ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 55-56 – 1999

A New Estimator for Panel Data Sample Selection Models

María Engracia ROCHINA-BARRACHINA*

ABSTRACT. – In this paper we are concerned with the estimation of a panel data sample selection model where both the selection and the regression equation contain individual effects allowed to be correlated with the observable variables. In this direction, some estimation techniques have been recently developed. We propose a new method for correcting for sample selection bias. Our estimation procedure is an extension of the familiar two-step sample selection technique to the case where one correlated selection rule in two different time periods generates the sample. Some non-parametric components are allowed. The finite sample properties of the estimator are investigated by Monte Carlo simulation experiments.

Un nouvel estimateur pour le modèle de sélection d'échantillonnage avec données de panel

RÉSUMÉ. – Dans cet article nous sommes intéressés par l'estimation d'un modèle de sélection d'échantillonnage avec données de panel et effets individuels correlationnés avec les variables explicatives. Nous proposons une nouvelle méthode pour corriger le biais de sélection d'échantillonnage. Nous permettons certaines composantes non paramétriques et nous recherchons selon les études de Monte Carlo le comportement de l'estimateur en échantillonnage limité.

* M.E. ROCHINA-BARRACHINA, University of Valencia, Spain.
This paper is part of my thesis dissertation at University College London, United Kingdom. Financial support from the Spanish foundation " Fundación Ramón Areces" is gratefully acknowledged. Thanks are also owed to M. ARELLANO, H. BIERENS, R. BLUNDELL, E. CHARLIER, B. HONORÉ, J. HOROWITZ, H. ICHIMURA, E. KYRIAZIDOU, M.J. LEE, D. MCFADDEN, C. MEGHIR, B. MELENBERG, S. THOMPSON, F. VELLA, F. WINDMEIJER, a co-editor and two anonymous referees for their very helpful comments on preliminary drafts. Earlier versions of the paper were presented at the "*Lunch Seminars*" at Tilburg University, June 1996, The Netherlands; at the 7th-Meeting of the European Conferences of the Econometrics Community (EC-Squared Conference) on Simulation Methods in Econometrics, December 1996, Florence, Italy; at the XXI Simposio de Análisis Económico, December 1996, Barcelona, Spain; at the 7th-International Conference on Panel Data, June 1997, Paris, France; and at the European Meeting of the Econometrics Society (ESEM'97), August 1997, Toulouse, France. The usual disclaimer applies.

1 Introduction

In this paper we are concerned with the estimation of a panel data sample selection model where both the binary selection indicator and the regression equation of interest contain unobserved individual-specific effects that may depend on the observable explanatory variables. For this case, not many estimators are available. The most recent ones are the estimators developed by WOOLDRIDGE [1995] and KYRIAZIDOU [1997]. Both of them are semiparametric in the sense that the model does not need to be fully specified. WOOLDRIDGE [1995] proposes a method under a parameterization of the sample selection mechanism, a conditional mean independence assumption for the time-varying errors in the main equation and some linear projections. A marginal normality assumption for both the individual effects and the idiosyncratic errors in the selection equation is imposed. KYRIAZIDOU [1997] proposes an estimator under much weaker conditions, in the sense that the distributions of all unobservables are left unspecified. The method allows for an arbitrary correlation between individual effects and regressors, but a joint *conditional exchangeability* assumption for the idiosyncratic errors in the model is needed.

The purpose of this paper is to propose an estimator that relaxes some of the assumptions in the above methods. Specifically, the estimator allows for an unknown conditional mean of the individual effects in the main equation, in contrast to WOOLDRIDGE [1995], and it also avoids the *conditional exchangeability* assumption in KYRIAZIDOU [1997]. We can see the estimator as complementary to those previously suggested, in the sense that it uses an alternative set of identifying restrictions to overcome the selection problem. In particular, the estimator imposes that the joint distribution of the time differenced regression equation error and the two selection equation errors, conditional upon the entire vector of (strictly) exogenous variables, is normal.

We assume that a large number of observations in the cross-section are available and the asymptotic properties hold as the number of individuals goes to infinity. "*Fixed-length*" panels are the most frequently encountered in practice. We base our analysis on two periods. Consequently, we get estimates based on each two waves we can form with the whole length of the panel, and then we combine them using a minimum distance estimator (see CHAMBERLAIN [1984]). This device allows us to focus on two-waves. In the paper we will discuss the extension of our estimation method to cover the more general situation.

The method follows the familiar two-step approach proposed by HECKMAN [1976, 1979] for sample selection models. HECKMAN [1976, 1979] proposed a two-stage estimator for the one selection rule case, and this has been extended to two selection rule problems with cross-section data by both HAM [1982] and POIRIER [1980]. In particular, our estimation procedure is an extension of HECKMAN's [1976, 1979] sample selection technique to the case where one correlated selection rule in two different time periods generates the sample. The idea of the estimator is to eliminate the individual effects from the equation of interest by taking time differences, and then to condition upon the

outcome of the selection process being "*one*" (observed) in the two periods. This leads to two correction terms, the form of which depends upon the assumptions made about the selection process and the joint distribution of the unobservables. With consistent first step estimates of these terms, simple least squares can be used to obtain consistent estimates in the second step.

We present two versions of the estimator depending on a varying degree of parametric assumptions for the first step estimator. The more semiparametric estimator generalises CHAMBERLAIN's [1980] approach to allow for correlation between the individual effects and the explanatory variables. This generalisation, as already pointed out by NEWEY [1994a] in the context of panel data probit models with semiparametric individual effects, allows for the conditional expectation of the individual effects in the selection equation to be unspecified. Given that the second step allows for temporal dependence and different variances for the errors in different time periods, we are interested in estimators for the first step that do not impose restrictions on the serial correlation and/or heteroskedasticity over time for the errors.

The results of this paper may be useful in a variety of situations that are analysed in practice. A classic example is female labour supply, where hours worked are observed only for those women who decide to participate in the labour force. Failure to account for sample selection is well known to lead to inconsistent estimation of the behavioural parameters of interest, as these are confounded with parameters that determine the probability of entry into the sample. The same problem appears when modelling company investment strategies and household consumption, where the analysis of these expenditures is conditioned to a prior investment or consumption decision, respectively.

The paper is organised as follows. Section 2 describes the model, sets out the estimation problem and presents the estimator. In Section 3 we consider estimation of the selection equation. Section 4 discusses the way the estimators based in two periods of a longer panel can be combined to get consistent and unique estimates for the whole panel. Section 5 reports results of a small Monte Carlo simulation study of finite sample performance. Section 6 gives concluding remarks. The Appendices provide formulae for the asymptotic variance of the estimator.

2 The Model and the Proposed Estimator

The model we consider is a panel data sample selection model with a binary selection equation. Both the sample selection rule and the regression equation of interest contain additive permanent unobservable individual effects possibly correlated with the explanatory variables.

The model can be written as follows,

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}; \quad i = 1,\ldots,N; \quad t = 1,\ldots,T, \tag{2.1}$$

(2.2) $\quad d^*_{it} = z_{it}\gamma - \eta_i - u_{it}; \quad d_{it} = 1[d^*_{it} \geqslant 0],$

where, $\beta \in \Re^k$ and $\gamma \in \Re^f$ are unknown parameter (column-) vectors, and x_{it}, z_{it} are vectors of strictly exogenous explanatory variables with possible common elements. α_i and η_i are unobservable time-invariant individual-specific effects, which are presumably correlated with the regressors. ε_{it} and u_{it} are idiosyncratic errors not necessarily independent of each other. Whether or not observations for y_{it} are available is denoted by the dummy variable d_{it}.

Estimation of β based on the observational equation (2.1) is confronted with two problems. First, the presence of the unobserved effect α_i, and second, the sample selection problem. By following an estimation procedure that just uses the available observations one is implicitly conditioning upon the outcome of the selection process, *i.e.*, upon $d_{it} = 1$. The problem of selectivity bias arises from the fact that this conditioning may affect the unobserved determinants of y_{it}.

The first problem is easily solved by noting that for those observations that have $d_{it} = d_{is} = 1(s \neq t)$, time differencing will eliminate the effect α_i from equation (2.1). This is analogous to the "*fixed-effects*" approach used in linear panel data models. Application of standard methods, *e.g.* OLS, on this time-differenced subsample will yield consistent estimates of β if the following condition holds:

(2.3) $\quad E(\varepsilon_{it} - \varepsilon_{is} | x_i, z_i, d_{it} = d_{is} = 1) = 0, \; s \neq t,$

where $x_i \equiv (x_{i1}, \ldots, x_{iT})$ and $z_i \equiv (z_{i1}, \ldots, z_{iT})$.

In general though, OLS estimation of model (2.1), using pairwise differences over time for individuals satisfying $d_{it} = d_{is} = 1$ $(s \neq t)$, would be inconsistent due to sample selectivity, since the conditional expectation in (2.3) would be, in general, unequal to zero.

The basic idea of the estimator relies on a parameterization of the conditional expectation in (2.3). To do that, some assumptions have to be imposed. There are two assumptions on the unobservables in the selection equation (A1 and A2 below). A third assumption (A3) imposes restrictions on the joint conditional distribution of the error terms in the two equations. The method is nonparametric with respect to the individual effects in the main equation and allows selection to depend on α_i in an arbitrary fashion. Under its less parametric version, the conditional mean of the individual effects in the selection equation is allowed to be an unknown function of the whole time span of the explanatory variables.

• ***A1 :***

A1A) With parametric individual effects: The regression function of η_i on z_i is linear. Following CHAMBERLAIN [1980], the method specifies the conditional mean of the individual effects in the selection equation as a linear projection on the leads and lags of the observable variables: $\eta_i = z_{i1}\delta_1 + \ldots + z_{iT}\delta_T + c_i$ where c_i, is a random effect.

A1B) With semiparametric individual effects: The conditional mean of η_i on z_i is left unrestricted: $\eta_i = E(\eta_i | z_i) + c_i$. This generalisation of CHAMBERLAIN [1980] is already used in NEWEY [1994a] for a panel probit model with strictly exogenous variables, and ARELLANO and CARRASCO

[1996] for binary choice panel data models with predetermined variables.

• ***A2 :*** *The errors in the selection equation,* $v_{it} = u_{it} + c_i$*, are normal* $(0,\sigma_t^2)$. This is a normality assumption for the underlying errors in the selection equation. Temporal dependence is allowed. This is important because, whether or not the u_{it} are independent across *t*, the v_{it} can never be counted on to be serially independent. Note also that the v_{it} are allowed to have different variances in different time periods.

