Panel Data
Sample Selection Models

María Engracia Rochina Barrachina

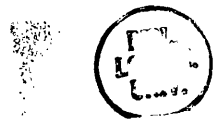
Thesis submitted for the Degree of

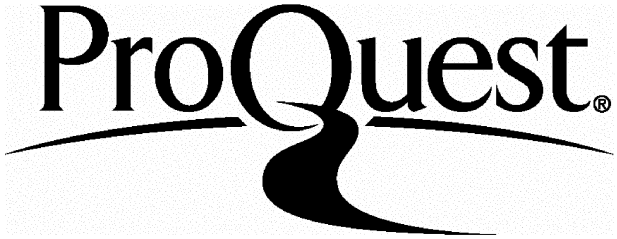
Doctor of Philosophy in Economics

University College London

University of London

2000





Preface

This thesis is the result of my research activities at the Department of Economics at University College London (UCL), at which I have been working as a Ph.D. student. I would like to thank some people and institutions that have supported me, and have somehow contributed to my work.

First of all, I would like to express my thanks to my parents, relatives and friends for their support and encouragement. I would specially like to mention my grandmother, which company I could only enjoy for one year more after my stay in London.

The supervision of both Richard Blundell and Costas Meghir is gratefully acknowledged.

Furthermore, I would like to thank all my fellow Ph.D. students and friends at UCL for their support, their company and the pleasant atmosphere they contributed to create. I owe special thanks to my friend and colleague Juan Alberto Sanchis-Llopis with whom I arrived to London, I moved out of London and I came back for submission of this thesis.

I would also like to thank several institutions for their support. UCL has contributed to my work with an stimulating research environment and the great possibility to attend seminars given by researchers of international prestige. Financial support from the Spanish foundation "Fundación Ramón Areces" is gratefully acknowledged. This institution made my Ph.D. possible and I would like to encourage the Foundation to continue with the priceless task of funding postgraduate

studies in foreign countries. I am very much indebted to my department of Applied Economics II at the University of Valencia, and specially to my "moral boss" there, José Antonio Martínez-Serrano, who started the policy of strongly encouraging young people to go abroad for postgraduate studies. Finally, many thanks to my colleagues and friends in this department.

María E. Rochina-Barrachina

Abstract

In this thesis estimators for "fixed-effects" panel data sample selection models are discussed, mostly from a theoretical point of view but also from an applied one. Besides the general introduction and conclusions (chapters 1 and 6, respectively) the thesis consists of four main chapters. In chapter 2 we are concerned about the finite sample performance of Wooldridge (1995) and Kyriazidou's (1997) estimators. Chapter 3 introduces a new estimator. The estimation procedure is an extension of the familiar two-steps sample selection technique to the case where one correlated selection rule in two time periods generates the sample. Some non-parametric components are introduced. We investigate the finite sample performance for the estimators in chapters 2 and 3 through Monte Carlo simulation experiments. In chapter 4 we apply the estimators in the previous chapters to estimate the return to actual labour market experience for females, using a panel of twelve years. All these estimators rely on the assumption of strict exogeneity of regressors in the equation of interest, conditional on individual specific effects and the selection mechanism. This assumption is likely to be violated in many applications. For instance, life history variables are often measured with error in survey data sets, because they contain a retrospective component. We show how non-strict exogeneity and measurement error can be taken into account within the methods. In chapter 5 we propose two semiparametric estimators under the assumption that the selection function depends on the conditional means of some observable variables. The first is a "weighted

double pairwise difference estimator" because it is based in the comparison of individuals in time differences. The second is a "single pairwise difference estimator" because only differences over time for a given individual are required. We investigate the finite sample properties of these estimators by Monte Carlo experiments.

Table of Contents

Chapter 1

Introduction

1.1 Panel Data and Sample Selection Models

In this thesis estimators for panel data sample selection models are discussed, mostly from a theoretical point of view but also from an applied one. The utilisation of panel data is commonly confronted with two problems, sample selectivity and unobserved heterogeneity, both of which give rise to specification bias. Sample selectivity arises in nonrandomly drawn samples, as a result of either self-selection by the individuals under investigation, or selection decisions made by data-analysts. As a consequence, in many problems of applied econometrics, the equation of interest is only defined for a subset of individuals from the overall population, while the parameters of interest are the parameters that refer to the whole population. Examples are the estimation of wage equations, or hours of work equations, where the dependent variable can only be measured when the individual participates in the labour market. Failure to account for sample selection is well known to lead to inconsistent estimation of the parameters of interest, as these are confounded with parameters that determine the probability of entry into the sample.

In contrast to sample selectivity, unobserved heterogeneity is a problem specific to panel data. Economic theory often suggests estimation equations that

contain an individual specific effect, which is unobserved, but correlated with the model regressors. Examples are unobserved ability components in wage equations, correlated with wages and education (see Card (1994) for details), or the estimation of Frisch demand functions in the consumption and labour supply literature (see, for instance, Browning, Deaton, and Irish (1985), Blundell and MaCurdy (1999) and MaCurdy (1981)). If unobserved individual specific (and time constant) effects affect the outcome variable, and are correlated with the model regressors, simple regression analysis does not identify the parameters of interest. In this thesis, if the individual effects are considered as nuisance parameters or if they are explicitly allowed to depend on the explanatory variables in a given or fully unrestricted way, then we call the panel data model a "fixed-effects" model.

In many applications with panel data, both sample selectivity and unobserved heterogeneity problems occur simultaneously. In this thesis we consider the problem of estimating panel data sample selection models with a binary selection equation. Both the sample selection rule and the regression equation of interest contain permanent unobservable individual effects possibly correlated with the explanatory variables.

The general model, which summarises the models along all the chapters in this thesis, can be written as follows,

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}; \quad i = 1,\ldots,N; \quad t = 1,\ldots,T, \tag{1.1}$$

$$d_{it}^{*} = f(z_{it}) - \eta_i - u_{it}; \quad d_{it} = 1\left[d_{it}^{*} \geq 0\right], \tag{1.2}$$

where x_{it} and z_{it} are vectors of explanatory variables (which may have components in common), $\beta \in \Re^k$ is an unknown parameter (column) vector, ε_{it} and u_{it} are unobserved disturbances, and α_i and η_i are individual-specific effects presumably correlated with the explanatory variables in the model. The index function $f(\cdot)$ in (1.2) is a scalar "aggregator" function which can accommodate different structures. In particular, we allow either for a linear parametric form of this function or an unrestricted one. Whether or not observations for y_{it} are available is denoted by the dummy variable d_{it}. 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}.

For the estimators considered in this thesis the individual effects α_i and η_i are treated as nuisance parameters or, alternatively, they are explicitly allowed to depend on the explanatory variables in a given or fully unrestricted way. Furthermore, each estimator imposes different stochastic restrictions for the error terms in the model. In this thesis we are interested in the estimation of the regression coefficients β in the model (1.1)-(1-2) above.

1.2 Contribution of This Thesis and Overview

The contribution of this thesis is to develop, apply, and learn about estimators for panel data sample selection models. Besides the theoretical approach to the estimators and the results from Monte Carlo simulations, applications can clarify the use of the estimators in practice.

This thesis consists of four main chapters. In the following, each of these chapters is discussed only briefly since each chapter is accompanied by its own introduction and conclusions. The emphasis is on the objectives and the interrelation between these chapters. The last chapter, chapter 6, provides a brief summary of the main results and conclusions from the various chapters.

In chapter 2, we examine Wooldridge (1995) and Kyriazidou's (1997) estimators for "fixed-effects" panel data sample selection models. The specification of the function $f(\cdot)$ in (1.2) is $f(z_{it}) = z_{it}\gamma$, where $\gamma \in \Re^f$ is an unknown parameter (column) vector. For Kyriazidou's (1997) estimator both α_i and η_t are treated as nuisance parameters. In Wooldridge (1995) they are explicitly allowed to depend on the leads and lags of the explanatory variables through a linear projection operator. Each estimation method relies on different stochastic restrictions for the error terms in the model. Particularly, Wooldridge's (1995) estimator relies on *conditional mean independence* assumptions while the one of Kyriazidou (1997) on a *conditional exchangeability* assumption. In this chapter we are concerned about the finite sample performance of both methods when estimating β under different settings. Although

Wooldridge (1995) is focused on the simplest consistent estimator of β in (1.1), a pooled OLS, we work with a more efficient minimum distance estimator. Furthermore, we characterise the asymptotic distribution for the minimum distance version of Wooldridge's (1995) estimator. The finite sample properties of the estimators are investigated by Monte Carlo experiments.

Chapter 3 is identical to Rochina-Barrachina (1999). In this chapter we introduce a new estimator for panel data sample selection models with "fixed-effects". The estimator relaxes some of the assumptions in the methods in chapter 2. Specifically, the estimator treats α_i as a nuisance parameter allowed to depend on the explanatory variables in an arbitrary fashion, in contrast to Wooldridge (1995), and it also avoids the *conditional exchangeability* assumption in Kyriazidou (1997). The new estimator 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 time differenced regression equation error and the two selection equation errors, 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. 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. 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.

We present two versions of the estimator depending on the treatment of the individual effects η_i. If η_i is explicitly allowed to depend on the explanatory variables in a linear way (as in Wooldridge (1995)) we have a version of the estimator referred to as the "more parametric new estimator". In this case, $f(z_{it}) - \eta_i$ in (1.2) is assumed to be equal to $z_i\gamma_t - c_t$, where $z_i \equiv (z_{i1},...,z_{iT})$, $\gamma_t \in \Re^{J \cdot T}$, and c_t is a random effect uncorrelated to the model regressors. However, if η_i is explicitly allowed to depend on the explanatory variables in a fully unrestricted way we call the estimator "less parametric new estimator". Under this alternative approach, to allow for semiparametric individual effects in the selection equation, the conditional mean of η_i is treated as an unknown function of the whole time span of the explanatory variables. The finite sample properties of both versions of the estimator are compared to those of Wooldridge (1995) and Kyriazidou's (1997) estimators by Monte Carlo experiments. In the Appendices of the chapter we provide formulae for the asymptotic variance of the new estimators.

The objective of chapter 4 is to learn about the performance of the methods in practice. Not many applications of these estimators exist in the literature. The first part of the chapter compares the three estimators in the previous chapters, points out the conditions under which each of them produces consistent estimates of β, and discusses problems of implementation. All these estimators rely on the assumption of

strict exogeneity of regressors in the equation of interest, conditional on individual specific effects and the selection mechanism. This assumption is likely to be violated in many applications. We show how non-strict exogeneity and measurement error can be taken into account within the estimation methods discussed. In the second part of the chapter, to learn about its performance we apply the estimators and their extensions to a typical problem in labour economics: The estimation of wage equations for female workers. The parameter we seek to identify is the effect of actual labour market experience on wages. Results for the participation equation for a selection of estimators are presented. The data for our empirical application is drawn from the German Socio-Economic Panel (GSOEP). The dataset used for estimation is based on the first 12 waves of the panel. The problems that arise in this application are non-random selection, and unobserved individual specific heterogeneity which might be correlated with the regressors. In addition, actual experience is predetermined, and the experience measure is likely to suffer from measurement error.

