-0.0055
	RMSE	0.1334	0.0872	0.0781	0.0747	0.0847
	Average se	0.0709	0.0736	0.0758	0.0731	0.0827
	Bootstrap se					0.1057
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=1000	Bias	0.0016	0.0020	0.0021	0.0021	0.0033
	RMSE	0.0390	0.0485	0.0508	0.0501	0.0531
	Average se	0.0383	0.0466	0.0490	0.0481	0.0470
	Bootstrap se					0.0605
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=1000	Bias	-0.0580	0.0014	0.00002	0.0008	0.0037
	RMSE	0.0751	0.0523	0.0557	0.0537	0.0627
	Average se	0.0480	0.0518	0.0546	0.0529	0.0524
	Bootstrap se					0.0675
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=1000	Bias	-0.1156	-0.0456	-0.0002	0.0005	-0.0055
	RMSE	0.1262	0.0691	0.0537	0.0514	0.0610
	Average se	0.0502	0.0521	0.0537	0.0518	0.0558
	Bootstrap se					0.0681
				$\sigma_a^2 = 0$, $\xi_1 = 0$, $\rho = 0$		
N=2500	Bias	-0.0001	-0.0009	-0.0012	-0.0012	-0.0005
	RMSE	0.0244	0.0291	0.0306	0.0300	0.0333
	Average se	0.0242	0.0294	0.0309	0.0303	0.0279
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$		
N=2500	Bias	-0.0582	-0.0008	-0.0012	-0.0007	0.0021
	RMSE	0.0657	0.0334	0.0356	0.0341	0.0385
	Average se	0.0303	0.0326	0.0343	0.0332	0.0280
				$\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0.5$		
N=2500	Bias	-0.1160	-0.0451	0.0008	0.0010	-0.0061
	RMSE	0.1203	0.0557	0.0332	0.0323	0.0369
	Average se	0.0317	0.0329	0.0338	0.0326	0.0372

computed bias is small for all estimators, while the RMSE is the smallest for the usual probit estimator (Table 1). When adding the correlated unobserved effects ($\sigma_a^2 = 0.3$, $\xi_1 = -0.3$, $\rho = 0$), the bias in the probit estimator noticeably increases, while this is not the case for the other four estimators. RMSEs increase for all estimators. Finally, when both the correlated unobserved heterogeneity and non-random selection are present, the probit estimators with and without time means have sizable computed biases. However, all correction procedures perform well.

The observed patterns are similar when the sample size increases. Both parametric and semiparametric estimators that implement the selection correction have small biases under all scenarios, while it is not the case for the methods that ignore selection. As expected, increasing the sample size decreases RMSEs for all estimators. Perhaps not surprisingly, the RMSEs are larger for the semiparametric estimator.

With regard to inference, the average analytical standard errors of the semiparametric estimator are often too small (noticeably smaller than the RMSE). On the other hand, the bootstrap standard errors are typically too large. Indeed, when testing $H_0 : \beta_2/\beta_1 = 0.6$, the null is rejected in about 7.4 - 16.7% of cases when using the analytical standard errors (depending on the parameter values and sample size), and in approximately 2.2 - 3.9% of the cases when using bootstrap. This suggests that in practice bootstrap standard errors should be used, which should result in a conservative test.

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 the online Supplement and are similar to the case where $u_{it1} \sim Normal(0, \sigma_u^2)$. Both the usual probit estimator and probit estimator of an augmented equation that includes time means have sizable biases when $\rho = 0.5$. In contrast, parametric and semiparametric estimators that account for nonrandom selection perform well under all scenarios. They have small biases and reasonable RMSEs. Similar to the trends observed in Table 1, RMSEs of all estimators decrease with N. For the semiparametric estimator, bootstrap standard errors appear to be more reliable than

analytical standard errors.

6 Empirical Application

As an empirical application, we study pension coverage among white women, which is useful in the analysis of gender differences in pension coverage (see, for example, Even and Macpherson, 1994). The importance of this topic is further underscored by the fact that participation in retirement programs is expected to affect individual savings and wealth (Abadie 2003, Chernozhukov and Hansen 2004).