• ***A3 :*** *The errors* $[(\varepsilon_{it} - \varepsilon_{is}), v_{it}, v_{is}]$ *are trivariate normally distributed conditional on* x_i *and* z_i. Additionally to the assumptions in the selection equation, one assumption about the relationship between $(\varepsilon_{it} - \varepsilon_{is})$ and (v_{it}, v_{is}) has to be imposed to obtain the functional form of the sample selection correction terms that correspond to the conditional expectation in (2.3). In particular, a trivariate normal distribution is assumed for the joint conditional distribution of the error terms. However, the normality assumption is unessential and could be replaced by other parametric assumptions. Different distributional assumptions will lead to a corresponding modification of the selectivity bias terms. [1] In any case, in this paper we derive the sample selection correction terms under normality. The multivariate normal distribution is the most commonly specified assumption on sample selection models. [2]

The assumptions above highlight the crucial deviation from KYRIAZIDOU'S [1997] work. There, the sample selection effects are considered as an unknown function of both the observed regressors and the unobservable individual effects in the selection equation. In her approach, the distributions of all unobservables are left unspecified, but an assumption on the underlying time-varying errors in the model is needed. That assumption is the *conditional exchangeability* condition and consists in the following. Given the model in (2.1) and (2.2), the joint distribution of the errors $(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is})$ and $(\varepsilon_{is}, \varepsilon_{it}, u_{is}, u_{it})$ is identical conditional on $\xi_i \equiv (z_{it}, z_{is}, x_{it}, x_{is}, \alpha_i, \eta_i)$. That is, $F(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is} | \xi_i) = F(\varepsilon_{is}, \varepsilon_{it}, u_{is}, u_{it} | \xi_i)$. Under this *conditional exchangeability* assumption, for an individual i with $z_{it}\gamma = z_{is}\gamma$, sample selection would be constant over time. For this individual, applying time-differences in equation (2.1), conditional on observability in the two periods, will eliminate both the unobservable effect α_i and the sample selection effects. The *conditional exchangeability* assumption implies a conditional stationarity assumption for the idiosyncratic errors in the selection equation. That means that u_{it} is stationary conditional on (z_{it}, z_{is}, η_i), that is $F(u_{it} | z_{it}, z_{is}, \eta_i) = F(u_{is} | z_{it}, z_{is}, \eta_i)$. The quoted assumption turns out to be

---

1. However, as pointed out in LEE [1982] it can happen that a particular binary choice model, for example, the arctan model, can be unsuitable to obtain selectivity bias terms for the regression equation. As the arctan probability model is based on the assumption that the distribution of the error term is Cauchy, the conditional mean for the dependent variable in the regression equation does not even exist.

2. Furthermore, to develop correction terms for selectivity bias based on the normal distribution does not necessarily imply lack of distributional flexibility. As pointed out in LEE [1982] even if the assumed distribution function for the disturbance in the probability choice model is not nornal, it is still possible to apply the correction terms for sample selection derived under multivariate normal disturbances. What is needed is to specify a strictly increasing transformation function $J = \Phi^{-1}F$, where Φ is the standard normal distribution function and F is the assumed distribution for a disturbance u, such that the transformed random variable $u^* = J(u)$ is standard normal.

restrictive since it also implies homoskedasticity over time of the idiosyncratic errors in the main equation. Under this assumption time effects, if any, are absorbed into the conditional mean, but they cannot affect the error structure of the model. Here we attempt to relax the assumption that the errors for a given individual are homoskedastic over time.

In our approach, we give a shape to the (generally unknown) sample selection effects. This requires explicitly allowing for statistical dependence of the individual effects in the selection equation on the observable variables. Assumption (AlA) specifies a functional form for that relation, although under our less parametric assumption (AlB) a particular specification is not needed. Furthermore, we also specify the full distribution of the differenced time varying errors in the main equation and the error terms in the selection equation.

Under assumptions Al-A3, the form of the selection term, to be added as an additional regressor to the differenced equation in (2.1), can be worked out (see, for instance, TALLIS [1961]). Consequently, the conditional mean in (2.3) can be written as:

$$E(\varepsilon_{it}-\varepsilon_{is}|x_i,z_i,v_{it}\leqslant H_{it},v_{is}\leqslant H_{is}) = \sigma_{(\varepsilon_t-\varepsilon_s)(v_t/\sigma_t)}\cdot\lambda_{its}+\sigma_{(\varepsilon_t-\varepsilon_s)(v_s/\sigma_s)}\cdot\lambda_{ist}, \tag{2.4}$$

where $H_{i\tau}=z_{i\tau}\gamma-E(\eta_i|z_i)$ for $\tau=t,s$, are the reduced form indices in the selection equation for period t and s. Our lambda terms are as follows:

$$\begin{aligned}\lambda_{its}&=\phi(M_{it})\Phi(M^*_{its})/\Phi_2(M_{it},M_{is},\rho_{ts}),\\ \lambda_{ist}&=\phi(M_{is})\Phi(M^*_{ist})/\Phi_2(M_{it},M_{is},\rho_{ts}),\end{aligned} \tag{2.5}$$

where

$$\begin{aligned}M_{it}&=\frac{H_{it}}{\sigma_t},\ M_{is}=\frac{H_{is}}{\sigma_s}\\ M^*_{its}&=(M_{is}-\rho_{ts}M_{it})/(1-\rho^2_{ts})^{1/2},\\ M^*_{ist}&=(M_{it}-\rho_{ts}M_{is})/(1-\rho^2_{ts})^{1/2},\end{aligned} \tag{2.6}$$

and $\rho_{ts}=\rho_{(v_t/\sigma_t)(v_s/\sigma_s)}$ is the correlation coefficient between the errors in the selection equation. $\phi(\cdot)$ is the standard normal density function, and $\Phi(\cdot)$, $\Phi_2(\cdot)$ are the standardised univariate and bivariate normal cumulative distribution functions, respectively. [3]

The estimation equation is given by

---

3. The terms M^*_{its}, M^*_{ist}, appear because in the bivariate normal distribution with density function $\phi_2(M_{it},M_{is},\rho_{ts})$ if we fix, for instance, the value of M_{is} we can write $\phi_2(M_{it},M_{is},\rho_{ts})=\phi(M_{is})\,\phi((M_{it}-\rho_{ts}M_{is})/(1-\rho^2_{ts})^{1/2})$. The following also holds:

$$\int_{-\infty}^{M_{it}}\phi(M_{is})\phi\Big((M_{it}-\rho_{ts}M_{is})/(1-\rho^2_{ts})^{1/2}\Big)dM_{it}=\phi(M_{is})\Phi\left((M_{it}-\rho_{ts}M_{is})/(1-\rho^2_{ts})^{1/2}\right)$$

A corresponding expression will be obtained if M_{it}, is the fixed element. To calculate λ_{its} we have conditioned on M_{it} being fixed and we integrate over M_{is}. To calculate λ_{ist} we do the reverse. The factor that appears in the denominator of both lambdas, $\Phi_2(M_{it},M_{is},\rho_{ts})$, is just a normalising factor.

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + \ell_{ts} \cdot \lambda(M_{it}, M_{is}, \rho_{ts}) + \ell_{st} \cdot \lambda(M_{is}, M_{it}, \rho_{ts}) + e_{its}, \tag{2.7}$$

where $e_{its} \equiv (\varepsilon_{it} - \varepsilon_{is}) - [\ell_{ts} \cdot \lambda(M_{it}, M_{is}, \rho_{ts}) + \ell_{st} \cdot \lambda(M_{is}, M_{it}, \rho_{ts})]$ is a new error term, which by construction satisfies $E(e_{its}|x_i, z_i, v_{it} \leqslant H_{it}, v_{is} \leqslant H_{is}) = 0$. Now, the solution to the problem is immediate. Assuming that we can form consistent estimates of λ_{its} and λ_{ist}, least squares estimation (with modified standard errors) applied to (2.7) can be used to obtain consistent estimates of β, ℓ_{ts} and ℓ_{st}. [4] A test of the restrictions H_0: $\ell = 0$, for $\ell = (\ell_{ts}, \ell_{st})'$, is easily carried out by constructing a Wald statistic. This can be used as a test for selection bias. To be able to estimate λ_{its} and λ_{ist}, we need to get consistent estimates of M_{it}, M_{is} and ρ_{ts}. The way of getting these values is what is going to make the distinction between a parametric first step estimator and a semiparametric one.

3 Estimation of the Selection Equation

To construct estimates of the $\lambda(\cdot)$ terms in (2.7) we have two alternatives. In the more parametric approach we assume that the $E(\eta_i|z_i)$ is specified as a linear projection on the leads and lags of observable variables (as in CHAMBERLAIN [1980], VERBEEK and NIJMAN [1992], and WOOLDRIDGE [1995]). [5] With this parameterization of the individual effects we go from the structural equation in (2.2) to the following reduced form selection rule: [6]

$$d_{it} = 1\{\gamma_{t0} + z_{i1}\gamma_{t1} + \ldots + z_{iT}\gamma_{tT} - v_{it} \geqslant 0\} \equiv 1\{H_{it} - v_{it} \geqslant 0\}. \tag{3.1}$$

We can form a likelihood function based in the fact that we observe four possible outcomes per each two time periods. Those individuals with $d_{it} = d_{is} = 1$; those with $d_{it} = d_{is} = 0$; those with $d_{it} = 1$ and $d_{is} = 0$; and those with $d_{it} = 0$ and $d_{is} = 1$. The probabilities that enter the likelihood

4. However, notice that to be able to identify ℓ_{ts} from ℓ_{st} time variation of M is needed. M may vary even when z is constant, if σ_t/σ_s is not unity (σ_t/σ_s could be identified; see CHAMBERLAIN [1982]). But even if ℓ_{ts} and ℓ_{st}, are not identified (only $\ell_{ts} + \ell_{st}$ is), this does not matter for the identification of the parameter of interest β.

5. Our main interest in introducing individual effects is motivated by the possibility of existence of missing variables that are correlated with z_i. If one mistakenly models η_i as independent of z_i, then the omitted variable bias is not eliminated. Then, we want to specify a conditional distribution for η_i given z_i that allows for dependence. A convenient possibility is to assume that the dependence is only via a linear regression function.

6. In fact, as WOOLDRIDGE [1995] pointed out, the mechanism described by (3.1) can be the reduced form also for other structural selection equations. For example, consider the dynamic model

   (a) $$d_{it} = 1\{\gamma_0 + \theta d^*_{i,t-1} + z_{it}\gamma - u_{it} \geqslant 0\},$$

   where u_{it} is a mean zero normal random variable independent of z_i. Then, assuming that d^*_{i0} given z_i is normally distributed with linear conditional expectation, (a) can be written as (3.1). The same conclusion holds if an unobserved individual effect of the form given in assumption (A1A) on the main text is added to (a).

function are Prob $(D_t = d_{it}, D_s = d_{is}) = \Phi_2(q_{it}M_{it}, q_{is}M_{is}, \rho^*_{its})$, where the new terms q_{it}, q_{is} and ρ^*_{its} are defined by $q_{it} = 2d_{it} - 1$, $q_{is} = 2d_{is} - 1$ and $\rho^*_{its} = q_{it}q_{is}\rho_{ts}$, respectively. This notational shorthand accounts for all the necessary sign changes needed to compute probabilities for d's equal to zero and one.