In chapter 5, estimation of the coefficients in a "double-index" selectivity bias model is considered under the assumption that the selection correction function depends only on the conditional means of some observable selection variables. We present two alternative methods. The first is referred to as a "weighted double pairwise difference estimator" (WDPDE) because of being based in the comparison of individuals in time differences. On the resulting model we apply a weighted least squares regression with decreasing weights to pairs of individuals with larger differences in their "double index" variables, and then larger differences in the selection correction terms. We extend Ahn and Powell's (1993) semiparametric

estimator of cross-section censored selection models to "fixed-effects" panel data models. We call the second method a "single pairwise difference estimator" (SPDE) because only differences over time for a given individual are required. On the model in time differences we take out its conditional expectation on the selection variables (the "double index"). This generalisation of Robinson's (1988) "partially linear" model to the case of panel data sample selection models with "fixed-effects" is estimated by least squares regression.

The estimators in this chapter have similar desirable properties as the estimator in chapter 3 (specially as the version called "less parametric new estimator"). They treat α_i as a nuisance parameter (as in Kyriazidou (1997) and our estimator in chapter 3) and η_i is explicitly allowed to depend on the explanatory variables in a fully unrestricted way (as chapter's 3 estimator under its less parametric version). However, they are distributionally free estimators compared with our earlier estimator in chapter 3 and Wooldridge's (1995) estimator. Furthermore, no *conditional exchangeability* assumption or parametric sample selection index in (1.2) is required compared with Kyriazidou (1997). In fact, by explicitly replacing in the model in chapter 5 $f(z_{it}) - \eta_i$ in (1.2) with $f_t(z_i) - c_i$ we do not only allow for semiparametric individual effects, presumably correlated with the explanatory variables, and/or for a lagged endogenous variable in the selection equation, but also for a semiparametric $f(z_{it})$ in (1.2). Although not explicitly there, the same implications hold for the "less parametric new estimator" in chapter 3. We extend the WDPDE and the SPDE to allow for endogeneity of some components of the regressors in the main equation.

The finite sample properties of the estimators are investigated by Monte Carlo experiments, and we provide in the Appendices formulae for its asymptotic variance-covariance matrices.

Chapter 2

Finite Sample Performance of Two Estimators for Panel Data Sample Selection Models with Correlated Heterogeneity*

2.1 Introduction

The utilisation of panel data is commonly confronted with two problems, sample selectivity and unobserved heterogeneity, both of which give rise to specification bias. Sample selectivity arises in nonrandomly drawn samples, as a result of either self-selection by the individuals under investigation, or selection decisions made by data-analysts. 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. In contrast to sample selectivity, unobserved heterogeneity is a problem specific to panel data. These permanent individual characteristics are commonly unobservable. Failure to account for such individual-specific effects may result in biased estimates of the behavioural parameters of interest. If the individual effect is related to some of the regressors, then we call the panel model "fixed-effects" type model; otherwise, we call

---

* Earlier versions of this chapter were presented at the XX Simposio de Análisis Económico, December 1995, Barcelona, Spain; and at the ENTER Meeting, January 1996, Toulouse, France.

it "random-effects" model. We consider the problem of estimating panel data models where both the (binary) sample selection rule and the relationship of interest contain unobservable individual-specific effects allowed to be correlated with the observable variables.

There are some estimators for panel data sample selection models that treat the individual effects in the selection equation as "random effects" uncorrelated with the observable variables (see, for instance, Verbeek (1990)). The estimator proposed by Zabel (1992) offers an alternative estimator that alleviates this problem by specifying the individual effects in the selection equation as a function of the means of time varying variables. These estimators share the reliance on distributional assumptions, the inability to incorporate serial dependence and time heteroskedasticity due to the time-varying errors, and the estimation of the models by maximum likelihood. Given the computational demands of estimating by maximum likelihood, induced by the requirement to evaluate multiple integrals, it is important to consider available two-step procedures. In particular, we are interested in two-step methods, for a "fixed-effects" type panel data sample selection model, which are semiparametric, in the sense that the model does not need to be fully specified, and relax some of the assumptions in the previous work on this area.

Fully parametric approaches to correct for selectivity bias in panel data faces the same problem that appears with cross-section data. One potential drawback to the application of these techniques is their sensitivity to the assumed parametric distribution of the unobservable error terms in the model. This chapter reviews 2 two-step "fixed effects" type estimators (with a varying degree of parametric assumptions)

for the panel data sample selection model. We focus on the recently developed methods by Wooldridge (1995) and Kyriazidou (1997), which extend the work in Nijman and Verbeek (1992), and Zabel (1992). Kyriazidou's (1997) estimator is less parametric as it does not restrict the functional form of the expectations of the individual effects conditional on the explanatory variables, but, on the other hand, Wooldridge's (1995) estimator does not impose the *conditional exchangeability* assumption characteristic in the work of Kyriazidou (1997). As the estimator of Kyriazidou (1997) imposes as few assumptions as possible on the shape of the distributions it is therefore likely to be hampered by larger standard errors. Wooldridge (1995) estimator relies on *conditional mean independence* assumptions while the one of Kyriazidou (1997) relies on a joint *conditional exchangeability* assumption for the errors in the model.

In this chapter we are concerned about the finite sample performance of Wooldridge (1995) and Kyriazidou's (1997) methods when estimating the parameters of interest under different settings. As the methods have not made assumptions about the distribution of some unobservables in the model, the finite sample distribution of the parameters is unknown. Therefore their properties are based on asymptotic behaviour. Each method relies on assumptions under which large sample properties of estimators are derived. In practice, we are not just interested in the choice of the method which asymptotically yields the most efficient and unbiased estimator but in the small sample properties of the estimators. Results from Monte Carlo simulation experiments are presented.

The chapter is organised as follows. Section 2 describes the model, the estimators and their asymptotic properties. Section 3 reports results of a small Monte Carlo simulation study of finite sample performance. Section 4 gives concluding remarks.

2.2 The Model and the Estimators

In this section, we examine Wooldridge (1995) and Kyriazidou's (1997) estimators for "fixed-effects" type panel data sample selection models.

The model can be written as follows,

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

$$d_{it}^{*} = z_{it}\gamma + \eta_i - u_{it}; \quad d_{it} = 1\left[d_{it}^{*} \geq 0\right], \tag{2.2}$$

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}.

2.2.1 Wooldridge's Estimator

The method developed by Wooldridge (1995) does not impose any distributional assumption on the individual effects and the idiosyncratic errors in the equation of interest. In this sense the estimator is semiparametric given that the model does not need to be fully specified. However, it imposes a marginal normality on the random component of the individual effects and the idiosyncratic error in the selection equation. Furthermore, it assumes a *conditional mean independence* on the equation of interest and it parameterizes some conditional means as linear projections. For instance, the individual effects in both equations are allowed to be correlated with the observable variables through these linear projections. Additionally, the conditional mean of the idiosyncratic error in the main equation on the random error term in the selection equation also follows a linear projection functional form.

Technically, Wooldridge's (1995) estimator does not require exclusion restrictions. However, in this chapter, we consider the variables in the main equation to be a subset of the variables in the sample selection rule.

In what follows, we formally state the assumptions that guarantee consistency and asymptotic normality of the estimator:

ASSUMPTION 1: $E\left(\varepsilon_{it}\middle|\alpha_i, x_i, z_i\right) = 0$, $t = 1,\ldots,T$, $x_i \equiv \left(x_{i1},\ldots,x_{iT}\right)$ and $z_i \equiv \left(z_{i1},\ldots,z_{iT}\right)$. This is an assumption of strict exogeneity of the explanatory variables with respect to ε_{it} conditional on the individual effect.

ASSUMPTION 2: For all t, (a) $E(\eta_i|z_i)$ is equal to a linear function of z_i; (b) the random error term in the selection equation $\eta_i - E(\eta_i|z_i) - u_{it} = c_i - u_{it} = -v_{it}$ follows a normal $(0,\sigma_t^2)$ and it is independent of z_i. Under conditions (a) and (b) we get the reduced form selection rule $d_{it} = 1\{\gamma_{t0} + z_{i1}\gamma_{t1} + ... + z_{iT}\gamma_{tT} - v_{it} \geq 0\}$.

ASSUMPTION 3: For the main equation, (a) $E(\varepsilon_{it}|x_i, z_i, v_{it}) = E(\varepsilon_{it}|v_{it}) = \rho_t v_{it}$. The first equality represents the mean independence of ε_{it} from the observable explanatory variables given v_{it}, or the strict exogeneity of these variables for ε_{it} given v_{it}. The second equality is just a linearity assumption for the conditional mean; (b) $E(\alpha_i|x_i, z_i, v_{it}) = x_{i1}\psi_1 + ... + x_{iT}\psi_T + \phi_t v_{it}$, which means that the regression function of α_i on x_i and v_{it} is linear. Notice that v_{it} is included in the conditioning set and in the linear projection.

Under assumptions 1 to 3, we can write (2.1) as

$$y_{it} = x_{i1}\psi_1 + ... + x_{iT}\psi_T + x_{it}\beta + \ell_t v_{it} + e_{it}, \tag{2.3}$$

where $\ell_t = \phi_t + \rho_t$ and the new error term e_{it} has conditional expectation $E(e_{it}|x_i, z_i, v_{it}) = 0$. With a first step binary choice selection equation we cannot get estimates of the residuals v_{it} and then we need the $E(y_{it}|x_i, z_i, d_{it} = 1)$, which is obtained by integrating

$$E\left(y_{it}|x_i,z_i,v_{it}\right)=x_{i1}\psi_1+\ldots+x_{iT}\psi_T+x_{it}\beta+\ell_t v_{it}, \tag{2.4}$$

over $v_{it} \leq \gamma_{t0}+z_{i1}\gamma_{t1}+\ldots+z_{iT}\gamma_{tT}$, to get

$$E\left(y_{it}|x_i,z_i,d_{it}=1\right)=x_{i1}\psi_1+\ldots+x_{iT}\psi_T+x_{it}\beta+\ell_t\lambda\left(H_{it}/\sigma_t\right), \tag{2.5}$$

where $H_{it}=z_{i1}\gamma_{t1}+\ldots+z_{iT}\gamma_{tT}=z_i\gamma_t$ is the reduced form index in the selection equation for period t and $\lambda\left(H_{it}/\sigma_t\right)=E\left[v_{it}|x_t,z_i,d_{it}=1\right]$. We assume $E\left(v_{it}^2\right)=\sigma_t^2=1$. To get estimates for $\lambda(\cdot)$, a probit is estimated for each t. For the second step, Wooldridge (1995) pointed out that two procedures are feasible. Either a pooled OLS procedure or minimum distance estimation consistently estimate β. Although Wooldridge (1995) is focused on the simplest consistent estimator, the pooled OLS, we will focus here on the more efficient minimum distance estimator. We will present the estimator that relies on OLS for each t and then it uses a minimum distance step to impose cross equation restrictions.