Estimation is performed using data from the National Longitudinal Survey of Youth 1979 (NLSY79), years 1990-1994. During the considered period, all respondents were at least 25 years old. The oldest respondent was 37 in 1994. Agricultural workers and those in the military service were excluded, as were observations that had missing values for the variables employed in the analysis and cases with inconsistent data entries (e.g. reporting pension coverage while not being employed). The final sample consists of 1,668 women who remained in the survey during the considered period. Between about 81 and 84 percent of women worked in a given year, so that the information on pension coverage is missing for a sizable part of the sample. The percent of working women who had pension coverage through their employer ranged from 48 to 58 percent in different years.

The dependent variable in the selection equation is hours worked within the last 52 weeks, which was generated as the total number of hours worked since the last interview divided by the total number of weeks since last interview and multiplied by 52. The dependent variable in the main equation is an indicator equal to one if the respondent's current or most recent job offered retirement benefits other than social security. Covariates in the selection equation include age, years of schooling, Armed Forces Qualification Test (AFQT) score from 1979,[13] and marital status. The individual time-mean of the marital status was included to account for the correlation with the unobserved effect. Time means of education and age were not included,[14] but their potential correlations with the unobserved effect should be captured through the

[13]Prior to performing the estimation, AFQT scores were standardized to have zero mean and unit variance

[14]There was almost no variation in years of schooling over time. Including the individual time mean of age would cause perfect collinearity with year indicators.

AFQT score. Marital status is assumed to influence the employment outcome, but not the pension coverage, which constitutes an exclusion restriction. Otherwise, covariates in the main equation are the same as in the selection equation. Sample summary statistics are reported in the online Supplement.

The data were used to estimate parametric and semiparametric models with censored selection discussed previously. In the semiparametric estimation tobit was used at the first step because standard errors could not be obtained for the CLAD and symmetrically trimmed censored least squares estimators.[15] Similar to simulations, observed covariates were orthogonalized prior to implementing the semiparametric estimation; h_g and h_ω were chosen using the cross-validation criterion. [16] Standard errors were obtained using panel bootstrap, as it performed better in simulations.

Moreover, we consider a flexible parametric model with the conditional probability $\mathrm{P}(y_{it} = 1|\mathbf{z}_i, v_{it2})$ given by

$$\Phi\left(\eta_{1\rho} + \mathbf{x}_{it}\boldsymbol{\beta}_\rho + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{1\rho} + \gamma_\rho v_{it2} + \theta_\rho(\bar{\mathbf{z}}_i\boldsymbol{\xi}_{1\rho})^2 + \psi_\rho v_{it2}^2 + \varphi_\rho(\bar{\mathbf{z}}_i\boldsymbol{\xi}_{1\rho}) \cdot v_{it2}\right), \tag{60}$$

where all coefficients are scaled. This model should be valid under weaker distributional assumptions because it allows the cumulative distribution to be a flexible function of the index and error v_{it2}.[17] Maintaining the assumption that $\mathbf{x}_{it}$ is conditionally independent of v_{it1}, and the Mundlak-Chamberlain model is correct, the effects of explanatory variables are still equal to β. In practice, v_{it2} was replaced by the residual from the first-step Tobit regression, and fully-robust standard errors corrected for the first-step estimation were obtained using panel bootstrap. As expected, the increased distributional flexibility comes at a price. Similar to the semiparametric estimator in Section 4, model (60) can be used to estimate relative effects only. Moreover, partial effects cannot be estimated because v_{it2} is not observed for $s_{it} = 0$.

Results are reported in Table 2. Coefficient estimates from parametric models are displayed in the upper panel, while estimated coefficient ratios are presented in the lower panel of the Table. The first column shows the results for a single-equation probit model that ignores selec-

[15] When performing bootstrap, both CLAD and symmetrically trimmed censored least squares failed to converge after a few replications. Such a problem typically occurs when the fraction of censored observations is large, which might have been the case in our bootstrap samples.

[16] Results were similar when reasonable changes in the bandwidth were considered ($2h_g$, $2h_\omega$ and $0.5h_g$, $0.5h_\omega$).