The reduced form coefficients (γ_t, γ_s) will be jointly determined with ρ_{ts} through the maximisation of a bivariate probit for each combination of time periods. See Appendix I for the variance-covariance matrix of β, ℓ_{ts} and ℓ_{st}, when we follow this parametric first step approach.

In our alternative approach, to allow for semiparametric individual effects in the selection equation, the conditional expectations $E(d_{i\tau}|z_i) = \Phi(M_{i\tau})$ for $\tau = t,s$ are replaced with nonparametric estimators $\widehat{h}_\tau(z_i) = \widehat{E}(d_{i\tau}|z_i)$, such as kernel estimators. [7] The way to recover estimated values for the M_i's is given by the inversion $\widehat{M}_{i\tau} = \Phi^{-1}[\widehat{h}_\tau(z_i)]$. In contrast to MANSKI's [1987] or KYRIAZIDOU's [1997] semiparametric individual effects models, u_{it} is allowed to be heteroskedastic over time. [8] Even if the method leaves $E(\eta_i|z_i)$ unrestricted, it may implicitly restrict other features of the distribution of $\eta_i|z_i$ by assuming that the distribution of $(\eta_i + u_{it})|z_i$ is parametric. [9] A likelihood function like the one above is now maximised just with respect to the parameter ρ_{ts}. See Appendix II for the variance-covariance matrix of, β, ℓ_{ts}, and ℓ_{st}, when we follow this semiparametric first step estimator.

In order to compute the $\widehat{h}_\tau(z_i)$ values in an application we will use the so-called NADARAYA-WATSON kernel regression function estimator, named after NADARAYA [1964] and WATSON [1964]. The NADARAYA-WATSON estimator has an obvious generalisation to multivariate and high order kernels. According to it, the corresponding nonparametric regression function estimator of $\widehat{h}_\tau(z_i)$ is

$$\widehat{h}_\tau(z_i) = \frac{\sum_{j=1}^{N} d_{jt} K[(z_i - z_j)/c_N]}{\sum_{j=1}^{N} K[(z_i - z_j)/c_N]} \tag{3.2}$$

where $d_t \in \{0,1\}$ and $z \in \Re^{T \cdot f}$. The d's are the dependent variables and the z's are $T \cdot f-$ component vectors of regressors.

For practical issues one needs to choose the kernel function K and a parti-

---

7. In fact, with this way to get estimated probabilities we do not longer need a parametric assumption about the form of the selection indicator index. The linearity assumption would be needed if we were interested not just in the index value, $M_{i\tau}$, but also in recovering the parameters in the selection equation. As our concern is the consistent estimation of the sample selection correction terms, the latter is not needed. This flexibility is convenient because although the form of this function may not be derived from some underlying behavioural model, the set of conditioning variables that govern the selection probability may be known in advance.

8. The advantage of the proposed estimator is that it allows the variance of the errors to vary over time. We relax the assumption that the errors for a given individual are homoskedastic over time. The price we pay is in terms of $(\eta_i + u_{it})|z_i$ being parametrically distributed. Also the amount of heteroskedasticity across individuals is restricted.

9. Potentially, a test for the linear correlation of the individual effects in the selection equation with respect to the explanatory variables could be performed.

cular bandwidth parameter c_N. Related literature advises the use of high order bias-reducing kernels that can be constructed following BIERENS [1987]. A simple way to construct kernels in $K_{T\cdot f,R+1}$ for arbitrary $T\cdot f \geqslant 1$ and even $R+1 \geqslant 2$ (where R is an odd integer $\geqslant 1$) is the following. [10] For $z \in \Re^{T\cdot f}$ and $\dfrac{R+1}{2} \geqslant 1$ let

(3.3)

$$K_{T\cdot f,R+1}\left(\frac{z_i - z_j}{c_N}\right) = \sum_{p=1}^{\frac{R+1}{2}} \frac{\theta_p \exp\left(-\frac{1}{2}\left(\frac{z_i - z_j}{c_N}\right)' \Omega^{-1}\left(\frac{z_i - z_j}{c_N}\right)/\mu_p^2\right)}{(\sqrt{2\pi})^{T\cdot f}\cdot|\mu_p|^{T\cdot f}\sqrt{\det(\Omega)}}$$

where Ω is a positive definite matrix and the parameters θ_p and μ_p are such that

$$\sum_{p=1}^{\frac{R+1}{2}} \theta_p = 1; \quad \sum_{p=1}^{\frac{R+1}{2}} \theta_p \cdot \mu_p^{2v} = 0 \tag{3.4}$$

for $v = 1,2,\ldots,\dfrac{R+1}{2} - 1$. We should specify $\Omega = \widehat{V}$, where $\widehat{V}$ is the sample variance matrix; that is, $\widehat{V} = \dfrac{1}{N}\sum_{j=1}^{N}(z_j - \overline{z})'(z_j - \overline{z})$ with $\overline{z} = \dfrac{1}{N}\sum_{j=1}^{N} Z_j$. Thus, for $R+1 = 2,4,6,\ldots$, we get

(3.5)

$$\widehat{K}_{T\cdot f,R+1}\left(\frac{z_i - z_j}{c_N}\right) = \sum_{p=1}^{\frac{R+1}{2}} \frac{\theta_p \exp\left(-\frac{1}{2}\left(\frac{z_i - z_j}{c_N}\right)' \widehat{V}^{-1}\left(\frac{z_i - z_j}{c_N}\right)/\mu_p^2\right)}{(\sqrt{2\pi})^{T\cdot f}\cdot|\mu_p|^{T\cdot f}\sqrt{\det(\widehat{V})}}$$

We will now focus on the problem of bandwidth selection. We need the convergence rate of $\widehat{h}$ to be faster enough. To build up such convergence rates we will use the uniform convergence rates of BIERENS [1987]. From the point of view of uniform consistency and under conditions satisfied by the high-order bias reducing kernels of BIERENS,

$$\min\left(c_N^{T\cdot f}\sqrt{N}, c_N^{-(R+1)}\right)\cdot \sup_{z\in\{z\in\Re^{T\cdot f}\}} |\widehat{h}_\tau(z) - h_\tau(z)| \tag{3.6}$$

is stochastically bounded. Clearly, the best uniform consistency rate is obtained for c_N such that $\min(c_N^{T\cdot f}\sqrt{N}, c_N^{-(R+1)})$ is maximal. This is the

10. For an integer q, let $m_q(K) = \int u^q K(u)du$. Then, the order $(R+1)$ of the kernel $K(\cdot)$ is defined as the first nonzero moment: $m_q = 0,\ q = 1,\ldots,R$; $m_q \neq 0$. Positive kernels can be at most of order $2(R=1)$.

case if $c_N \propto N^{-1/[2(R+1)+2T\cdot f]}$. We then have $\min(c_N^{T\cdot f}\sqrt{N}, c_N^{-(R+1)}) \propto N^{(R+1)/[2(R+1)+2T\cdot f]}$. Thus, the sequence of bandwidths C_N, used in the estimation is of the form $c_N = c \cdot N^{-1/[2(R+1)+2T\cdot f]}$ for a value of R that should satisfy the inequality $R+1 \geqslant T \cdot f$.

Now, the bandwidth selection problem is reduced to choose the constant c. A natural way to proceed (HÄRDLE and LINTON [1994], HÄRDLE [1990] and SILVERMAN [1986]) is to choose c so as to minimise some kind of measure of the "*distance*" of the estimator from the true value (according to some performance criterion). If we are, for example, interested in the quadratic loss of the estimator at a single point z, which is measured by the mean squared error, $MSE\{\widehat{h}_t(z)\}$, then we will minimise the MSE over c in order to get an approximately optimal value for c.

By following this minimisation method the optimal bandwidth depends on elements that we do not know unless we know the optimal bandwidth. In practice, these quantities have to be estimated on the basis of some preliminary smoothing process which raises a second-order bandwidth selection problem. A first step bandwidth c should be chosen to estimate the quantities that determine the optimal bandwidth c^*.

4 Single Estimates for the Whole Panel

In section 2 and 3 we have introduced the proposed second and first stage estimators. There, the analysis was based on two periods. In this section we will illustrate how to combine the estimators coming from each two waves of the panel to come up with a single estimate.

The estimators used in the first step are obtained by maximising the likelihood function of $\binom{T}{2}$ bivariate probits separately, each of them based in a different combination of two time periods. Once estimates of the correction terms are included, equation (2.7) can be estimated using least squares for each combination of panel waves (t,s), $t \neq s$, [11] this gives a total of $\binom{T}{2}$ pairs for a panel of length T. A minimum distance procedure, with the corresponding weighting matrix, can then be applied to combine these estimates. To estimate the weighting matrix an estimate for the covariance matrix of the estimators for the different time periods is required. The block diagonal matrices are simply the corresponding covariance matrices estimates for each

11. Notice that for sample selection models we gain in efficiency by considering all possible pairs in place of just first differences. Different combinations of individuals appear in different pairs because of the observability rule driven by $d_{it} = 1$ or $d_{it} = 0$.

pair. To get the block off-diagonal matrices of the weighting matrix we just need to combine the corresponding two *influence functions* for each combination of two pairs. These covariances among pairs do not converge to zero. In the minimum distance step we restrict the estimates for β to be the same for each combination (*t*,*s*) and we estimate $\binom{T}{2} \times 2$ coefficients for the correction terms in all the pairs (two correction terms per pair). [12] The latter group of parameters is left unconstrained in the minimum distance step. The number of parameters associated to the correction terms is a function of T and it grows faster than T. This is so because the estimator allows the variance of the time varying errors in both equations to vary over time and it does not restrict the correlations between the time-differenced errors in the main equation and the errors in the selection rule to be time-invariant. As we focus on the case where the data consist of a large number of individuals observed through a small (fixed) number of time periods and look at asymptotics as the number of individuals approaches infinity, the growth in parameters does not impose a problem, in principle.

Alternatively, for the more parametric first step, we can use a strategy that might asymptotically be more efficient. A minimum distance estimator can be obtained from the $\binom{T}{2}$ sets of bivariate probits estimates that we can form with a panel of length T. Lambda terms based on the resulting estimates can be plugged into equation (2.7) that is again estimated by pairs. A second minimum distance step is computed in the same way it was applied for the estimator based on the results of bivariate probits estimated separately. However, although this strategy might asymptotically be more efficient, the other one is easier from a practical point of view and it still provides consistent estimates.