Rewrite (2.5) as

$$\begin{aligned}E\left(y_{it}|x_i,z_i,d_{it}=1\right)&=x_{i1}\psi_1+\ldots+x_{it-1}\psi_{t-1}+x_{it}\left(\beta+\psi_t\right)+x_{it+1}\psi_{t+1}+\ldots+x_{iT}\psi_T+\ell_t\lambda\left(z_i\gamma_t\right)\\&=x_i\Psi_t+\ell_t\lambda\left(z_i\gamma_t\right),\end{aligned} \tag{2.6}$$

where $\Psi_t\equiv\left(\psi_1',\ldots,\psi_{t-1}',\left(\beta+\psi_t\right)',\psi_{t+1}',\ldots,\psi_T'\right)'$. By following the minimum distance approach, for the subsample with $d_{it}=1$, we do least squares regression of y_{it} on x_i

and $\hat{\lambda}_{it}$ to get estimates of the reduced form parameters $\pi_t \equiv \left(\Psi_t', \ell_t\right)'$. Although estimation of the reduced form parameters requires just one wave, estimating β requires at least two waves.

Wave t provides an estimator $\hat{\pi}_t \equiv \left(\hat{\Psi}_t', \hat{\ell}_t\right)'$ for the parameter vector π_t. Define $\hat{\pi} \equiv \left(\hat{\pi}_1', \ldots, \hat{\pi}_T'\right)'$ and $\pi \equiv \left(\pi_1', \ldots, \pi_T'\right)'$. The cross equation restrictions to be exploited by the minimum distance estimator are

$$\pi = \mathrm{Re}\,st \cdot \theta, \tag{2.7}$$

where π is the stacked vector of reduced form parameters for all the waves, $\theta = \left(\psi', \beta', \ell_1, \ldots, \ell_T\right)'$, with $\psi \equiv \left(\psi_1', \ldots, \psi_T'\right)'$, is the vector of structural parameters we want to recover in the minimum distance step, and $\mathrm{Re}\,st$ is the matrix of restrictions that relates the reduced form parameters to the structural ones. Subtracting $\hat{\pi}$ from both sides of (2.7) and multiplying by -1 we get

$$\hat{\pi} - \pi = \hat{\pi} - \mathrm{Re}\,st \cdot \theta \tag{2.8}$$

The minimum distance estimator is obtained by minimising

$$\left(\hat{\pi} - \mathrm{Re}\,st \cdot \theta\right)' W^{-1} \left(\hat{\pi} - \mathrm{Re}\,st \cdot \theta\right) \tag{2.9}$$

with respect to θ, where W is a positive definite matrix. The optimal choice of W corresponds to the variance $V(\hat{\pi} - \mathrm{Re}\,st \cdot \theta)$, equal to $V(\hat{\pi} - \pi)$ according to (2.8). To get a consistent estimate $\hat{W}$ for the matrix W we need to get the influence function for $\hat{\pi}_t$. Recall (2.6) and define $R_{it} \equiv (x_i, \lambda_{it})'$ and $e_{it} = y_{it} - x_i\Psi_t - \ell_t\lambda_{it}$. The sample moment condition for $\hat{\pi}_t \equiv (\hat{\Psi}_t', \hat{\ell}_t)'$ in the second step of the estimation procedure is

$$\frac{1}{N}\sum_{i=1}^{N} d_{it}\left\{y_{it} - x_i\hat{\Psi}_t - \hat{\ell}_t\hat{\lambda}_{it}\right\}R_{it} = 0, \tag{2.10}$$

the first order condition of a two stage *extremum estimator* with finite dimensional first stage parameters. Observe that[1]

$$\sqrt{N}(\hat{\gamma}_t - \gamma_t) =^{p} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} I_{\gamma_t}^{-1}\left\{z_i'\frac{[d_{it} - \Phi(z_i\gamma_t)]\phi(z_i\gamma_t)}{\Phi(z_i\gamma_t)[1-\Phi(z_i\gamma_t)]}\right\} \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Lambda_{it}, \tag{2.11}$$

where I_{γ_t} is the probit information matrix for γ_t.

The so called *delta method* yields[2]

---

[1] The notation $=^{p}$ denotes convergence in probability.

[2] Look at the section for two-stage *extremum estimators* with finite dimensional first-stage nuisance parameters in Lee (1996).

$$\sqrt{N}\left(\hat{\pi}_t - \pi_t\right) =^{p} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} E^{-1}\left(d_t R_t R_t^{'}\right)\left\{d_{it} e_{it} R_{it} + A_t \Lambda_{it}\right\} \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \delta_{it} , \tag{2.12}$$

where

$$A_t \equiv -\ell_t E\left\{d_t\left[-\left(z\gamma_t\right)\cdot\lambda\left(z\gamma_t\right) - \lambda^2\left(z\gamma_t\right)\right]R_t z\right\}, \tag{2.13}$$

and the expression in $[\cdot]$ is the partial derivative of $\lambda\left(z\gamma_t\right)$ with respect to $\left(z\gamma_t\right)$. The term $A_t\Lambda_{it}$ is the effect of the first stage on the second. It is clear from (2.12) that the influence function for $\hat{\pi}_t$ is δ_{it}. Define $\delta_t = \left(\delta_{1t},\ldots,\delta_{it},\ldots,\delta_{Nt}\right)^{'}$ and $\delta \equiv \left(\delta_1^{'},\ldots,\delta_t^{'},\ldots,\delta_T^{'}\right)^{'}$; then $W = E\left(\delta\delta^{'}\right)/N$. The T positive definite block-on-diagonal matrices in W are equal to $E\left(\delta_\tau\delta_\tau^{'}\right)/N$, for $\tau = 1,\ldots T$, respectively. These matrices are the corresponding variance-covariance matrices of the reduced form parameters for each wave of the panel. The $T(T-1)/2$ *distinct* block-off-diagonal matrices in W are equal to $E\left(\delta_\tau\delta_s^{'}\right)/N$, for the distinct combinations of panel waves we can get with a panel of length T and being $\tau \neq s$. These matrices are the variance-covariance matrices between the reduced form parameter estimates in two different waves. Estimates for all these matrices are obtained by replacing the parameters with their estimates and the expectations involved by their sample analogous. For instance, the estimate of the inverse of the probit information matrix in (2.11) is given by

$$\hat{I}_{\gamma_t}^{-1} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\phi^2(z_i\hat{\gamma}_t)}{\Phi(z_i\hat{\gamma}_t)[1-\Phi(z_i\hat{\gamma}_t)]} z_i' z_i \right\}^{-1} . \tag{2.14}$$

With an estimate of W at hand we can provide the closed form solution to the minimisation problem in (2.9):

$$\hat{\theta}_{MDE} = \left(\mathrm{Re}\,st'\hat{W}^{-1}\,\mathrm{Re}\,st\right)^{-1}\left(\mathrm{Re}\,st'\hat{W}^{-1}\hat{\pi}\right),\ \sqrt{N}\left(\hat{\theta}_{MDE}-\theta\right) =^{d} N\left(0,\ \left(\mathrm{Re}\,st'\hat{W}^{-1}\,\mathrm{Re}\,st\right)^{-1}\right), \tag{2.15}$$

where the last term is the asymptotic distribution for the minimum distance estimator of Wooldridge's (1995) panel data sample selection model. The results for the alternative pooled OLS procedure are provided in Wooldridge's (1995) paper.

2.2.2 Kyriazidou's Estimator

The method developed by Kyriazidou (1997) does not impose any distributional assumption on the individual effects and the idiosyncratic errors in both equations in the model. The estimator is semiparametric. In contrast to Wooldridge's (1995) it is a distributionally free method that allows for individual heteroskedasticity of unknown form and it avoids the need to parameterize the functional form of any conditional mean. The price is in terms of being computationally more demanding than Wooldridge's (1995) estimator, with a convergence rate slower than $\sqrt{N}$, and a

joint-*conditional exchangeability* assumption which involves all the idiosyncratic errors in the model.

Technically, Kyriazidou's (1997) estimator requires an exclusion restriction, which implies that at least one of the variables in the selection equation, z_{it}, is not contained in the main equation regressors, x_{it}. As in Kyriazidous's (1997) analysis, we present the estimator based on a panel with two time periods. The method can be generalised to cover the case of a longer panel.

In what follows, we state the main assumptions under which the estimator is derived:

ASSUMPTION 1: $(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is})$ and $(\varepsilon_{is}, \varepsilon_{it}, u_{is}, u_{it})$ are identically distributed conditional on the vector of (observed and unobserved) explanatory variables $(x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i)$.

The joint *conditional exchangeability* assumption implies stationary marginal distributions for the time varying errors in the model.

ASSUMPTION 2: Each period sample selection effect is a sufficiently smooth function of the indices $z_{it}\gamma$, $z_{is}\gamma$ and the joint conditional distribution of the errors. This smoothness condition ensures that once Assumption 1 holds, $z_{it}\gamma = z_{is}\gamma$ implies that the selection terms are the same in the two time periods and they cancel each other by time differencing the main equation in the model. Differencing between

periods s and t will entirely remove, at the same time, the time constant individual effect.

Under assumptions 1 and 2, an OLS estimator applied to

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + e_{its}, \tag{2.16}$$

for individuals satisfying $d_{it} = d_{is} = 1, s \neq t$ and $z_{it}\gamma = z_{is}\gamma$, is consistent. The resulting error $e_{its} \equiv (\varepsilon_{it} - \varepsilon_{is}) - (\lambda_{its} - \lambda_{ist})$, where λ_{its} and λ_{ist} are the selection terms for periods t and s, respectively, has a conditional expectation that satisfies $E(e_{its} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_t, d_{it} = d_{is} = 1) = 0$. For each time period the selection terms are

$$\begin{aligned} \lambda_{its} &= \mathrm{E}(\varepsilon_{it} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, u_{it} \leq z_{it}\gamma + \eta_i, u_{is} \leq z_{is}\gamma + \eta_i) \\ \lambda_{ist} &= \mathrm{E}(\varepsilon_{is} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, u_{is} \leq z_{is}\gamma + \eta_i, u_{it} \leq z_{it}\gamma + \eta_i) \end{aligned} \tag{2.17}$$

The estimation procedure has several steps. The estimator requires that there are individuals with $z_{it}\gamma = z_{is}\gamma$ with probability one, which is rare in a given sample. To implement the estimator, Kyriazidou (1997) constructs kernel weights, which are a declining function of the distance $|z_{it}\gamma - z_{is}\gamma|$, and estimates time differenced

equations by weighted OLS[3]. For a fixed sample size, observations with less selectivity bias are given more weight, while asymptotically, only those observations with zero bias are used. Thus, in the first step, the unknown coefficients of the selection equation are estimated by the smoothed conditional maximum score estimator (SCMSE) considered in an earlier version of Kyriazidou's (1997) paper (Kyriazidou (1994)) and also in Charlier et al. (1995). This estimator is a mixture of the panel version of the maximum score estimator of Manski (1975, 1985), proposed by Manski (1987), and of the smoothed maximum score estimator of Horowitz (1992) for cross-section data.