[17] A cubic (instead of a quadratic) specification was also considered, but higher-order terms were jointly insignificant at the 5% significance level. Also, the estimates of $\boldsymbol{\beta}_\rho$ were almost unchanged

tion. Both education and ability are associated with a higher probability of having the pension coverage, while the estimated effect of age is negative. The latter result is likely due to the specificities of the sample (these are young women of childbearing age). Similarly, hours worked are negatively related to the age and marital status, but increase with education and cognitive ability (column 2 in Table 2). Once the parametric selection correction is implemented (column 3), all coefficient estimates become slightly larger. Estimated relative effects of education and AFQT also increase. One standard deviation increase in the AFQT score is estimated to have a substantially larger effect on the probability of pension coverage than an additional year of schooling. The flexible semiparametric estimator (column 4) produces even larger estimates

Table 2: Estimated Coefficients and Coefficient Ratios.

	One-step Probit y_{it} (pension) (1)	First-step Tobit s_{it} (work hours) (2)	Second-step Probit y_{it} (pension) (3)	Second-step Flexible param. y_{it} (pension) (4)	Two-step Semiparam. y_{it} (pension) (5)
Age	-0.0401	-35.0103	-0.0527	-0.0455	
	(0.0125)	(5.4823)	(0.0125)	(0.0121)	
Education	0.0460	67.6352	0.0643	0.0730	
	(0.0142)	(6.2844)	(0.0147)	(0.0155)	
AFQT	0.1583	81.2810	0.2410	0.2433	
	(0.0348)	(0.5746)	(0.0351)	(0.0360)	
Married		-125.5764			
		(59.0559)			
$\hat{\beta}_2/\hat{\beta}_1$	-1.1469		-1.2197	-1.6057	-1.0696
	(0.9186)		(0.6577)	(1.0367)	(0.9381)
$\hat{\beta}_3/\hat{\beta}_1$	-3.9436		-4.5697	-5.3530	-2.7571
	(2.2079)		(1.7822)	(2.7131)	(1.0733)
Significance of $\hat{v}_{it2}$			$t = 12.67$		
			(0.000)		
Joint significance of $\hat{v}_{it2}$, $\hat{v}_{it2}^2$, $(\bar{\mathbf{z}}_i\boldsymbol{\xi}_{1\rho})\hat{v}_{it2}$				$\chi_2^2 = 335.50$ (0.000)	

All parametric models include year indicators.
All models include the time mean of the marital status indicator.
Standard errors (reported under the estimated coefficients and coefficient ratios) were obtained using panel bootstrap with 200 replications.
P-value is reported in parentheses under the test statistic.

of relative effects. This finding is probably not surprising, given that the first-step residuals are highly statistically significant in both fully parametric and flexible parametric regressions,

which indicates the existence of a selection bias. Finally, the semiparametric estimates of relative effects (column 5) are smaller than those produced by the other methods. Notice, however, that the estimated relative effect of the AFQT score as compared to education is the largest for the semiparametric estimator ($\hat{\beta}_{educ}/\hat{\beta}_{AFQT} = 0.388$) and is most similar to the flexible parametric estimate ($\hat{\beta}_{educ}/\hat{\beta}_{AFQT} = 0.300$). Generally, notable differences between the traditional parametric and semiparametric estimates suggest that, in this application, the underlying error distributions are likely different from normal, so that using semiparametric methods should be more appropriate.

7 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 rather well. The bias of the semiparametric estimator is somewhat larger than that of the parametric correction methods, which 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.

Finally, we use the proposed estimators for studying the determinants of pension coverage among women. Test results indicate that the selection bias is present. When correction methods are applied, the estimated relative effects have the same signs in all models. However, the semiparametric estimates are rather different from the parametric ones, indicating that the

normality assumption likely fails in this application.

Appendix: Estimating Treatment Effects

Consider a parametric model, where s_{it} is not a selection variable, but a binary treatment indicator that appears as an additional explanatory variable:

$$\begin{aligned} y_{it}^{*} &= \mathbf{x}_{it}\boldsymbol{\beta} + \psi s_{it} + c_{i1} + u_{it1}, \\ y_{it} &= 1[y_{it}^{*} > 0], \quad t = 1, \ldots, T. \end{aligned} \tag{61}$$

If parts (ii)-(iv) of Assumption 3.3 hold, the estimating equation is

$$\begin{aligned} y_{it} &= 1[\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \psi s_{it} + \gamma v_{it2} + e_{it1} > 0], \\ e_{it1}|\mathbf{z}_i, v_{it2}, s_{it} &\sim Normal(0, 1-\rho^2), \quad t = 1, \ldots, T, \end{aligned} \tag{62}$$

which uses Chamberlain's modeling approach and the argument outlined in Section 3.1. However, because s_{it} is endogenous, its individual time mean should not be included in $\bar{\mathbf{z}}_i$.