To test for the assumption of the $\binom{T}{2} \times 2$ correction terms being jointly significant is easily carried out by constructing a Wald statistic. This can be used as a test for selection bias. A test of overidentifying restrictions in the minimum distance step can be also performed. The latter, in fact, implies testing whether the imposed restrictions (β being constant over time) cannot be rejected.

A *curse of dimensionality* problem, well known in the nonparametric literature, can appear in the case of many continuous variables in the sample selection equation and/or a large number of time periods. This affects the quality of the nonparametric estimator $\widehat{h}_\tau(z_i) = \widehat{E}(d_{i\tau}|z_i)$ in section 3 obtained by using high-dimensional smoothing techniques. [13]

To overcome this difficulty in nonparametric estimation some dimension reduction approaches have been proposed. Common restrictions are additive or multiplicative separability among the variables in the selection equation

12. In the minimum distance step we can recover a time trend coefficient or time dummies coefficients (reflecting changing economy conditions common to all individuals).

13. Estimation precision decreases as the dimension of z_i increases.

14. For some of these approaches see HÄRDLE and CHEN [1995] and HOROWITZ [1998].

that would allow to move from high-dimension multivariate kernels to lower-dimension or even univariate kernels. [14] Current literature in nonparametric methods is trying to find alternative definitions of separability to overcome this problem. Another alternative, that can also be applied to the more parametric first step, consists on assuming that the individual effect in the selection equation depends only on the time average of the time varying variables (see for example MUNDLAK [1978], NIJMAN and VERBEEK [1992], and ZABEL [1992]) This economises on parameters or dimension but also imposes restrictions on the relationship between η_i and z_i that could be violated, especially if the z_{it} are trending.

5 Monte Carlo Experiments

In this section we report the results of a small simulation study to illustrate the finite sample performance of the proposed estimators. Each Monte Carlo experiment is concerned with estimating the scalar parameter β in the model

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}; \quad i = 1,\ldots,N; \quad t = 1,2,$$
$$d_{it}^* = z_{1it}\gamma_1 + z_{2it}\gamma_2 - \eta_i - u_{it}; \quad d_{it} = 1[d_{it}^* \geqslant 0],$$

where y_{it} is only observed if $d_{it} = 1$. The true value of β, γ_1, and γ_2 is 1; z_{1it} and z_{2it} follow a $N(0,1)$; x_{it} is equal to the variable z_{2it} (we have imposed one exclusion restriction); otherwise stated something different the individual effects are generated as [15] $\eta_i = -[(z_{1i1} + z_{1i2})/2 + (z_{2i1} + z_{2i2})/2 + N(0,1) + 0.07]$ and $\alpha_i = (x_{i1} + x_{i2})/2 + \sqrt{2} \cdot N(0,1) + 1$; the different types of errors in each experiment are shown at the top of the corresponding tables. For all the experiments the errors in the main equation are generated as a linear function of the errors in the selection equation, which guarantees the existence of non-random selection into the sample.

The results with 100 replications and different sample sizes are presented in *Tables 1-5'*. All tables report the estimated mean bias for the estimators, the small sample standard errors (SE), and the standard errors predicted by the asymptotic theory (ASE). As not all the moments of the estimators may exist in finite samples some measures based on quantiles, as the median bias, and

15. Their particular design is driven by the fact that at this stage we want to keep both a linear correlation with respect to the explanatory variables and a normality assumption for the remaining random terms. The reason is that methods like the one proposed by WOOLDRIDGE [1995] and our proposed estimator assume normality for the remaining random terms in the selection equation. This means that the difference between η_i and its conditional mean is a random normal error. At the same time, both WOOLDRIDGE [1995] and our more parametric new estimator are developed under the assumption of a linear correlation between individual effects in the selection equation and the leads and lags of the explanatory variables. In particular, we have assumed that this linear correlation follows MUNDLAK's [1978] formulation. Furthermore, WOOLDRIDGE [1995] also imposes the linearity assumption (that is not needed for our estimator) for the individual effects in the main equation. It is also quite common to assume that there is a constant term in the individual effects.

the median absolute deviation (MAD) are also reported. In Panel A for all tables we report the finite sample properties of the estimator that ignores sample selection and is, therefore, inconsistent. The purpose in presenting these results is to make explicit the importance of the sample selection problem for each of our experiments. This estimator is obtained by applying least squares to the model in time differences for the sample of individuals who are observed in both time periods, *i.e.* those that have $d_{i1} = d_{i2} = 1$.

As we pointed out earlier, with our estimator we aimed to relax some of the assumptions of the currently available methods. Specifically, we wanted to avoid the misspecification problems that could appear by breaking the linear projection assumption for the individual effects in the main equation on the explanatory variables in the case of WOOLDRIDGE's [1995] estimator. Furthermore, we also wanted to avoid the *conditional exchangeability* assumption in KYRIAZIDOU [1997] and to allow the time-varying errors to be heteroskedastic over time.

In *Table 1* we compare the two versions of our estimator with WOOLDRIDGE's [1995] and KYRIAZIDOU's [1997] estimators when the *conditional exchangeability* assumption breaks down. We allow for no-constant variances over time for the error terms and different degrees for the sample selection problem. The latter comes through different correlation coefficients over time, for the idiosyncratic errors in the selection equation and the idiosyncratic errors in the main equation. Varying the mean of the idiosyncratic errors in the main equation from 0 (in period 2) to -5 (in period 1), does not affect, in principle, any of the estimators. For this to be true, a constant term is included in the estimators in differences to pick up the change in means. For the estimator that does not rely on time-differences, WOOLDRIDGE's [1995] estimator, the minimum distance step allows for a time-varying intercept.[16] For WOOLDRIDGE's [1995] estimator, although other procedures could be used – such as pooled least squares (the simplest consistent estimator) – we have applied minimum distance estimation. Panel C reports the results for KYRIAZIDOU's [1997] estimator. We report the results for the estimator using the true γ and the one estimated by smoothed conditional maximum score in the construction of the kernel weights.[17] For the former we implement second (R = 1), fourth (R = 3), and sixth (R = 5) higher order bias reducing kernels of BIERENS [1987] according to section 3 above. They correspond to a normal, to a mixture of two normals and to a mixture of three normals, respectively. For the latter we just used a second order kernel (R = 1). The bandwidth sequence is[18] $c_N = c \cdot N^{-1/[2(R+1)+1]}$, where the optimal constant c^* is obtained by

16. The problem could have been solved by the inclusion of time dummies in the main equation.
17. For details on the latter, see HOROWITZ [1992], KYRIAZIDOU [1994] and CHARLIER *et al.* [1995].
18. By following the maximum rates of convergence in distribution for the univariate case according to BIERENS [1987].
19. The bias correction removes only asymptotic bias, so the bias-corrected estimator needs not be unbiased in finite samples. According to the corollary in KYRIAZIDOU's [1997], to construct the bias corrected estimator we have to compute another estimator with window width $c_{N,\delta} = c \cdot N^{-\delta/[2(R+1)+1]}$. We select $\delta = 0.5$.

TABLE 1

Invalidating the Exchangeability Assumption in KYRIAZIDOU

U1= 0.8 *N(0.1) ; U2 = 2*N(0,1) ; ε 1 = 0.1*U1 – 5 + 0.6*N(0.1) ; ε 2 = 0.9*U2+ 0.6*N(0.1);

PANEL A

Ignoring Correction for Sample Selection

N	Mean Bias	Median Bias	SE	ASE	MAD
250	0.1608	0.1589	0.2223	0.1437	0.1668
500	0.1908	0.2007	0.2147	0.1021	0.2007
750	0.1813	0.1867	0.1978	0.0831	0.1867
1000	0.1644	0.1709	0.1807	0.0718	0.1709

PANEL B

	WOOLDRIDGE's Estimator					**More Parametric New Estimator**					**Less Parametric New Estimator**				
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	-0.0203	0.0061	0.1491	0.1601	0.1000	-0.0179	-0.0155	0.1941	0.1689	0.1382	0.0636	0.0502	0.1958	0.1658	0.1282
500	0.0061	-0.0029	0.1166	0.1110	0.0836	0.0250	0.0425	0.1391	0.1188	0.0936	0.0649	0.0795	0.1513	0.1218	0.1021
750	0.0033	-0.0051	0.0985	0.0915	0.0754	0.0123	0.0068	0.0868	0.0994	0.0582	0.0443	0.0527	0.1048	0.1028	0.0801
1000	-0.0004	0.0109	0.0924	0.0785	0.0648	-0.0047	-0.0073	0.0996	0.0844	0.0717	0.0225	0.0197	0.1041	0.1262	0.0620

PANEL C

KYRIAZIDOU's Estimator

	Estimated First Step Parameters					**True First Step Parameters**													
	***R* = 1**					***R* = 1**					***R* = 3**					***R* = 5**			
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.05	0.06	0.24	0.19	0.20	0.04	0.06	0.28	0.18	0.17	0.06	0.07	0.22	0.17	0.15	0.07	0.07	0.21	0.17	0.13
500	0.07	00.5	0.31	0.14	0.10	0.06	0.07	0.20	0.14	0.11	0.09	0.10	0.18	0.13	0.12	0.11	0.11	0.18	0.13	

the *plug-in* method described in HÄRDLE and LINTON [1994] with an initial $c = 1$. In both cases, we present the bias corrected estimator. [19] For our less parametric new estimator (LPNE) we also used second order kernels. [20] The first step probabilities $h_1(z_i)$and $h_2(z_i)$ are estimated by *leave-one-out* kernel estimators (this is theoretically convenient) constructed as in section 3 but without z_i being used in estimating $\widehat{h}_\tau(z_i)$. The summation in (3.2) should read $j \neq i$. The bandwidth sequence for the LPNE is[21] $c_N = c \cdot N^{-1/[2(R+1)+T\cdot f]}$. The constant part of the bandwidth was chosen equal to 1. There was no serious attempt at optimal choice but we avoided values which entailed extreme bias or variability.