The SCMSE is obtained by maximising the following expression conditional on $d_{it} \neq d_{is}$

$$I(\gamma;\sigma_N) = \frac{1}{N}\sum_{i=1}^{N}\left[2 \cdot 1\{d_{it} - d_{is} = 1\} - 1\right] \cdot L\left(\frac{(z_{it} - z_{is})\gamma}{\sigma_N}\right), \tag{2.18}$$

where σ_N is a sequence of strictly positive real numbers satisfying $lim_{N\to\infty}\sigma_N = 0$, $1\{\cdot\}$ is an indicator function and L is a continuous function, analogous to a cumulative distribution function but it also might take on values larger than one or lower than zero and it need not be increasing. Two examples of functions $L(\cdot)$ satisfying the requirements for this smoothing function (see, for instance, Horowitz (1992) or

[3] The estimator is arbitrarily close to root n-consistency depending on the degree of smoothness one is willing to assume for the kernel function.

Charlier et al. (1995)) are $L_2(\cdot) = \Phi(\cdot)$, where Φ is the cumulative standard normal distribution function, and

$$L_4(v) = \begin{cases} 0 & \text{if } v < -1 \\ 0.5 + \left(\frac{105}{64}\right)\left[(v) - \left(\frac{5}{3}\right)(v)^3 + \left(\frac{7}{5}\right)(v)^5 - \left(\frac{3}{7}\right)(v)^7\right] & \text{if } -1 \le v \le 1. \\ 1 & \text{if } v > 1 \end{cases} \qquad (2.19)$$

L_4 is the integral of a fourth order kernel for nonparametric density estimation (Müller, 1984). In the Monte Carlo experiments we will restrict our attention to $L_4(\cdot)$.

The parameters γ are identified up to scale, under the normalisation $|\gamma_k| = 1$, with γ_k being a nonzero coefficient of an absolute continuous element of the vector $(z_{it} - z_{is})$.

The asymptotic distribution of the SCMSE is given by

$$N^{\frac{R_1+1}{[2(R_1+1)+1]}} \cdot (\hat{\gamma} - \gamma) =^d N\left[-(\vartheta^*)^{\frac{R_1+1}{[2(R_1+1)+1]}} \cdot C^{-1}A,\ (\vartheta^*)^{-\frac{1}{[2(R_1+1)+1]}} \cdot C^{-1}DC^{-1}\right], \qquad (2.20)$$

where $(R_1 + 1)$ is the order of the kernel associated to the function $L(\cdot)$ [4]. We can see from (2.20) that the fastest possible rate of convergence in distribution for $\hat{\gamma}$ is

[4] 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).

$N^{-\frac{R_1+1}{[2(R_1+1)+1]}}$, slower than $N^{-\frac{1}{2}}$. A sufficient condition to obtain this optimal rate of convergence is $\sigma_N = \left(\vartheta/N\right)^{\frac{1}{[2(R_1+1)+1]}}$ with $0 < \vartheta < \infty$. Based on the asymptotic result of (2.20) the asymptotic optimal value for ϑ, in the sense of minimising the Mean Square Error (MSE)= $E\left[(\hat{\gamma}-\gamma)'\Omega(\hat{\gamma}-\gamma)\right]$, is $\vartheta = \vartheta^* = \dfrac{trace\left[C^{-1}\Omega C^{-1}D\right]}{2(R_1+1)A'C^{-1}\Omega C^{-1}A}$, where Ω is any nonstochastic, positive semidefinite matrix such that $A'C^{-1}\Omega C^{-1}A \neq 0$. By choosing R_1 large enough the rate of convergence can be made arbitrarily close to $N^{-\frac{1}{2}}$. From (2.20) the bias corrected estimator is $\tilde{\gamma} = \hat{\gamma} + \left(\vartheta^*/N\right)^{\frac{R_1+1}{[2(R_1+1)+1]}} \cdot C^{-1}A$.

Finally, to make the results useful in applications, it is necessary to estimate consistently the matrices A, D and C. The structure of the asymptotic covariance matrix is similar to that of an *extremum estimator*. Let $\hat{\gamma}$ be a consistent smoothed conditional maximum score estimator based on $\sigma_N = \left(\vartheta/N\right)^{\frac{1}{[2(R_1+1)+1]}}$. For $|\gamma_k| = 1$, define

$$r_{its}(\gamma;\sigma) = \left[2 \cdot 1\{d_{it} - d_{is} = 1\} - 1\right] \cdot L'\left(\frac{(z_{it}-z_{is})\gamma}{\sigma}\right) \cdot \frac{(z_{it}-z_{is})^{(-k)}}{\sigma}, \qquad (2.21)$$

where $(z_{it}-z_{is})^{(-k)}$ excludes the k-element from the vector $(z_{it}-z_{is})$. Let $\sigma_{N,\delta_1} = \left(\vartheta/N\right)^{\frac{\delta_1}{[2(R_1+1)+1]}}$, where $0 < \delta_1 < 1$. Then

$$\hat{A} \equiv \left(\sigma_{N,\delta_1}\right)^{-(R_1+1)} \frac{\partial I\left(\hat{\gamma};\sigma_{N,\delta_1}\right)}{\partial \gamma} =^{p} A, \tag{2.22}$$

where $\frac{\partial I\left(\hat{\gamma};\sigma_{N,\delta_1}\right)}{\partial \gamma}$ is the first order derivation of the objective function in (2.18) with respect to $\gamma^{(-k)}$ evaluated at $\left(\hat{\gamma},\sigma_{N,\delta_1}\right)$ [5];

$$\hat{D} \equiv \left(\sigma_N / N\right) \sum_{i=1}^{N} r_{its}\left(\hat{\gamma};\sigma_N\right) r_{its}\left(\hat{\gamma};\sigma_N\right)' =^{p} D; \tag{2.23}$$

$$\hat{C} \equiv \frac{\partial^2 I\left(\hat{\gamma};\sigma_N\right)}{\partial \gamma \gamma'} =^{p} C, \tag{2.24}$$

where $\frac{\partial^2 I\left(\hat{\gamma};\sigma_N\right)}{\partial \gamma \gamma'}$ is the second order derivation of the objective function in (2.18) with respect to $\gamma^{(-k)}$ evaluated at $\left(\hat{\gamma},\sigma_N\right)$.

For a complete revision of the assumptions and regularity conditions that guarantee consistency and asymptotic normality of the SCMSE see Manski (1987), Horowitz (1992), Kyriazidou (1994) and Charlier et al. (1995).

For the second step, the weighted OLS estimator is given by

---

[5] Note that $\frac{\partial I\left(\hat{\gamma};\sigma_{N,\delta}\right)}{\partial \gamma} \neq 0$, even though $\frac{\partial I\left(\hat{\gamma};\sigma_N\right)}{\partial \gamma} = 0$ by the first order condition of the optimisation problem in (2.18), because $\sigma_{N,\delta}$ converges to 0 more slowly than σ_N.

$$
\begin{aligned}
&\hat{\beta} = \hat{S}_{xx}^{-1}\hat{S}_{xy}, \\
&\hat{S}_{xx} \equiv \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{its}(x_{it}-x_{is})'(x_{it}-x_{is})d_{it}d_{is}, \\
&\hat{S}_{xy} \equiv \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{its}(x_{it}-x_{is})'(y_{it}-y_{is})d_{it}d_{is}, \\
&\hat{\psi}_{its} \equiv \frac{1}{c_N}K\left(\frac{(z_{it}-z_{is})\hat{\gamma}}{c_N}\right),
\end{aligned}
\tag{2.25}
$$

where $K(\cdot)$ is a "kernel density" function and c_N is a sequence of bandwidths which tends to zero as $n \to \infty$.

In order to derive the asymptotic properties of the estimator $\hat{\beta}$, Kyriazidou (1997) makes, among others, the following additional assumption:

ASSUMPTION 3: $c_N = c \cdot N^{-\mu}$, where $0 < c < \infty$, and $1-2p < \mu < p/2$, where p is the rate of convergence of the first step estimator $\hat{\gamma}$.

Under the whole set of assumptions and if $\sqrt{Nc_N}\,c_N^{R_2+1} \to \tilde{c}$ with $0 \le \tilde{c} < \infty$ the asymptotic distribution of the estimator is[6]

$$
\sqrt{Nc_N}\left(\hat{\beta}-\beta\right) =^{d} N\left(\tilde{c}\,\Sigma_{xx}^{-1}\Sigma_{x\lambda},\ \Sigma_{xx}^{-1}\Sigma_{xv}\Sigma_{xx}^{-1}\right). \tag{2.26}
$$

[6] For a complete revision including all the regularity conditions see Kyriazidou's (1997) paper.

It is shown by Kyriazidou (1997) that the asymptotic distribution of $\hat{\beta}$ coincides with the asymptotic distribution of the unfeasible estimator which uses the true γ in the kernel weights. Then, asymptotically, the first step estimator does not affect the asymptotic distribution of the second step estimator. The key for this result to hold is that from $(Nc_N)^{-1/2} = c^{-1/2} N^{-1/2+\mu/2}$ the rate of convergence of $\hat{\beta}$ is $N^{-1/2+\mu/2}$, slower than N^{-p}, since $1-2p < \mu$ by Assumption 3. The maximal rate of convergence in distribution of $\hat{\beta}$ is achieved by setting μ as small as possible, that is $\mu = 1/[2(R_2+1)+1]$. A bias-corrected estimator, with respect to the estimator in (2.26), is obtained by following Bierens (1987):

$$\hat{\hat{\beta}} \equiv \frac{\hat{\beta} - N^{-\frac{(1-\delta_2)(R_2+1)}{[2(R_2+1)+1]}} \cdot \hat{\beta}_{\delta_2}}{1 - N^{-\frac{(1-\delta_2)(R_2+1)}{[2(R_2+1)+1]}}} \tag{2.27}$$

where $0 < \delta_2 < 1$ and the estimators $\hat{\beta}$ and $\hat{\beta}_{\delta_2}$ have the associated bandwidths $c_N = c \cdot N^{-1/[2(R_2+1)+1]}$ and $c_{N,\delta_2} = c \cdot N^{-\delta_2/[2(R_2+1)+1]}$, respectively. The asymptotic distribution of $\hat{\hat{\beta}}$ is

$$N^{\frac{R_2+1}{[2(R_2+1)+1]}}\left(\hat{\hat{\beta}} - \beta\right) =^{d} N\left(0,\ c^{-1} \Sigma_{xx}^{-1} \Sigma_{xv} \Sigma_{xx}^{-1}\right). \tag{2.28}$$

The bias-corrected estimator preserves the maximal rate of convergence which can be arbitrarily close to $N^{-1/2}$ depending on R_2 and provided that γ is estimated faster than β, that is $p > (R_2 + 1)/[2(R_2 + 1) + 1]$.