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

$$\begin{aligned} L_{it} \equiv f(y_{it}, s_{it}|\mathbf{z}_i) &= \left[\int_{-\infty}^{q_{it}} \Phi(r_{it})\phi(\nu)d\nu\right]^{y_{it}s_{it}} \\ &\times \left[\int_{-\infty}^{q_{it}} [1-\Phi(r_{it})]\phi(\nu)d\nu\right]^{(1-y_{it})s_{it}} \\ &\times \left[1-\int_{-\infty}^{q_{it}} \Phi(r_{it})\phi(\nu)d\nu\right]^{y_{it}(1-s_{it})} \\ &\times \left[1-\int_{-\infty}^{q_{it}} [1-\Phi(r_{it})]\phi(\nu)d\nu\right]^{(1-y_{it})(1-s_{it})}, \end{aligned} \tag{63}$$

where $r_{it} = (\eta_1 + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \psi s_{it} + \rho\nu)\left(1-\rho^2\right)^{-1/2}$. Similar to Section 3.3, the estimator is partial MLE. Statistical inference should generally account for serial correlation in the score functions.[18] 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 $\hat{\eta}_1$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\xi}}_1$, and $\hat{\psi}$

[18]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.

are obtained:

$$\widehat{\text{ATE}} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\Phi \left(\hat{\eta}_1 + \mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_1 + \hat{\psi} \right) - \Phi \left(\hat{\eta}_1 + \mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_1 \right) \right]. \tag{64}$$

We can also obtain ATEs for different subpopulations by fixing $\mathbf{x}_t$ at different values (which means dropping the i subscript in (64)).

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.

References

Abadie, A., 2003, Semiparametric Instrumental Variable Estimation of Treatment Response Models. *Journal of Econometrics* 113, 231-263.

Abrevaya, J. and Christian M. D., 2008, The Effects of Birth Inputs on Birthweight. *Journal of Business and Economic Statistics* 26, 379-397.

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.

Charlier, E., Melenberg, B., and van Soest, A., 2001, An analysis of housing expenditure using semiparametric models and panel data. *Journal of Econometrics* 101, 71-107.

Chernozhukov, V., and Hansen, C., 2004, The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis. *Review of Economics and Statistics* 86, 735-751.

Even, W. E. and Macpherson, D.A., 1994, Gender Differences in Pensions. *Journal of Human Resources* 29, 555-587.

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.

Jäckle, R.t, and Himmler, O. 2010. Health and Wages: Panel Data Estimates Considering Selection and Endogeneity. *Journal of Human Resources* 45(2), 364-406.

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

Maurer, J., Klein, R., and Vella, F. 2011. Subjective Health Assessments and Active Labor Market Participation of Older Men: Evidence from a Semiparametric Binary Choice Model with Nonadditive Correlated Individual-specific Effects. *Review of Economics and Statistics* 93(3), 764-774.

Meng, C. and P. Schmidt. 1985. On the Cost of Partial Observability in the Bivariate Probit Model, *International Economic Review* 26, 71-85.

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

Murphy, K.M. and R.H. Topel, 1985, Estimation and Inference in Two-Step Econometric Models, *Journal of Business and Economic Statistic* 3, 370-379.

Nam, S. and J.M. Wooldridge, 2016, On Computing Average Partial Effects in Models with Endogeneity or Heterogeneity, manuscript, Michigan State University, Department of Economics.

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.

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

Rothe, C., 2009, Semiparametric Estimation of Binary Response Models with Endogenous Regressors. *Journal of Econometrics* 153, 51-64.

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.

Terza, J.V., Basu, A., and Rathouz, P.J., 2008, Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. *Journal of Health Economics* 27, 531-543.

Vella, F., 1992, Simple Tests for Sample Selection Bias in Censored and Discrete Choice Models. *Journal of Applied Econometrics* 7, 413-421.

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.

Wooldridge, J.M., 2014, Quasi-maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables. *Journal of Econometrics* 182, 226-234.