From Table 1 we see that all the estimators are less biased than the estimator ignoring correction for sample selection. KYRIAZIDOU's [1997] estimator shows a bias that does not generally go away with sample size. Furthermore, the estimator becomes quite imprecise if we look at the standard errors. The bias are all positive and increase a bit as the kernel order increases. Standard errors are always worse than the asymptotic ones. As can be seen in Panel B, both versions of our proposed estimator behave quite well in terms of all the considered measures. Both our estimator and WOOLDRIDGE's [1995] are robust to violations of the *conditional exchangeability* assumption. The relative SE's and MAD's of WOOLDRIDGE and the more parametric new estimator (MPNE) illustrate the efficiency gains or losses associated with the use of the semiparametric components employed in the LPNE. The LPNE has a larger finite-sample bias than WOOLDRIDGE's [1995] estimator and the MPNE, but this bias decreases with sample size. We do not need extremely large samples for our estimators to achieve reasonable agreement between the finite-sample standard errors and the results of asymptotic theory. At this stage, it is important to notice that for the experiments in Tables 1, 2, and 5, we observe some anomalous results in the ASE for the LPNE. Specifically, the estimated ASE for sample size equal to 1000 seem to be too high both with respect to the SE and in relation to their own evolution as sample size grows. The complexity of the variance-covariance matrix for the LPNE (see Appendix II), which estimate involves derivatives of the kernel functions, advises a more careful treatment in a particular application. Without further research, our intuition points at the appearance of anomalous

---

20. Even if for theoretical reasons it is sometimes useful to consider kernels that take on negative values (kernels of order higher than 2), in most applications K is a positive probability density function. In the Monte Carlo experiments we restrict our attention to second order kernel estimators. The reasons to support this decision are as follows. First, the results of MARRON and WAND [1992] advise caution against the application of higher order kernels unless quite large sample sizes are available because the merits of bias reduction methods are based on asymptotic approximations. Second, higher order kernels were generating some estimates for $\widehat{h}_\tau(z_i)$ not in between zero and one for some τ and i, so that the inverses $\Phi^{-1}[\widehat{h}_\tau(z_i)]$ did not exist. Solutions like the accumulation of these individuals in the corresponding extremes had severe consequences for the estimate of the asymptotic variance covariance matrix, that relies on derivatives of the functions $\Phi^{-1}[\widehat{h}_\tau(z_i)]$.

21. By following the best uniform consistency rate in BIERENS [1987] for multivariate kernels. If we were focused on convergence in distribution the optimal rate would have been obtained by setting $c_N = c \cdot N^{-1/[2(R+1)+T\cdot f]}$.

TABLE 2

Keeping the Exchangeability Assumption in KYRIAZIDOU and Generating a Misspecification Problem for WOOLDRIDGE

$$U = N(0,1);$$

$$\varepsilon = 0.8*U + 0.6*N(0,1);$$

$$\alpha = (X1+X2)/2+(X1^2+X2^2)/2+\sqrt{2}*N(0,1)+1;$$

PANEL A

Ignoring Correction for Sample Selection

N	Mean Bias	Median Bias	SE	ASE	MAD
250	0.1227	0.1016	0.1644	0.1070	0.1022
500	0.1108	0.1187	0.1363	0.0768	0.1187
750	0.1177	0.1215	0.1327	0.0636	0.1215
1000	0.1085	0.1159	0.1183	0.0542	0.1159

PANEL B

	WOOLDRIDGE's Estimator					**More Parametric New Estimator**					**Less Parametric New Estimator**				
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.3394	0.3664	0.3744	0.1555	0.3664	0.0057	0.0024	0.1517	0.1223	0.0848	0.0335	0.0187	0.1237	0.1257	0.0909
500	0.3496	0.3587	0.3671	0.1126	0.3587	0.0158	0.0180	0.1111	0.0876	0.0820	0.0195	0.0290	0.1089	0.0910	0.0818
750	0.3456	0.3557	0.3584	0.0943	0.3557	0.0042	0.0028	0.0777	0.0713	0.0550	0.0062	0.0100	0.0742	0.0789	0.0464
1000	0.3427	0.3477	0.3515	0.0821	0.3477	0.0010	0.0010	0.0660	0.0627	0.0422	-0.0029	-0.0027	0.0690	0.1017	0.0455

observations for the calculus of the ASE as sample size increases. Therefore, a trimming solution is called for.

In Table 2 we generate a misspecification problem for WOOLDRIDGE's [1995] estimator. The linear projection assumption for the individual effects in the main equation has been violated. As can be seen in the top part of that table we have generated the true α_i's by adding to our benchmark specification quadratic terms on the x's. Under this design WOOLDRIDGE's [1995] estimator is clearly inconsistent and it suffers from a misspecification bias problem. Both versions of our estimator are robust against any type of design for the individual effects in the main equation. As the estimation method is based on time-differences, its properties are independent of the particular shape for the individual effects in that equation. The MPNE and the LPNE are well behaved in terms of all the considered measures. The results for KYRIAZIDOU's [1997] estimator have not been included because this method is also independent of the particular shape of α_i.

Table 3 varies the standard sample sizes with respect to the other tables and incorporates a different design for the experiment. We have a new error structure for the errors in the main equation and dependent data over time has been introduced through the correlation of the variables in period 2 with the variables in period 1. Both estimators perform well with the new type of explanatory variables.

Table 4 presents results under a different design for the individual effects in the selection equation. We expect the LPNE to perform better than the MPNE. The reason is that the former allows for an unrestricted conditional mean of η_i. The MPNE was developed under the assumption of a linear conditional mean. By invalidating this linearity assumption we are generating a misspecification problem for the first step of the MPNE. The results in Table 4 confirm our prior. This holds for all the considered measures.

Finally, in Tables 5 and 5' we compare WOOLDRIDGE's [1995] estimator, KYRIAZIDOU's [1997] estimator and the proposed MPNE and LPNE when the joint conditional normality assumption (assumption A3 in section 2) breaks down. Table 5 reports the results when normalised and central χ^2 distributions with 2 degrees of freedom are considered. In Table 5' we adopt uniform distributions normalised to have mean 0 and unit variance. By looking at the estimates ignoring sample selection we see that the bias induced by sample selection is bigger in the case of uniformly distributed errors. Both uniform or χ^2 distributions do not seem to affect too badly our proposed estimators in relation to WOOLDRIDGE's [1995] and KYRIAZIDOU's [1997] estimators. WOOLDRIDGE's [1995] estimator does not need joint normality for the errors in both equations. It is sufficient to have a marginal normality for the errors in the selection equation and a linear projection of the errors in the main equation on the errors in the selection equation. Usually, it is the case that for sample selection models, and in terms of robustness of the estimators against misspecification of the error distribution, it is more critical the normality assumption for the errors in the main equation than in the selection rules. For example, it is known that in HECKMAN's [1976,1979] two-stage estimator the first-stage probit seems to be rather robust to violations of the normality assumption. As the results in our experiments are conditional to a sample selection problem design that comes through a linear projection of the errors in the main equation on the errors in the selection equation, the MPNE and

TABLE 3
Dependent Data

$$U = N(0,1);$$
$$\varepsilon = 0.6*U + 0.8*N(0,1);$$
$$Z2 = 0.7*Z1 + N(0,1);$$

PANEL A

Ignoring Correction for Sample Selection					
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.0799	0.0628	0.1601	0.1340	0.0950
500	0.0728	0.0694	0.1189	0.0962	0.0874
1000	0.0816	0.0726	0.1026	0.0679	0.0749
2000	0.0812	0.0669	0.0952	0.0483	0.0669

PANEL B

More Parametric New Estimator						Less Parametric New Estimator				
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.0071	0.0035	0.1593	0.1524	0.1047	0.0286	0.0024	0.1705	0.1586	0.1108
500	0.0105	0.0095	0.1248	0.1078	0.0969	0.0140	0.0039	0.1215	0.1137	0.0739
1000	0.0105	-0.0067	0.0833	0.0778	0.0478	0.0043	0.0049	0.0777	0.0847	0.0508
2000	-0.0038	-0.0044	0.0494	0.0553	0.0347	-0.0073	-0.0056	0.0554	0.0586	0.0359

TABLE 4
A Misspecification Problem in the MPNE

$$U = N(0,1);$$
$$\varepsilon = 0.8*U + 0.6*N(0,1);$$
$$\eta = -(Z11^2 * Z12^2) + (Z21^2 * Z22^2) - N(0,1);$$

PANEL A

Ignoring Correction for Sample Selection					
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.1246	0.1120	0.1932	0.1271	0.1366
500	0.1193	0.1173	0.1630	0.0893	0.1208
750	0.1200	0.1179	0.1382	0.0750	0.1179
1000	0.1258	0.1241	0.1409	0.0644	0.1241

PANEL B

More Parametric New Estimator						Less Parametric New Estimator				
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.0406	0.0372	0.1496	0.1445	0.1013	0.0426	0.0325	0.1667	0.1519	0.1216
500	0.0577	0.0504	0.1322	0.0998	0.0823	0.0242	0.0154	0.1240	0.1049	0.0776
750	0.0412	0.0545	0.0942	0.0834	0.0693	0.0108	0.0070	0.0859	0.1022	0.0644
1000	0.0419	0.0466	0.0801	0.0714	0.0525	0.0086	0.0132	0.0748	0.0756	0.0487

TABLE 5

Breaking the Normality Assumption with χ^2 Distributions

$U = \chi^2_2(0.1)$; $\varepsilon = 0.8*U + 0.6*\chi^2_2(0,1)$; $\alpha = (X1 + X2)/2 + \sqrt{2}*\chi^2_2(0,1)+1$; $\eta = -[(Z11+Z12)/2 + (Z21+Z22)/2 + \chi^2_2(0,1) + 0.07]$;

PANEL A

Ignoring Correction for Sample Selection

N	Mean Bias	Median Bias	SE	ASE	MAD
250	0.0859	0.0747	0.1344	0.0920	0.0857
500	0.0788	0.0858	0.1083	0.0671	0.0872
750	0.0730	0.0739	0.0887	0.0546	0.0739
1000	0.0722	0.0782	0.0856	0.0474	0.0782

PANEL B

	WOOLDRIDGE's Estimator					**KYRIAZIDOU's Estimator**																			
	Estimated First Step Parameters					**True First Step Parameters**																			
	R=1					**R=1**					**R=3**					**R=5**									
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.02	0.02	0.12	0.12	0.09	0.04	0.05	0.19	0.12	0.11	0.02	0.01	0.17	0.12	0.11	0.04	0.03	0.14	0.11	0.10	0.04	0.03	0.14	0.11	0.09
500	0.02	0.02	0.10	0.08	0.08	0.02	0.04	0.11	0.09	0.08	0.04	0.05	0.12	0.09	0.09	0.05	0.06	0.10	0.08	0.08	0.05	0.05	0.10	0.08	0.08
750	0.01	0.004	0.07	0.07	0.04	0.01	0.01	0.09	0.07	0.06	0.02	0.03	0.10	0.08	0.06	0.04	0.04	0.08	0.07	0.05	0.04	0.04	0.07	0.07	0.05
1000	0.02	0.02	0.06	0.06	0.04	0.04	0.05	0.09	0.07	0.07	0.02	0.03	0.09	0.06	0.06	0.03	0.03	0.07	0.06	0.05	0.04	0.04	0.07	0.06	0.05

PANEL C

	More Parametric New Estimator					**Less Parametric New Estimator**				
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.0205	0.0298	0.1122	0.1054	0.0795	0.0373	0.0355	0.1266	0.1147	0.0870
500	0.0304	0.0279	0.0947	0.0782	0.0600	0.0281	0.0259	0.0876	0.0810	0.0620
750	0.0152	0.0190	0.0618	0.0639	0.0428	0.0116	0.0161	0.0578	0.0657	0.0413
1000	0.0314	0.0360	0.0653	0.0543	0.0427	0.0168	0.0203	0.0615	0.1281	0.0428