In Kyriazidou (1997) it is proposed to choose c so as to minimise the asymptotic MSE of the estimator based on the asymptotic result of (2.26):

$$\begin{aligned} MSE &= E\left[\left(\hat{\beta}-\beta\right)'\Gamma\left(\hat{\beta}-\beta\right)\right] = trace\left[\Gamma E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right]\right] \\ &= N^{-2(R_2+1)/[2(R_2+1)+1]}\, trace\left[\Gamma \Sigma_{xx}^{-1}\left(c^{-1}\Sigma_{xv} + c^{2(R_2+1)}\Sigma_{x\lambda}\Sigma_{x\lambda}'\right)\Sigma_{xx}^{-1}\right], \end{aligned} \tag{2.29}$$

for any nonstochastic positive semidefinite matrix Γ that satisfies $\Sigma'_{x\lambda}\Sigma_{xx}^{-1}\Gamma\Sigma_{xx}^{-1}\Sigma_{x\lambda} \neq 0$. Thus, the MSE is minimised by setting

$$c = c^* = \left(\frac{trace\left[\Sigma_{xx}^{-1}\Gamma\Sigma_{xx}^{-1}\Sigma_{xv}\right]}{2(R_2+1)\Sigma'_{x\lambda}\Sigma_{xx}^{-1}\Gamma\Sigma_{xx}^{-1}\Sigma_{x\lambda}}\right)^{1/[2(R_2+1)+1]} \tag{2.30}$$

Finally, to make the results useful in applications, it is necessary to estimate consistently the matrices Σ_{xx}, Σ_{xv} and $\Sigma_{x\lambda}$. Define $\left(\hat{\varepsilon}_{it} - \hat{\varepsilon}_{is}\right) \equiv \left(y_{it} - y_{is}\right) - \left(x_{it} - x_{is}\right)\hat{\beta}$. Then

$$\hat{S}_{xx} =^{p} \Sigma_{xx}; \tag{2.31}$$

$$\hat{\Sigma}_{xv} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{1}{c_N}K\left(\frac{(z_{it}-z_{is})\hat{\gamma}}{c_N}\right)^2 (x_{it}-x_{is})'(x_{it}-x_{is})(\hat{\varepsilon}_{it}-\hat{\varepsilon}_{is})^2 d_{it}d_{is} =^p \Sigma_{xv}; \qquad (2.32)$$

$$\hat{\Sigma}_{x\lambda} \equiv c_{N,\delta_2}^{-(R+1)} \cdot \frac{1}{N}\sum_{i=1}^{N}\frac{1}{c_{N,\delta_2}}K\left(\frac{(z_{it}-z_{is})\hat{\gamma}}{c_{N,\delta_2}}\right)(x_{it}-x_{is})'(\hat{\varepsilon}_{it}-\hat{\varepsilon}_{is})d_{it}d_{is} =^p \Sigma_{x\lambda}. \qquad (2.33)$$

An extension of the method to cover the case of a longer panel is briefly mentioned in Kyriazidou (1997). The estimator is of the form

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i-1}\sum_{s<t}\hat{\psi}_{its}(x_{it}-x_{is})'(x_{it}-x_{is})d_{it}d_{is}\right]^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i-1}\sum_{s<t}\hat{\psi}_{its}(x_{it}-x_{is})'(y_{it}-y_{is})d_{it}d_{is} \qquad (2.34)$$

for all $s,t = 1,\ldots,T_i$, where T_i denotes the number of waves for the individual i. However, an easier way to generalise the estimator to the case of more than two time periods is as follows: given some estimates for the selection equation[7], the main equation can be estimated using (2.25) for each two waves in the panel, and then a minimum distance estimator can be used to combine all the estimates. The asymptotic distribution of the minimum distance estimator for the Kyriazidou's (1997) panel data model with more than two time periods is derived by Charlier et al. (1997). A consistent estimate for the weighting matrix for the minimum distance is required. The $T(T-1)/2$ positive definite block-on-diagonal matrices of this matrix are the

[7] Obtained, for instance, as in Charlier et al. (1995), where it can be found an extension to a longer panel of the SCMSE in Kyriazidou (1994).

corresponding variance-covariance matrices for each different pair of waves in the panel. The $\left(T(T-1)/4\right)\left[\left(T(T-1)/2\right)-1\right]$ *distinct* block-off-diagonal matrices are the variance-covariance matrices between the Kyriazidou's (1997) estimators based on two different pairs of panel waves. These block-off-diagonal matrices converge to zero due to the fact that the bandwidth tends to zero as $N \to \infty$. The proof can be found in Charlier et al. (1997). The minimum distance estimator is therefore a weighted average of the estimators for each pair (t,s), $t \neq s$, with weights given by the inverse of the corresponding variance-covariance matrix estimate.

2.3 Monte Carlo Experiments

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

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i + \varepsilon_{it}; \quad i = 1,\ldots,N; \quad t = 1,2, \\ d_{it}^{*} &= z_{1it}\gamma_1 + z_{2it}\gamma_2 + \eta_i - u_{it}; \quad d_{it} = 1\left[d_{it}^{*} \geq 0\right], \end{aligned} \tag{3.1}$$

where y_{it} is only observed if $d_{it} = 1$. The true value of β, γ_1, and γ_2 is 1. For the baseline Monte Carlo design 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 $\eta_i = (z_{1i1} + z_{1i2})/2 + (z_{2i1} + z_{2i2})/2 + c_i - 1$ and $\alpha_i = (x_{i1} + x_{i2})/2 + \sqrt{2} \cdot N(0,1) + 1$, where $c_i = 0.6 \cdot N(0,1)$ is a random effect[8]; the idiosyncratic errors are as follows: $u_{it} \sim 0.8 \cdot N(0,1)$, $v_{it} = c_i - u_{it}$, and $\varepsilon_{it} = 0.8 \cdot N(0,1) + 0.6 \cdot v_{it}$. 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 generated data in the basic Monte Carlo design are compatible with the assumptions in both methods.

The results with 100 replications and sample sizes equal to 250, 500, 1000, 2000, 4000, 8000 and 14000 are presented in Tables 1 to 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 the median absolute deviation (MAD) are also reported.

Table 1 presents the finite sample properties of the two estimators under our basic Monte Carlo design. The estimates will be consistent since all of the assumptions in the methods hold. Going from top to bottom of Table 1, Table 1A reports the results for Kyriazidou's (1997) bias corrected estimator (with $\delta_2 = 0.9$) using the true γ in the construction of the kernel weights. We use a second ($R_2 = 1$),

[8] The individual effects design is driven by the fact that we want to keep both its linear correlation with respect to the explanatory variables, and a normality assumption for its random component. The reason is that Wooldridge's (1995) estimator assumes normality for the 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, Wooldridge's (1995) estimator is developed under the assumption of a linear correlation between the individual effects in the selection equation and the leads and lags of the explanatory variables. Furthermore, Wooldridge (1995) also imposes the linearity assumption 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.

TABLE 1: Basic Monte Carlo Design

Table 1A

Kyriazidou's Estimator: Real First Step Parameters															
	R_2=1					R_2=3					R_2=5				
N	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.0379	-0.0386	0.1270	0.1024	0.0842	-0.0511	-0.0475	0.1218	0.0997	0.0869	-0.0550	-0.0502	0.1211	0.1001	0.0876
500	-0.0293	-0.0295	0.0997	0.0770	0.0571	-0.0457	-0.0482	0.0883	0.0729	0.0568	-0.0501	-0.0505	0.0870	0.0725	0.0585
1000	-0.0368	-0.0391	0.0782	0.0590	0.0560	-0.0571	-0.0555	0.0798	0.0537	0.0572	-0.0629	-0.0613	0.0820	0.0527	0.0627
2000	-0.0228	-0.0247	0.0606	0.0437	0.0460	-0.0449	-0.0425	0.0634	0.0385	0.0453	-0.0514	-0.0538	0.0668	0.0374	0.0548
4000	-0.0195	-0.0212	0.0478	0.0328	0.0319	-0.0416	-0.0397	0.0531	0.0285	0.0408	-0.0499	-0.0492	0.0582	0.0273	0.0492
8000	-0.0079	-0.0083	0.0306	0.0251	0.0205	-0.0337	-0.0329	0.0403	0.0208	0.0329	-0.0452	-0.0454	0.0494	0.0195	0.0454
14000	-0.0133	-0.0139	0.0302	0.0198	0.0236	-0.0365	-0.0372	0.0412	0.0161	0.0372	-0.0476	-0.0470	0.0504	0.0149	0.0470

Table 1B

Kyriazidou's Estimator: Estimated First Step Parameters R_2=1, R_1=3						Wooldridge's Estimator				
N	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.0355	-0.0249	0.1431	0.1067	0.1032	-0.0104	-0.0165	0.1217	0.1275	0.0888
500	-0.0433	-0.0413	0.0945	0.0777	0.0653	0.0123	0.0022	0.0935	0.0897	0.0594
1000	-0.0168	-0.0138	0.0696	0.0583	0.0420	-0.0085	-0.0120	0.0619	0.0637	0.0405
2000	-0.0114	-0.0097	0.0638	0.0446	0.0494	0.0041	0.0007	0.0489	0.0451	0.0333
4000	-0.0127	-0.0098	0.0360	0.0331	0.0283	-0.0048	-0.0061	0.0289	0.0320	0.0210
8000	-0.0086	-0.0146	0.0390	0.0259	0.0297	-0.0007	-0.0004	0.0208	0.0225	0.0156
14000	-0.0034	-0.0083	0.0305	0.0205	0.0213	-0.0006	0.0004	0.0164	0.0170	0.0106

a fourth ($R_2 = 3$) and a sixth ($R_2 = 5$) order bias-reducing kernel function[9]. The bandwidth sequence is $c_N = cN^{-1/[2(R_2+1)+1]} = cN^{-1/5}$, $cN^{-1/9}$, $cN^{-1/13}$, respectively. We chose the initial c equal to 1. Then, we compute $\hat{\beta}$ based on c_N, and construct $(\hat{\varepsilon}_{it} - \hat{\varepsilon}_{is})$ as defined in section 2 above. We use $\hat{\beta}$ and $(\hat{\varepsilon}_{it} - \hat{\varepsilon}_{is})$ to compute the estimates of Σ_{xx}, Σ_{xv}, and $\Sigma_{x\lambda}$ as in (2.31), (2.32), and (2.33), respectively. Then we estimate c^* by $\hat{c}$, using equation (2.30) with Σ_{xx}, Σ_{xv}, and $\Sigma_{x\lambda}$ replaced by their consistent estimates. The asymptotic bias-corrected estimator is computed as in (2.27) using $\hat{c}$ as the constant in the definition of c_N and c_{N,δ_2}. On the left-hand side of Table 1B we focus on the case of a second order bias-reducing kernel function and γ

---

[9] We use high ($R_2 + 1$) order bias reducing kernels constructed following Bierens (1987). For $z, \gamma \in \Re$ and $[(R_2+1)/2] \geq 1$ let

$$K_{R_2+1}[(z_{it} - z_{is})\gamma / c_N] = \sum_{p=1}^{(R_2+1)/2} \frac{\theta_p \exp\left\{-(1/2)[(z_{it} - z_{is})\gamma / c_N]' \bar{\Omega}^{-1}[(z_{it} - z_{is})\gamma / c_N] / \mu_p^2\right\}}{(\sqrt{2\pi})|\mu_p|\sqrt{\det(\bar{\Omega})}}$$

where $\bar{\Omega}$ is a positive definite matrix and the parameters θ_p and μ_p are such that $\sum_{p=1}^{(R_2+1)/2} \theta_p = 1$ and $\sum_{p=1}^{(R_2+1)/2} \theta_p \cdot \mu_p^{2\upsilon} = 0$, for $\upsilon = 1,2,\ldots,[(R_2+1)/2] - 1$. We should specify $\bar{\Omega} = \hat{V}$, where $\hat{V}$ is the sample variance matrix; that is, $\hat{V} = (1/N)\sum_{j=1}^{N}[(z_{jt} - z_{js})\hat{\gamma} - \bar{Z}]'[(z_{jt} - z_{js})\hat{\gamma} - \bar{Z}]$ with $\bar{Z} = (1/N)\sum_{j=1}^{N}(z_{jt} - z_{js})\hat{\gamma}$. Thus, for $R_2 + 1 = 2,4,6,\ldots$, we get

$$\hat{K}_{R_2+1}[(z_{it} - z_{is})\hat{\gamma} / c_N] = \sum_{p=1}^{(R_2+1)/2} \frac{\theta_p \exp\left\{-(1/2)[(z_{it} - z_{is})\hat{\gamma} / c_N]' \hat{V}^{-1}[(z_{it} - z_{is})\hat{\gamma} / c_N] / \mu_p^2\right\}}{(\sqrt{2\pi})|\mu_p|\sqrt{\det(\hat{V})}}.$$

For $R_2 = 1$ we set $\theta_1 = \mu_1 = 1$; for $R_2 = 3$ we set $\theta_1 = 2$, $\theta_2 = -1$, $\mu_1 = 1$, and $\mu_2 = \sqrt{2}$; for $R_2 = 5$ we set $\theta_1 = 3$, $\theta_2 = -3$, $\theta_3 = 1$, $\mu_1 = 1$, $\mu_2 = \sqrt{2}$, and $\mu_3 = \sqrt{3}$.

is estimated by the SCMSE. The SCMSE is computed by maximising (2.18) with respect to γ_2 given $\gamma_1 = 1$ by scale normalisation in all of the Monte Carlo replications[10]. For L we use L_4 in (2.19), so $R_1 + 1 = 4$ and the optimal rate of convergence for $\hat{\gamma}$ is $N^{-4/9}$, faster than the rate of convergence for $\hat{\beta}$ ($N^{-2/5}$). The results are those for $\delta_1 = 0.1$. The bandwidth parameter for the SCMSE was constructed as follows. Given $R_1 + 1 = 4$, we chose $\sigma_N = N^{-1/[2(R_1+1)+1]}$ and $\sigma_{N,\delta_1} = N^{-\delta_1/[2(R_1+1)+1]}$. We then compute the SCMSE $\hat{\gamma}$ based on σ_N, and use $\hat{\gamma}$ and σ_{N,δ_1} to compute $\hat{A}$, $\hat{D}$, and $\hat{C}$ by (2.22), (2.23), and (2.24), respectively. Then we estimate ϑ^* by $\hat{\vartheta}$, where $\hat{\vartheta}$ is obtained from ϑ^* in section 2 by replacing A, D, and C with $\hat{A}$, $\hat{D}$, and $\hat{C}$. For Ω the identity matrix was used. Table 1B reports on the right-hand side the results for the minimum distance version of Wooldridge's (1995) estimator.

From Table 1 we see that Wooldridge's (1995) estimator is less biased than Kyriazidou's (1997) estimator both with and without estimated first step parameters. Furthermore, Wooldridge's (1995) estimator reaches its asymptotic behaviour faster due to the fact that this estimator is $\sqrt{N}$-consistent. Kyriazidou's (1997) estimator is consistent at a rate slower than $N^{-1/2}$ and for this reason behaves well for bigger

[10] I thank Charles Manski and Scott Thompson for providing me with their computer program for maximum score estimation. Ekaterini Kyriazidou kindly provided a computer program of Joel Horowitz for smoothed maximum score estimation. In this study we use an extension to panel data of the latter program. Given that the smoothed conditional score function can have multiple extrema, the program uses a nonconvex optimisation technique. In particular, the simulated annealing algorithm of Szu and Hartley (1987). The SCMSE is bias corrected.

sample sizes. For this estimator the more satisfactory results are obtained with the normal kernel density function ($R_2 = 1$).

In Table 2 we generate a misspecification problem for Wooldridge's (1995) estimator. In Table 2A the linear projection functional form for the individual effects in the main equation has been violated. We have generated the true $\alpha_i's$ by adding to our benchmark specification quadratic terms on the $x's$:

$$\alpha_i = (x_{i1} + x_{i2})/2 + (x_{i1}^2 + x_{i2}^2)/2 + \sqrt{2} \cdot N(0,1) + 1. \qquad (3.2)$$

Under this design Wooldridge's (1995) estimator is clearly inconsistent and it suffers from a misspecification bias problem. The estimator behaves badly in terms of all the considered measures. In Table 2B we invalidate the linearity assumption for the individual effects in the selection equation by adding quadratic terms on the $z's$:

$$\eta_i = (z_{1i1} + z_{1i2})/2 + (z_{2i1} + z_{2i2})/2 + (z_{1i1}^2 + z_{1i2}^2)/2 + (z_{2i1}^2 + z_{2i2}^2)/2 + c_i - 1. \qquad (3.3)$$

The inconsistency of the first step parameter estimates hardly influences the bias for the second step estimates. Experiments for Kyriazidou's (1997) estimator are not included in Table 2 because this estimator is robust against any type of design for the individual effects in both equations. As the method is based in estimation with time differences its properties are independent of the particular shape of the individual effects.

TABLE 2: Generating A Misspecification Problem For Wooldridge

Table 2A

$$\alpha_i = (x_{i1} + x_{i2})/2 + (x_{i1}^2 + x_{i2}^2)/2 + \sqrt{2} \cdot N(0,1) + 1$$

N	Mean Bias	Median Bias	SE	ASE	MAD
250	0.4336	0.4293	0.4662	0.1691	0.4293
500	0.4645	0.4637	0.4819	0.1212	0.4637
1000	0.4485	0.4426	0.4576	0.0865	0.4426
2000	0.4518	0.4550	0.4567	0.0610	0.4550
4000	0.4539	0.4528	0.4559	0.0434	0.4528
8000	0.4576	0.4610	0.4585	0.0307	0.4610
14000	0.4549	0.4572	0.4556	0.0231	0.4572

Table 2B

$$\eta_i = (z_{1i1} + z_{1i2})/2 + (z_{2i1} + z_{2i2})/2 + (z_{1i1}^2 + z_{1i2}^2)/2 + (z_{2i1}^2 + z_{2i2}^2)/2 + c_i - 1$$

N	Mean Bias	Median Bias	SE	ASE	MAD
250	0.0290	-0.0155	0.3151	0.3198	0.1647
500	0.0235	0.0431	0.2478	0.2257	0.1960
1000	0.0009	0.0101	0.1568	0.1584	0.1154
2000	-0.0060	-0.0122	0.1048	0.1139	0.0813
4000	-0.0015	-0.0027	0.0840	0.0799	0.0608
8000	0.0050	0.0093	0.0519	0.0564	0.0354
14000	0.0097	0.0081	0.0463	0.0426	0.0356

In Table 3 we generate a predetermined variable which affects the set of explanatory variables in both equations in (3.1). We invalidate the strict exogeneity assumption underlying both methods. The predetermined variable $x_{i2} = z_{2i2}$ for the new experiment has been generated as

$$z_{2i2} = N(0,1) + 0.5\varepsilon_{i1}. \tag{3.4}$$

We report the results for both estimators. Note that the results are very similar when Wooldridge's (1995) estimator is used relative to the case where Kyriazidou's (1997) estimator is applied. Both estimators show a strong negative bias, and huge SE and MAD very far from the ASE predicted by the asymptotic theory.

In Tables 4 and 5 we compare Wooldridge (1995) and Kyriazidou's (1997) estimators when the *conditional exchangeability* assumption breaks down. Differently to the baseline design, for Table 4 we have:

$$\begin{aligned} &u_{i1} \sim 0.5 \cdot N(0,1);\ u_{i2} \sim 2 \cdot N(0,1); \\ &c_i = 0.6 \cdot N(0,1); \\ &v_{i1} = c_i - u_{i1};\ v_{i2} = c_i - u_{i2}; \\ &\varepsilon_{i1} = 0.8 \cdot N(0,1) + 0.1 \cdot v_{i1};\ \varepsilon_{i2} = 0.8 \cdot N(0,1) + 0.9 \cdot v_{i2} \end{aligned} \tag{3.5}$$

For Table 5 we substitute ε_{i1} in (3.5) by $\varepsilon_{i1} = 0.8 \cdot N(0,1) + 0.1 \cdot v_{i1} - 5$.