TABLE 5'

Breaking the Normality Assumption with Uniform Distributions

U = Uniform (0,1) ; ε = 0.8*U +0.6* Uniform (0,1) ; α = (X1+ X2)/2 + $\sqrt{2}$* Uniform (0,1)+1 ; η = -[(Z11+Z12)/2+ (Z21+Z22)/2 + Uniform (0,1) + 0.07] ;

PANEL A

Ignoring Correction for Sample Selection

N	Mean Bias	Median Bias	SE	ASE	MAD
250	0.1023	0.0914	0.1482	0.1082	0.0945
500	0.1278	0.1299	0.1461	0.0769	0.1299
750	0.1214	0.1294	0.1320	0.0612	0.1294
1000	0.1250	0.1273	0.1358	0.0538	0.1273

PANEL B

	WOOLDRIDGE's Estimator					**KYRIAZIDOU's Estimator**																				
	Estimated First Step Parameters					**True First Step Parameters**																				
	R=1					**R=1**					**R=3**					**R=5**										
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	0.001	-0.0001	0.15	0.13	0.10	0.05	0.07	0.19	0.14	0.13	0.05	0.06	0.18	0.14	0.10	0.07	0.05	0.16	0.13	0.09	0.07	0.05	0.15	0.13	0.10
500	0.01	0.01	0.10	0.09	0.07	0.02	0.02	0.14	0.10	0.08	0.01	0.01	0.13	0.11	0.09	0.05	0.05	0.11	0.10	0.08	0.06	0.06	0.11	0.10	0.08
750	0.003	0.005	0.08	0.08	0.05	0.02	0.03	0.12	0.09	0.07	0.03	0.03	0.11	0.09	0.07	0.06	0.06	0.10	0.08	0.07	0.07	0.07	0.10	0.08	0.07
1000	0.005	0.009	0.06	0.07	0.04	0.03	0.05	0.12	0.08	0.08	0.03	0.05	0.11	0.08	0.08	0.06	0.07	0.10	0.07	0.09	0.07	0.08	0.10	0.07	0.08

PANEL C

	More Parametric New Estimator					**Less Parametric New Estimator**				
N	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**ASE**	**MAD**
250	-0.0047	0.0023	0.1557	0.1253	0.1121	0.0076	0.0072	0.1510	0.1293	0.1061
500	0.0034	0.0003	0.0817	0.0897	0.0470	0.0184	0.0220	0.0890	0.0918	0.0488
750	0.0030	0.0043	0.0616	0.0719	0.0413	0.0123	0.0201	0.0684	0.0758	0.0469
1000	-0.0048	-0.0008	0.0615	0.0626	0.0421	0.0049	0.0083	0.0615	0.0659	0.0466

the LPNE do not need a trivariate normal distribution for the errors in both equations but just a bivariate normal distribution for the errors in the selection equation. As a result, it may be the case that invalidating joint normality (given linearity) does not have strong consequences for WOOLDRIDGE's [1995] and our proposed estimators. We defer for future research to look at the effects of breaking down at the same time linearity and normality assumptions. KYRIAZIDOU's [1997] estimator is a distributionally free method and therefore it is robust to any distributional assumption that preserves the *conditional exchangeability* assumption. It is fair to say that we will probably need larger sample sizes than the ones included in our experiments to exploit the properties of this estimator. The sample size in WOOLDRIDGE's [1995] estimator is given by the observability rule $d_{it} = 1$; our proposed methods use individuals with $d_{it} = d_{is} = 1$; and KYRIAZIDOU's [1997] estimator uses individuals with $d_{it} = d_{is} = 1$ and $z_{it}\gamma \cong z_{is}\gamma$ in equation (2.2). Thus, the latter method uses the smallest sample size of all the methods.

6 Concluding Remarks and Extensions

In this paper we are concerned with the estimation of a panel data sample selection model where both the binary selection indicator and the regression equation of interest contain unobserved individual-specific effects that may depend on the observable explanatory variables. For this case, not many estimators are available. The most recent ones are the estimators developed by WOOLDRIDGE [1995] and KYRIAZIDOU [1997]. We introduce an estimator that can be seen as complementary to those previously suggested, in the sense that it uses an alternative set of identifying restrictions to overcome the selection problem. In particular, the estimator imposes that the joint distribution of the error terms, conditional upon the entire vector of (strictly) exogenous variables, is normal. The estimation procedure is an extension of HECKMAN's [1976,1979] sample selection technique to the case where one correlated selection rule in two different time periods generates the sample. We present two versions of the estimator depending on a varying degree of parametric assumptions for the first step estimator.

The finite sample properties of the estimator are investigated by Monte Carlo experiments. The results of our small Monte Carlo simulation study show the following. First, the estimator is robust to violations of the *conditional exchangeability* assumption in KYRIAZIDOU's [1997] method. Second, the estimator is free from misspecification problems affecting the individual effects in the main equation, in contrast to WOOLDRIDGE's [1995] one. Furthermore, under its less parametric version, the estimator is also exempt from misspecification problems about the individual effects in the sample selection equation. Third, the estimator performs well with dependent data introduced through correlation over time for the variables in the model.

Finally, violations of the normality assumption (given linearity) do not seem to affect too badly the proposed estimator.

Our analysis rests on the strict exogeneity of the explanatory variables in both equations, although it would be possible to relax this assumption in the main equation by maintaining only the strict exogeneity of the regressors in the selection equation and taking an instrumental variables approach. We also maintain a joint normality assumption. [22] We defer for future research to look at the effects of breaking down at the same time linearity and normality assumptions. More research is also needed in the search for trimming solutions to overcome the anomalous effect of particular observations in the estimates of the variance-covariance matrix for semiparametric two and three-stage estimators.

• References

ARELLANO M., CARRASCO R., (1996). – "Binary Choice Panel Data Models with Predetermined Variables", CEMFI *Working Paper*, N° 9618.

BERNDT E., HALL B., HALL R., HAUSMAN J., (1974). – "Estimation and Inference in Non-Linear Structural Models" *Annals of Economic and Social Measurement*, 3/4, pp. 653-665.

BIERENS H.J., (1987). – "Kernel Estimators of Regression Functions", in *Advances in Econometrics*, Fifth World Congress, Volume I, Econometric Society Monographs, N° 13 Ed. T F. Bewley, Cambridge University Press.

CHAMBERLAIN G., (1980). – "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, XLVII, pp. 225-238.

CHAMBERLAIN G., (1982). – "Multivariate Regression Models for Panel Data", *Journal of Econometrics*, 18, pp. 5-46.

CHAMBERLAIN G., (1984). – "Panel Data", in Z. Griliches and M. Intriligator, Eds., *Handbook of Econometrics*, Volume II, North-Holland Publishing CO, Amsterdam, Ch.22.

CHARLIER E., MELENBERG B., VAN SOEST A.H.O. (1995). – "A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model with an Application to Labour Force Participation", *Statistica Nederlandica*, 49, pp. 324-342.

HAM J.C., (1982). – "Estimation of a Labour Supply Model with Censoring Due to Unemployment and Underemployment", *Review of Economic Studies*, XLIX, pp. 335-354.

HÄRDLE W., (1990). – *Applied Nonparametric Regression*, Cambridge University Press.

HÄRDLE W., LINTON O., (1994). – "Applied Nonparametric Methods", in R. F. Engle and D. L. McFadden, Eds., *Handbook of Econometrics*, Volume IV, Elsevier Science.

HÄRDLE W., CHEN R., (1995). – "Nonparametric Time Series Analysis, a Selective Review with Examples", *Discussion Paper*, 14, Humboldt-Universität zu Berlin.

HECKMAN J., (1976). – "The Common Structure of Statistical Models of Truncation, and Simple Estimates for Such Models", *Annals of Economics and Social Measurement*, 15, pp. 475-492.

HECKMAN J., (1979). – "Sample Selection Bias as a Specification Error", *Econometrica*, 47, pp.153-161.

HOROWITZ J., (1992). – "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, pp. 505-531.

HOROWITZ J., (1998). – *Semiparametric Methods in Econometrics*, lecture notes in statis-

22. The exploration of the possibility of relaxing normality or any orther parametric assumption for the errors distributions is being currently undertaken by the author.

tics-131, Springer.

KYRIAZIDOU E., (1994). – "Estimation of a Panel Data Sample Selection Model", *unpublished manuscript*, Northwestern University.

KYRIAZIDOU E., (1997). – "Estimation of a Panel Data Sample Selection Model", *Econometrica*, Vol. 65, N°6, pp. 1335-1364.

LEE L. F. (1982). – "Some Approaches to the Correction of Selectivity Bias", *Review of Economic Studies*, XLIX, pp. 355-372.

LEE M.J., (1996). – *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*, Springer.

MANSKI C., (1987). – "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, pp. 357-362.

MARRON J.S., WAND M.P. (1992). – "Exact Mean Integrated Squared Error", *Annals of Statistics*, 20, pp. 712-736.

MUNDLAK, Y. (1978). – "On the Pooling of Time Series and Cross-Section Data", *Econometrica*, 46, pp. 69-85.

NADARAYA, E.A. (1964). – "On Estimating Regression", *Theory Prob. Appl*., 10, pp.186-190.

NEWEY, W.K. (1992). – "Partial Means, Kernel Estimation, and a General Asymptotic Variance Estimator", *Mimeo*, MIT.

NEMEY, W.K. (1994a). – "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, Vol. 62, N° 6, pp. 1349-1382.

NEWEY, W.K., (1994b). "Kernel Estimation of Partial Means and a General Variance Estimator", *Econometric Theory*, 10, pp. 233-253.

NEWEY, W.K., MCFADDEN D. (1994c). "Large Sample Estimation and Hypothesis Testing", in R. F. Engle and D.L. McFadden, Eds., *Handbook of Econometrics*, Volume IV, Elsevier Science.

NIJMAN, T. VERBEEK M. (1992). – "Nonresponse in Panel Data: the Impact on Estimates of a Life Cycle Consumption Function", *Journal of Applied Econometrics*, 7, pp. 243-257.

POIRIER, D.J. (1980). – "Partial Observability in Bivariate Probit Models", *Journal of Ecometrics*, 12, pp. 209-217.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

TALLIS, G.M. (1961), "The Moment Generating Function of the Truncated Multi-Normal Distribution", *Journal of the Royal Statistical Society*, 23, Series b, pp. 223-229.

VERBEEK, M., NIJMAN T. (1992). – "Testing for Selectivity Bias in Panel Data Models", *International Economic Review*, 33, pp. 681-703.