We allow for non-constant variances over time for the error terms and different degrees for the sample selection problem. The latter comes through different

TABLE 3: Invalidating Strict Exogeneity

$$z_{2i2} = N(0,1) + 0.5\varepsilon_{i1}$$

Table 3A: Wooldridge's Estimator

N	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.2287	-0.2323	0.2556	0.1138	0.2323
500	-0.2103	-0.2164	0.2289	0.0789	0.2164
1000	-0.2124	-0.2115	0.2197	0.0571	0.2115
2000	-0.2086	-0.2072	0.2129	0.0405	0.2072
4000	-0.2104	-0.2106	0.2130	0.0288	0.2106
8000	-0.2061	-0.2035	0.2071	0.0202	0.2035
14000	-0.2069	-0.2067	0.2076	0.0153	0.2067

Table 3B: Kyriazidou's Estimator

N	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.2019	-0.2244	0.2450	0.0990	0.2285
500	-0.2168	-0.2191	0.2322	0.0723	0.2191
1000	-0.2050	-0.2197	0.2150	0.0542	0.2197
2000	-0.2026	-0.2012	0.2089	0.0411	0.2012
4000	-0.2087	-0.2095	0.2127	0.0309	0.2095
8000	-0.1992	-0.2012	0.2011	0.0232	0.2012
14000	-0.1900	-0.1928	0.1913	0.0192	0.1928

TABLE 4: Invalidating The Exchangeability Assumption In Kyriazidou

$$u_{i1} \sim 0.5 \cdot N(0,1),\ u_{i2} \sim 2 \cdot N(0,1)$$
$$c_i = 0.6 \cdot N(0,1)$$
$$v_{it} = c_i - u_{it}$$
$$\varepsilon_{i1} = 0.8 \cdot N(0,1) + 0.1 \cdot v_{i1},\ \varepsilon_{i2} = 0.8 \cdot N(0,1) + 0.9 \cdot v_{i2}$$

Table 4A

Kyriazidou's Estimator: Real First Step Parameters															
	R_2=1					R_2=3					R_2=5				
N	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.0540	-0.0819	0.2057	0.1587	0.1646	-0.0909	-0.0950	0.1984	0.1526	0.1489	-0.1010	-0.1004	0.1963	0.1526	0.1468
500	-0.0584	-0.0705	0.1621	0.1199	0.1158	-0.0979	-0.1047	0.1533	0.1118	0.1196	-0.1091	-0.1148	0.1569	0.1107	0.1231
1000	-0.0672	-0.0847	0.1450	0.0926	0.1078	-0.1133	-0.1188	0.1485	0.0839	0.1188	-0.1285	-0.1339	0.1549	0.0818	0.1339
2000	-0.0399	-0.0508	0.0983	0.0696	0.0781	-0.0917	-0.0949	0.1143	0.0598	0.0949	-0.1094	-0.1087	0.1261	0.0576	0.1087
4000	-0.0447	-0.0416	0.0875	0.0510	0.0574	-0.0888	-0.0906	0.1031	0.0436	0.0906	-0.1091	-0.1122	0.1180	0.0414	0.1122
8000	-0.0196	-0.0115	0.0530	0.0403	0.0378	-0.0727	-0.0713	0.0811	0.0322	0.0713	-0.0963	-0.0965	0.1009	0.0298	0.0965
14000	-0.0157	-0.0188	0.0461	0.0322	0.0325	-0.0661	-0.0640	0.0730	0.0253	0.0640	-0.0933	-0.0913	0.0971	0.0229	0.0913

Table 4B

Kyriazidou's Estimator: Estimated First Step Parameters R_2=1, R_1=3						Wooldridge's Estimator				
N	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.0991	-0.1099	0.2636	0.1656	0.1553	-0.0113	-0.0160	0.1444	0.1575	0.0797
500	-0.0770	-0.0761	0.2179	0.1275	0.1180	0.0151	0.0026	0.1201	0.1108	0.0852
1000	-0.0387	-0.0657	0.1706	0.0926	0.0930	-0.0010	-0.0124	0.0802	0.0785	0.0598
2000	-0.0587	-0.0604	0.1076	0.0675	0.0788	0.0018	0.0033	0.0585	0.0557	0.0439
4000	-0.0227	-0.0205	0.0776	0.0537	0.0672	-0.0070	-0.0062	0.0348	0.0392	0.0219
8000	-0.0146	-0.0192	0.0566	0.0412	0.0454	-0.0023	-0.0055	0.0251	0.0276	0.0163
14000	-0.0156	-0.0135	0.0463	0.0328	0.0331	0.0026	-0.0013	0.0232	0.0209	0.0177

TABLE 5: Invalidating The Exchangeability Assumption In Kyriazidou

$$u_{i1} \sim 0.5 \cdot N(0,1),\ u_{i2} \sim 2 \cdot N(0,1)$$
$$c_i = 0.6 \cdot N(0,1)$$
$$v_{it} = c_i - u_{it}$$
$$\varepsilon_{i1} = 0.8 \cdot N(0,1) + 0.1 \cdot v_{i1} - 5,\ \varepsilon_{i2} = 0.8 \cdot N(0,1) + 0.9 \cdot v_{i2}$$

Table 5A

Kyriazidou's Estimator: Real First Step Parameters															
	R_2=1					R_2=3					R_2=5				
N	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.0503	-0.0482	0.1986	0.1610	0.1234	-0.0983	-0.0934	0.1857	0.1534	0.1192	-0.1113	-0.1083	0.1868	0.1528	0.1237
500	-0.0666	-0.0866	0.1713	0.1221	0.1393	-0.1109	-0.1217	0.1702	0.1133	0.1344	-0.1214	-0.1373	0.1730	0.1123	0.1460
1000	-0.0511	-0.0672	0.1198	0.0911	0.0824	-0.0995	-0.0976	0.1313	0.0822	0.0976	-0.1139	-0.1148	0.1398	0.0802	0.1148
2000	-0.0359	-0.0345	0.0948	0.0693	0.0679	-0.0874	-0.0880	0.1113	0.0600	0.0880	-0.1078	-0.1124	0.1251	0.0573	0.1124
4000	-0.0406	-0.0416	0.0775	0.0512	0.0639	-0.0873	-0.0890	0.0992	0.0436	0.0890	-0.1077	-0.1075	0.1158	0.0414	0.1075
8000	-0.0278	-0.0327	0.0613	0.0392	0.0382	-0.0768	-0.0774	0.0854	0.0320	0.0774	-0.1007	-0.1006	0.1062	0.0296	0.1006
14000	-0.0155	-0.0214	0.0453	0.0324	0.0319	-0.0665	-0.0711	0.0739	0.0253	0.0711	-0.0938	-0.0963	0.0979	0.0229	0.0963

Table 5B

Kyriazidou's Estimator: Estimated First Step Parameters R_2=1, R_1=3						Wooldridge's Estimator				
N	Mean Bias	Median Bias	SE	ASE	MAD	Mean Bias	Median Bias	SE	ASE	MAD
250	-0.0797	-0.0947	0.2749	0.1611	0.1670	-0.0089	-0.0281	0.1521	0.1812	0.1013
500	-0.0691	-0.0909	0.2614	0.1284	0.1329	0.0045	-0.0140	0.1424	0.1303	0.1025
1000	-0.0445	-0.0490	0.1629	0.0936	0.0875	-0.0005	-0.0077	0.1009	0.0937	0.0742
2000	-0.0513	-0.0603	0.1068	0.0678	0.0756	-0.0008	0.0057	0.0703	0.0642	0.0506
4000	-0.0180	-0.0313	0.0756	0.0539	0.0543	0.0011	-0.0021	0.0446	0.0454	0.0287
8000	-0.0072	-0.0161	0.0619	0.0412	0.0431	-0.0039	-0.0071	0.0318	0.0320	0.0227
14000	-0.0146	-0.0150	0.0443	0.0330	0.0272	0.0007	0.0012	0.0240	0.0240	0.0167

correlation coefficients over time for the idiosyncratic errors in the main equation and the random terms in the selection equation. There exists a difference between the methods on the way the time-varying error in the main equation relates to the time-varying error in the selection equation. To satisfy the *conditional exchangeability* assumption in Kyriazidou (1997) constraints the correlation between ε_{i1} and u_{i1} to be equal to the correlation between ε_{i2} and u_{i2}. However, the *conditional mean independence* assumption in Wooldridge (1995) imposes no restriction on this correlation and, accordingly, it can be different over time. The difference between Tables 4 and 5 is that in the latter a varying mean of the idiosyncratic errors in the main equation from 0 (in period 2) to -5 (in period 1), is introduced. This does not affect, in principle, any of the estimators. For this to be true, in Kyriazidou's (1997) estimator we have included a constant term in the model in differences to treat properly the case of a change in means over time. For Wooldridge's (1995) estimator, the minimum distance step allows for a time-varying intercept.

Tables 4A and 5A report the results for Kyriazidou's (1997) bias corrected estimator using the true γ in the construction of the kernel weights. On the left-hand side of Tables 4B and 5B we focus on the case where γ is estimated by SCMSE. Tables 4B and 5B report on the right-hand side the results for Wooldridge's (1995) estimator.

From Tables 4 and 5 we see that Kyriazidou's (1997) estimator has larger finite-sample bias than Wooldridge's (1995) estimator. Furthermore, the former estimator becomes quite imprecise if we look at the standard errors. The bias are all negative and increase as the kernel order increases. Standard errors are always worse

than the asymptotic ones. The main effect of breaking down the exchangeability condition is in terms of precision in the estimates. As it can be seen in Tables 4B and 5B, Wooldridge's (1995) behaves well in terms of all the considered measures. Wooldridge's (1995) is robust to violations of the *conditional exchangeability* assumption. We do not need large samples for Wooldridge's (1995) estimator to achieve reasonable agreement between the finite-sample standard errors and the results of asymptotic theory.

2.4 Concluding Remarks and Extensions

This chapter reviews 2 two-step "fixed effects" type estimators for the panel data sample selection model. We focus on the recently developed methods by Wooldridge (1995) and Kyriazidou (1997). The chapter is concerned about the finite sample performance of both methods when estimating the parameters of interest under different settings.

The finite sample properties of the estimators are investigated by Monte Carlo experiments. The results of our small Monte Carlo simulation study show the following. First, and for the Monte Carlo design where all of the assumptions in the methods hold, Wooldridge's (1995) estimator is less biased than Kyriazidou's (1997) estimator and it reaches faster its asymptotic behaviour. Second, Wooldridge's (1995) suffers from an important misspecification bias problem when the linear projection functional form for the individual effects in the main equation is invalidated.

However, breaking down the linearity assumption for the individual effects in the selection equation hardly influences the bias for the second step estimates. In contrast to Wooldridge's (1995) estimator, Kyriazidou's (1997) method is free from misspecification problems affecting the individual effects in both equations. Third, the estimators are not robust to the violation of the underlying strict exogeneity assumption. Finally, Wooldridge's (1995) estimator is robust to violations of the *conditional exchangeability* assumption. Under this scenario, the main effect on Kyriazidou's (1997) estimator is in terms of precision in the estimates, with SE always worse than the ASE. Furthermore, we get larger finite-sample bias than in Wooldridge's (1995) estimator.

It would be interesting to develop complementary estimators relaxing some of the assumptions in the above methods. In particular, the need to parameterize the conditional mean of the individual effects in the main equation, as it occurs in Wooldridge's (1995) method, and the need of a *conditional exchangeability* assumption for the idiosyncratic errors in the model, as it is the case in Kyriazidou (1997). This is left for chapter 3.

Chapter 3

A New Estimator for Panel Data Sample Selection Models*

3.1 Introduction

In this chapter 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

* Thanks are owed to Manuel Arellano, Herman Bierens, Richard Blundell, Erwin Charlier, Bo Honoré, Joel Horowitz, Hidehiko Ichimura, Ekaterini Kyriazidou, Myoung-jae Lee, Daniel McFadden, Costas Meghir, Bertrand Melenberg, Scott Thompson, Francis Vella, Frank Windmeijer, a co-editor and two anonymous referees for their very helpful comments on preliminary drafts of this chapter. Earlier versions of the chapter 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.

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 chapter 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 chapter 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 chapter 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 chapter 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.