WATSON, G.S. (1964), "Smooth Regression Analysis", *Sankhya* Series A, 26 pp. 359-372.

WOOLDRIDGE, J.M. (1995). – "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions", *Journal of Econometrics*, 68 pp. 115-132.

ZABEL, J.E. (1992). "Estimating Fixed Effects and Random Effects with Selectivity", *Economics Letters*, 40, pp. 269-272.

APPENDIX I

The Variance-Covariance Matrix for the More Parametric New Estimator

Recall (2.7) : with $(y_{it} - y_{is}) = (x_{it} - x_{is})\ \beta + (\varepsilon_{it} - \varepsilon_{is})$ and $d_{i\tau} = 1\{z_i\gamma_\tau - v_{i\tau} \geqslant 0\}$ for $\tau = t,s$,

$$E[(y_{it} - y_{is}) \,|x_i, z_i, v_{it} \leqslant z_i\gamma_t, v_{is} \leqslant z_i\gamma_s] = (x_{it} - x_{is})\beta + \ell_{ts} \cdot \lambda_{its} + \ell_{st} \cdot \lambda_{ist}$$

$$\lambda_{its} = \frac{\phi[z_i\gamma_t] \cdot \Phi\left[\dfrac{z_i\gamma_s - \rho_{ts} \cdot z_i\gamma_t}{(1-\rho_{ts}^2)^{1/2}}\right]}{\Phi_2[z_i\gamma_t, z_i\gamma_s, \rho_{ts}]},$$

$$\lambda_{ist} = \frac{\phi[z_i\gamma_s] \cdot \Phi\left[\dfrac{z_i\gamma_t - \rho_{ts} \cdot z_i\gamma_s}{(1-\rho_{ts}^2)^{1/2}}\right]}{\Phi_2[z_i\gamma_t, z_i\gamma_s, \rho_{ts}]}$$

The two-stage estimation goes as follows. First, we estimate $\overline{\omega}_{ts} = (\gamma_t', \gamma_s', \rho_{ts})'$ by $\widehat{\overline{\omega}}_{ts} = (\widehat{\gamma}_t', \widehat{\gamma}_s', \widehat{\rho}_{ts})'$ using a bivariate probit with observations on (d_{it}, d_{is}, z_i). Second, for the subsample with $d_{it} = d_{is} = 1$, we do least squares estimation of $\Delta y_{its} = (y_{it} - y_{is})$ on $\Delta x_{its} = (x_{it} - x_{is})$ and $(\widehat{\lambda}_{its}, \widehat{\lambda}_{ist})$ to estimate the parameter of interest, β, and the coefficients accompanying the sample selection terms $(\ell_{ts}, \ell_{st})'$. Define $R_{its} \equiv (\Delta x_{its}, \lambda_{its}, \lambda_{ist})'$. The sample moment condition for $\widehat{\beta}$ and $(\widehat{\ell}_{ts}, \widehat{\ell}_{st})'$ in the second stage is

$$\textbf{(I.1)} \quad \frac{1}{N}\sum_i d_{it}d_{is}\{\Delta y_{its} - \Delta x_{its}\widehat{\beta} - \widehat{\ell}_{ts} \cdot \widehat{\lambda}_{its} - \widehat{\ell}_{st} \cdot \widehat{\lambda}_{ist}\}R_{its} = 0;$$

this is the first order condition of a two stage *extremum estimator* with finite dimensional first stage parameters.

Define

23. The notation $=^p$ denotes convergence in probability.

$$\widehat{\Pi}_{ts} \equiv (\widehat{\beta}', \widehat{\ell}_{ts}, \widehat{\ell}_{st})'$$
$$\Pi_{ts} \equiv (\beta', \ell_{ts}, \ell_{st})'$$
$$e_{its} \equiv \Delta y_{its} - \Delta x_{its}\beta - \ell_{ts} \cdot \lambda_{its} - \ell_{st} \cdot \lambda_{ist},$$

and observe[23]

(I.2)

$$\sqrt{N}(\widehat{\overline{\omega}}_{ts} - \overline{\omega}_{ts}) = {}^{p}\frac{1}{\sqrt{N}} \sum_i I_{\overline{\omega}_{ts}}^{-1} \cdot \begin{bmatrix} z_i'(q_{it}\phi_{it}\,\Phi_{its}/\phi_{2,its}) \\ z_i'(q_{is}\phi_{is}\,\Phi_{ist})/\phi_{2,its}) \\ q_{it}q_{is}\phi_{2,its}/\Phi_{2,its} \end{bmatrix} \equiv \frac{1}{\sqrt{N}} \sum_i \Lambda_i,$$

where $I_{\overline{\omega}_{ts}}$ is the bivariate probit information matrix for $\overline{\omega}_{ts}$,

$$\phi_{it} \equiv \phi[q_{it}z_i\gamma_t], \qquad \phi_{is} \equiv \phi[q_{is}z_i\gamma_s],$$
$$\Phi_{its} \equiv \Phi[(q_{is}z_i\gamma_s - \rho^*_{its}q_{it}z_i\gamma_t)/(1-\rho^{*2}_{its})^{1/2}],$$
$$\Phi_{ist} \equiv \Phi[(q_{it}z_i\gamma_t - \rho^*_{its}q_{is}z_i\gamma_s)/(1-\rho^{*2}_{its})^{1/2}],$$
$$\Phi_{2,its} \equiv \Phi_2\{q_{it}z_i\gamma_t, q_{is}z_i\gamma_s, \rho^*_{its}\} \text{ and } \phi_{2,its} \equiv \phi_2\{q_{it}z_i\gamma_t, q_{is}z_i\gamma_s, \rho^*_{its}\}.$$

q_{it}, q_{is}, and ρ^*_{its} are defined in section 3 of the paper.

The so called *delta method* yields[24]

(I.3)

$$\sqrt{N}(\widehat{\Pi}_{ts} - \Pi_{ts}) = {}^{p}\ E^{-1}(d_t d_s R_{ts} R_{ts}') \cdot \frac{1}{\sqrt{N}} \sum_i \{d_{it}d_{is}e_{its}R_{its} + A \cdot \Lambda_i\}$$

where

(I.4)

$$A \equiv E\left\{ d_t d_s \left[\left(-\ell_{ts}\frac{\partial \lambda_{ts}}{\partial z\gamma_t} - \ell_{st}\frac{\partial \lambda_{st}}{\partial z\gamma_t} \right) R_{ts} z \right.\right.$$
$$\left. \left(-\ell_{ts}\frac{\partial \lambda_{ts}}{\partial z\gamma_s} - \ell_{st}\frac{\partial \lambda_{st}}{\partial z\gamma_s} \right) R_{ts} z \quad \left(-\ell_{ts}\frac{\partial \lambda_{ts}}{\partial \rho_{ts}} - \ell_{st}\frac{\partial \lambda_{st}}{\partial \rho_{ts}} \right) R_{ts} \right] \Bigg\}$$

Then

(I.5)

$$\sqrt{N}(\widehat{\Pi}_{ts} - \Pi_{ts}) = {}^{d}\ N(0, \Gamma),$$
$$\Gamma = E^{-1}(d_t d_s R_{ts} R_{ts}') \cdot E\{(d_t d_s e_{ts} R_{ts} + A \cdot \Lambda)(d_t d_s e_{ts} R_{ts} + A \cdot \Lambda)'\}$$
$$\cdot E^{-1}(d_t d_s R_{ts} R_{ts}')$$

The term $A \cdot \Lambda$ is the effect of the first stage on the second. An estime for Γ is obtained by replacing the parameters with their estimates and the expectations by their sample analogs. As the Fisher information matrix in (I.2) contains the negatives of the *expected* values of the second derivatives, the complexity of the second derivatives in this case makes it an excellent candidate for the BERNDT *et al.* [1974] estimator of the inverse of the Fisher information matrix. This yields :

24. Look at the section for two-stage *extremum estimators* with finite dimensional first-stage nuisance parameters in LEE [1996].

(I.6)

$$\widehat{I}_{\overline{\omega}_{ts}}^{-1} = \left\{ \frac{1}{N} \sum_i \begin{bmatrix} z_i'(q_{it}\widehat{\phi}_{it}\widehat{\Phi}_{its}/\widehat{\Phi}_{2,its} \\ z_i'(q_{is}\widehat{\phi}_{is}\widehat{\Phi}_{ist}/\widehat{\Phi}_{2,its} \\ q_{it}q_{is}\widehat{\phi}_{2,its}/\widehat{\Phi}_{2,its} \end{bmatrix} \cdot \begin{bmatrix} z_i'(q_{it}\widehat{\phi}_{it}\widehat{\Phi}_{its}/\widehat{\Phi}_{2,its} \\ z_i'(q_{is}\widehat{\phi}_{is}\widehat{\Phi}_{ist}/\widehat{\Phi}_{2,its} \\ q_{it}q_{is}\widehat{\phi}_{2,its}/\widehat{\Phi}_{2,its} \end{bmatrix}' \right\}^{-1}.$$

APPENDIX II

The Variance-Covariance Matrix for the Less Parametric New Estimator

The three-stage semiparametric estimation goes as follows. First, we estimate the probabilities $E(d_{i\tau}|z_i)$ for $\tau = t,s$ by $\widehat{h}_\tau(z_i)$ using kernel estimators with observations on $(d_{i\tau},z_i)$. Second, we use the probability estimates to estimate ρ_{ts} by $\widehat{\rho}_{ts}$ using a bivariate probit with observations on $(d_{it},d_{is},\Phi^{-1}[\widehat{h}_t(z_i)],\Phi^{-1}[\widehat{h}_s(z_i)])$. Third, for the subsample with $d_{it} = d_{is} = 1$, we do least squares estimation of Δy_{its} on Δx_{its} and the estimated sample selection correction terms to estimate the parameters of interest, β, and the coefficients accompanying the sample selection terms $(\ell_{ts},\ell_{st})'$. The sample moment condition for $\widehat{\beta}$ and $(\widehat{\ell}_{ts},\widehat{\ell}_{st})'$ in the third stage is

(II.1)
$$\frac{1}{N}\sum_i d_{it}d_{is}\{\Delta y_{its} - \Delta x_{its}\widehat{\beta} - \widehat{\ell}_{ts}\cdot\widehat{\lambda}_{its} - \widehat{\ell}_{st}\cdot\widehat{\lambda}_{ist}\}R_{its} = 0;$$

where

$$\widehat{\lambda}_{its} = \frac{\phi[\Phi^{-1}(\widehat{h}_{it})]\cdot\Phi\left[\dfrac{\Phi^{-1}(\widehat{h}_{is}) - \widehat{\rho}_{ts}\cdot\Phi^{-1}(\widehat{h}_{it})}{(1-\widehat{\rho}_{ts}^2)^{1/2}}\right]}{\Phi_2[\Phi^{-1}(\widehat{h}_{it}),\Phi^{-1}(\widehat{h}_{is}),\widehat{\rho}_{ts}]},$$

$$\widehat{\lambda}_{ist} = \frac{\phi[\Phi^{-1}(\widehat{h}_{is})]\cdot\Phi\left[\dfrac{\Phi^{-1}(\widehat{h}_{it}) - \widehat{\rho}_{ts}\cdot\Phi^{-1}(\widehat{h}_{is})}{(1-\widehat{\rho}_{ts}^2)^{1/2}}\right]}{\Phi_2[\Phi^{-1}(\widehat{h}_{it}),\Phi^{-1}(\widehat{h}_{is}),\widehat{\rho}_{ts}]}$$

Once the first and the second step estimators have been consistently estimated, the third step estimator can be seen as another two stage semi-

25. They are infinite-dimensional because as $N \to \infty$ the number of terms also goes to ∞, given that the terms are individual specific and that the first step are functions rather than a finite-dimensional parameter.

parametric *extremum estimator* where the first stage estimators are given by the vector of infinite dimensional nuisance parameters [25] $\widehat{h}_{ts} = (\widehat{h}_t, \widehat{h}_s)'$ $(=^p h_{ts} = (h_t, h_s)')$ and the finite parameter $\widehat{\rho}_{ts} (=^p \rho_{ts})$. With this approach (II.1) is the first order condition of a two stage semiparametric *extremum estimator* with a combination of finite and infinite dimensional first stage parameters.

Observe that

$$\textbf{(II.2)}\quad \sqrt{N}(\widehat{\rho}_{ts} - \rho_{ts}) =^p \frac{1}{\sqrt{N}} \sum_i I_{\rho_{ts}}^{-1} \cdot \Big\{ (q_{it} q_{is} \phi_{2,its} / \Phi_{2,its}) + E\left[\partial \left(\frac{q_t q_s \phi_{2,ts}}{\Phi_{2,ts}} \middle| z_i \right) \Big/ \partial h_{ts}' \right] [d_{its} - E(d_{ts} \mid z_i)] \Big\} \equiv \frac{1}{\sqrt{N}} \sum_i \Lambda_i$$

where $I_{\rho_{ts}}$is the bivariate probit information matrix for ρ_{ts},

$$\Phi_{2,its} \equiv \Phi_2\{q_{it}\Phi^{-1}[\widehat{h}_t(z_i)], q_{is}\Phi^{-1}[\widehat{h}_s(z_i)], \widehat{\rho}^*_{its}\},$$
$$\phi_{2,its} \equiv \phi_2\{q_{it}\Phi^{-1}[\widehat{h}_t(z_i)], q_{is}\Phi^{-1}[\widehat{h}_s(z_i)], \widehat{\rho}^*_{its}\},$$

and $d_{its} = (d_{it}, d_{is})'$.

A *delta method* for an estimator with first step kernel estimators yields

$$\begin{aligned}\textbf{(II.3)}\quad \sqrt{N}(\widehat{\Pi}_{ts} - \Pi_{ts}) =^p\ & E^{-1}(d_t d_s R_{ts} R_{ts}') \\ & \cdot \frac{1}{\sqrt{N}} \sum_i \{ d_{it} d_{is} e_{its} R_{its} + A \cdot \Lambda_i \\ & + E[\partial \{ d_t d_s \{ \Delta y_{ts} - \Delta x_{ts}\beta - \ell_{ts} \cdot \lambda_{ts} - \ell_{st} \cdot \lambda_{st} \} R_{ts} | z_i \} / \partial h_{ts}'] \\ & \cdot [d_{its} - E(d_{ts} | z_i)] \}\end{aligned}$$

where

$$\textbf{(II.4)}\qquad A \equiv E\{ d_t d_s \left(-\ell_{ts} \frac{\partial \lambda_{ts}}{\partial \rho_{ts}} - \ell_{st} \frac{\partial \lambda_{st}}{\partial \rho_{ts}} \right) R_{ts} \}.$$

Then[26]

$$\textbf{(II.5)}\quad \sqrt{N}(\widehat{\Pi}_{ts} - \Pi_{ts}) =^d N(0, \Gamma),$$

$$\begin{aligned}\Gamma = {} & E^{-1}(d_t d_s R_{ts} R_{ts}') \\ & \cdot E\{ (d_t d_s e_{ts} R_{ts} + A \cdot \Lambda + E[\partial \{ d_t d_s \{ \Delta y_{ts} - \Delta x_{ts}\beta - \ell_{ts} \cdot \lambda_{ts} - \ell_{st} \cdot \lambda_{st} \} \\ & R_{ts} | z_i \} / \partial h_{ts}'] \cdot [d_{its} - E(d_{ts} | z_i)]) \\ & (d_t d_s e_{ts} R_{ts} + A \cdot \Lambda + E[\partial \{ d_t d_s \{ \Delta y_{ts} - \Delta x_{ts}\beta - \ell_{ts} \cdot \lambda_{ts} - \ell_{st} \cdot \lambda_{st} \} \\ & R_{ts} | z_i \} / \partial h_{ts}'] \cdot [d_{its} - E(d_{ts} | z_i)])' \} \cdot E^{-1}(d_t d_s R_{ts} R_{ts}')\end{aligned}$$

---

26. The result in (II.5) holds when a high-order kernel is used in the first stage, and the bandwidth is chosen to be smaller than the optimal bandwidth minimizing the asymptotic mean squared error; such a small bandwidth reduces the asymptotic bias faster than the optimal bandwidth. With a kernel estimator the result in (II.5) can be proved using high-order kernels, U-statistic theories, and the proper uniform consistency theorem.

The variance-covariance matrix should take into account the estimation errors coming directly by the effect of the kernel estimates on the sample selection correction terms and the effect of the $\widehat{\rho}_{ts}$ coefficient. For the latter, the influence function in (II.2) will already take into account the indirect effect of the estimation errors in the kernels on the sample selection correction terms through the estimated correlation coefficient. An estimate for Γ is generally obtained by replacing the parameters with their estimates and the expectations by their sample analogs. In our case, both

$$E[\partial\{d_t d_s\{\Delta y_{ts} - \Delta x_{ts}\beta - \ell_{ts}\cdot\lambda_{ts} - \ell_{st}\cdot\lambda_{st}\}R_{ts}|z_i\}/\partial h'_{ts}] \cdot[d_{its} - E(d_{ts}|z_i)]$$

in (II.3) and $E\left[\partial\left(\frac{q_t q_s \phi_{2,ts}}{\Phi_{2,ts}}|z_i\right)/\partial h'_{ts}\right]\cdot[d_{its} - E(d_{ts}|z_i)]$ in (II.2) are complex and difficult to calculate, making it hard to form their estimators.

(II.6)

$$\widehat{\vartheta}_i = \left.\frac{\partial}{\partial\zeta}\right|_{\zeta=0}\left\{\frac{1}{N}\sum_j \frac{q_{jt}q_{js}\phi_2\left\{q_{jt}\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], q_{js}\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], \widehat{\rho}^*_{jts}\right\}}{\Phi_2\left\{q_{jt}\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], q_{js}\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], \widehat{\rho}^*_{jts}\right\}}\right\}$$

There is an alternative estimator, developed in NEWEY [1992], that does not have these problems. [27] It uses only the form of the functions to derive and the kernels[28] to calculate the estimator. For a scalar ζ the estimator for $E[\partial\left(\frac{q_t q_s\phi_{2,ts}}{\Phi_{2,ts}}\middle|z_i\right)/\partial h'_{ts}]\cdot[d_{its} - E(d_{ts}|z_i)]$ is given by[29]

This estimator can be thought of as the influence of the i[th] observation through the kernel estimators. It can be calculated by either analytical or numerical differentiation with respect to the scalar ζ. We have combined both approaches. The derivative is then evaluated at $\zeta = 0$. Consistency is shown in NEWEY [1992].

27. See also NEWEY [1994b] and NEWEY and MCFADDEN [1994c].

28 . We have to make the decomposition of $h_\tau = g_\tau / f$, because NEWEY's [1992] results are given for first step estimators of the form $(1/N)\sum_i y_i(1/c_N^{T\cdot f})K((z_j - z_i)/c_N)$. A kernel estimator of the density of z_j will be a component of the expression before, where y is identically equal to 1.

29. In practice, we include $\widehat{\vartheta}_i - \sum_j \frac{\widehat{\vartheta}_j}{N}$.

30. In practice, we include $\widehat{\chi}_i - \sum_j \frac{\widehat{\chi}_j}{N}$.

NEWEY's [1992] estimator for

$$-\ell_{ts}\cdot\left\{\frac{\phi\left[\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right]\right]\cdot\Phi\left[\frac{\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right]-\widehat{\rho}_{ts}\cdot\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right]}{(1-\widehat{\rho}_{ts}^2)^{1/2}}\right]}{\Phi_2\left[\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right],\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right],\widehat{\rho}_{ts}\right]}\right\}$$

$$-\ell_{st}\cdot\left\{\frac{\phi\left[\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right]\right]\cdot\Phi\left[\frac{\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right]-\widehat{\rho}_{ts}\cdot\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right]}{(1-\widehat{\rho}_{ts}^2)^{1/2}}\right]}{\Phi_2\left[\Phi^{-1}\left[\frac{\widehat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right],\Phi^{-1}\left[\frac{\widehat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\widehat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right],\widehat{\rho}_{ts}\right]}\right\}$$

$$E[\partial\{d_t d_s\{\Delta y_{ts}-\Delta x_{ts}\beta-\ell_{ts}\cdot\lambda_{ts}-\ell_{st}\cdot\lambda_{st}\}R_{ts}|z_i\}/\partial h'_{ts}]\cdot[d_{its}-E(d_{ts}|z_i)]$$

is given by[30]

(II.7) $$\widehat{\chi}_i=\frac{\partial}{\partial\zeta}\bigg|_{\zeta=0}\frac{1}{N}\sum_j d_{jt}d_{js}R_{jts}\{\Delta y_{jts}-\Delta x_{jts}\beta$$

that is calculated by a mixture of analytical and numerical differentiation with respect to the scalar ζ. The derivative is then evaluated at $\zeta=0$.

As the Fisher information matrix in (II.2) contains the negatives of the expected values of the second derivatives, the complexity of the second derivatives in this case makes it an excellent candidate for the BERNDT *et al.* [1974] estimator of the inverse of the Fisher information matrix. This yields:

(II.8) $$\widehat{I}_{\rho_{ts}}^{-1} = \left\{ \frac{1}{N} \sum_{i} (q_{it} q_{is} \widehat{\phi}_{2,its} / \widehat{\Phi}_{2,its}) \cdot (q_{it} q_{is} \widehat{\phi}_{2,its} / \widehat{\Phi}_{2,its})' \right\}^{-1}$$