3.2 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,

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

$$d_{it}^{*} = z_{it}\gamma - \eta_i - u_{it}; \quad d_{it} = 1\left[d_{it}^{*} \geq 0\right], \tag{2.2}$$

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:

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

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

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 non-parametric 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[11]. In any case, in this chapter we derive the sample selection correction terms under normality. The multivariate normal distribution is the most commonly specified assumption on sample selection models[12].

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

[11] 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.

[12] 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 normal, 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.

$(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is})$ 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 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 (A1A) specifies a functional form for that relation, although under our less parametric assumption (A1B) 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 A1-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:

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

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

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

where

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

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[13].

The estimation equation is given by

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

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 $E(e_{its} | x_i, z_i, \nu_{it} \le H_{it}, \nu_{is} \le H_{is}) = 0$. 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}[14]. 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

---

[13] 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 $\phi_2(M_{it}, M_{is}, \rho_{ts}) = \phi(M_{is})\phi\left((M_{it} - \rho_{ts} M_{is}) / (1 - \rho^2_{ts})^{1/2}\right)$. The following also holds: $\int_{-\infty}^{M_{it}} \phi(M_{is})\phi\left((M_{it} - \rho_{ts} M_{is}) / (1 - \rho^2_{ts})^{1/2}\right) dM_{it} = \phi(M_{is})\Phi\left((M_{it} - \rho_{ts} M_{is}) / (1 - \rho^2_{ts})^{1/2}\right)$. 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.

[14] 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, 1984), Newey (1994a), and Arellano and Carrasco (1996)). 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.

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.

3.3 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))[15]. With this parameterization of the individual effects we go from the structural equation in (2.2) to the following reduced form selection rule[16]:

$$d_{it} = 1\{\gamma_{t0} + z_{i1}\gamma_{t1} + ... + z_{iT}\gamma_{tT} - v_{it} \geq 0\} \equiv 1\{H_{it} - v_{it} \geq 0\}. \qquad (3.1)$$

---

[15] 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.

[16] 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

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

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).

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 function are $\text{Prob}(D_t = d_{it}, D_s = d_{is}) = \Phi_2(q_{it} M_{it}, q_{is} M_{is}, \rho^*_{its})$, 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 $\hat{h}_\tau(z_i) = \hat{E}(d_{i\tau}|z_i)$, such as kernel estimators[17]. The way to recover estimated values for the $M_i's$ is given by the inversion $\hat{M}_{i\tau} = \Phi^{-1}[\hat{h}_\tau(z_i)]$. In contrast to Manski's (1987) or Kyriazidou's (1997) semiparametric individual effects models, u_{it} is allowed to be heteroskedastic over

[17] 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.

time[18]. 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[19]. 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 $\hat{h}_\tau(z_i)$ 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 $\hat{h}_\tau(z_i)$ is

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

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

---

[18] 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.

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

For practical issues one needs to choose the kernel function K and a particular 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 $T\cdot f \geq 1$ and even $R+1 \geq 2$ (where R is an odd integer ≥ 1) is the following[20]. For $z \in R^{T\cdot f}$ and $\frac{R+1}{2} \geq 1$ let

$$K_{T\cdot f,R+1}\left(\frac{z_i - z_j}{c_N}\right) = \sum_{p=1}^{\frac{R+1}{2}} \frac{\theta_p \exp\left(-\frac{1}{2}\left(\frac{z_i - z_j}{c_N}\right)' \Omega^{-1}\left(\frac{z_i - z_j}{c_N}\right) / \mu_p^2\right)}{\left(\sqrt{2\pi}\right)^{T\cdot f} \cdot |\mu_p|^{T\cdot f} \sqrt{\det(\Omega)}} \tag{3.3}$$

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

$$\begin{aligned} &\sum_{p=1}^{\frac{R+1}{2}} \theta_p = 1; \\ &\sum_{p=1}^{\frac{R+1}{2}} \theta_p \cdot \mu_p^{2\upsilon} = 0 \end{aligned} \tag{3.4}$$

[20] 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).

for $\upsilon = 1,2,\ldots,\frac{R+1}{2}-1$. We should specify $\Omega = \hat{V}$, where $\hat{V}$ is the sample variance matrix; that is, $\hat{V} = \frac{1}{N}\sum_{j=1}^{N}\left(z_j - \bar{z}\right)'\left(z_j - \bar{z}\right)$ with $\bar{z} = \frac{1}{N}\sum_{j=1}^{N} z_j$. Thus, for $R+1 = 2,4,6,\ldots$, we get

$$\hat{K}_{T\cdot f,R+1}\left(\frac{z_i - z_j}{c_N}\right) = \sum_{p=1}^{\frac{R+1}{2}} \frac{\theta_p \exp\left(-\frac{1}{2}\left(\frac{z_i - z_j}{c_N}\right)'\hat{V}^{-1}\left(\frac{z_i - z_j}{c_N}\right)/\mu_p^2\right)}{\left(\sqrt{2\pi}\right)^{T\cdot f}\cdot\left|\mu_p\right|^{T\cdot f}\sqrt{\det\left(\hat{V}\right)}} \tag{3.5}$$

We will now focus on the problem of bandwidth selection. We need the convergence rate of $\hat{h}$ 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,

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

is stochastically bounded. Clearly, the best uniform consistency rate is obtained for c_N such that $\min\left(c_N{}^{T\cdot f}\sqrt{N}, c_N{}^{-(R+1)}\right)$ is maximal. This is the case if $c_N \propto N^{-1/[2(R+1)+2T\cdot f]}$. We then have $\min\left(c_N{}^{T\cdot f}\sqrt{N}, c_N{}^{-(R+1)}\right) \propto N^{(R+1)/[2(R+1)+2T\cdot f]}$. 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 $R+1 \geq T \cdot f$.

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, $\mathrm{MSE}\{\hat{h}_\tau(z)\}$, 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^*.

3.4 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 $(t,\ s)$, $t \neq s$[21]; 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 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)[22]. The latter group of parameters is left unconstrained in the minimum distance step. The number of parameters associated to the correction

[21] 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 $d_{it} = 0$.

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

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 $\hat{h}_\tau(z_i) = \hat{E}(d_{i\tau}|z_i)$ in section 3 obtained by using high-dimensional smoothing techniques[23]. To overcome this difficulty in nonparametric estimation some dimension reduction approaches have been proposed. A common restriction is additive separability among the variables in the selection equation that would allow to move from high-dimension multivariate kernels to lower-dimension or even univariate kernels[24]. 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.

[23] Estimation precision decreases as the dimension of z_i increases.

[24] For some of these approaches see Härdle and Chen (1995) and Horowitz (1998).

3.5 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

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

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[25] $\eta_i = -\left[\left(z_{1i1} + z_{1i2}\right)/2 + \left(z_{2i1} + z_{2i2}\right)/2 + N(0,1) + 0.07\right]$ and $\alpha_i = \left(x_{i1} + x_{i2}\right)/2 + \sqrt{2}\cdot N(0,1) + 1$; 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.

[25] 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 the 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 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 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

TABLE 1: Invalidating The Exchangeability Assumption In Kyriazidou

$$U1 = 0.8*N(0,1);$$
$$U2 = 2*N(0,1);$$
$$\varepsilon 1 = 0.1*\ U1\ -5 + 0.6*N(0,1);$$
$$\varepsilon 2 = 0.9*\ U2 + 0.6*N(0,1);$$

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	0.05	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	0.12
750	0.05	0.04	0.14	0.12	0.10	0.03	0.04	0.14	0.12	0.08	0.07	0.08	0.12	0.11	0.10	0.08	0.09	0.12	0.11	0.10
1000	0.03	0.03	0.12	0.10	0.09	0.04	0.05	0.16	0.10	0.10	0.07	0.08	0.13	0.09	0.09	0.08	0.09	0.13	0.09	0.10

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[26]. 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[27]. 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[28] $c_N = c \cdot N^{-1/[2(R+1)+1]}$, where the optimal constant c^* is obtained by the *plug-in* method described in Härdle and Linton (1994) with an initial $c = 1$. In both cases, we present the bias corrected estimator[29]. For our less parametric new

[26] The problem could have been solved by the inclusion of time dummies in the main equation.
[27] For details on the latter, see Horowitz (1992), Kyriazidou (1994) and Charlier et al. (1995).
[28] By following the maximum rates of convergence in distribution for the univariate case according to Bierens (1987).
[29] 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 $\delta = 0.5$.

estimator (LPNE) we also used second order kernels[30]. 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 $\hat{h}_\tau(z_i)$. The summation in (3.2) should read $j \neq i$. The bandwidth sequence for the LPNE is[31] $c_N = c \cdot N^{-1/[2(R+1)+2T\cdot f]}$. 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

---

[30] 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 $\hat{h}_\tau(z_i)$ not in between zero and one for some τ and i, so that the inverses $\Phi^{-1}\left[\hat{h}_\tau(z_i)\right]$ 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 $\Phi^{-1}\left[\hat{h}_\tau(z_i)\right]$.

[31] 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 $c_N = c \cdot N^{-1/[2(R+1)+T\cdot f]}$.

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 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 $\alpha_i's$ 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

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

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

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 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

$$U = \text{Uniform}(0,1);$$

$$\varepsilon = 0.8*U + 0.6*\text{Uniform}(0,1);$$

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

$$\eta = -[(Z11 + Z12)/2 + (Z21 + Z22)/2 + \text{Uniform}(0,1) + 0.07];$$

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 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

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 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.

3.6 Concluding Remarks and Extensions

In this chapter 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[32]. 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.

[32] In chapter 5 we present new estimators with some similar properties as the estimator in this chapter but which relax normality or any other parametric assumption for the errors distribution.

3.7 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 $d_{i\tau}=1\{z_i\gamma_\tau-v_{i\tau}\geq 0\}$ for $\tau=t,s$, $E[(y_{it}-y_{is})|x_i,z_i,v_{it}\leq z_i\gamma_t,v_{is}\leq z_i\gamma_s]=(x_{it}-x_{is})\beta+\ell_{ts}\cdot\lambda_{its}+\ell_{st}\cdot\lambda_{ist}$,

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

The two-stage estimation goes as follows. First, we estimate $\varpi_{ts}=(\gamma_t',\gamma_s',\rho_{ts})'$ by $\hat{\varpi}_{ts}=(\hat{\gamma}_t',\hat{\gamma}_s',\hat{\rho}_{ts})'$ 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 $(\hat{\lambda}_{its},\ \hat{\lambda}_{ist})$ 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 $\hat{\beta}$ and $(\hat{\ell}_{ts},\hat{\ell}_{st})'$ in the second stage is

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

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

Define

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

and observe[33]

$$\sqrt{N}\left(\hat{\varpi}_{ts} - \varpi_{ts}\right) =^{p} \frac{1}{\sqrt{N}} \sum_{i} I_{\varpi_{ts}}^{-1} \cdot \begin{bmatrix} z_i'\left(q_{it}\phi_{it}\Phi_{its}/\Phi_{2,its}\right) \\ z_i'\left(q_{is}\phi_{is}\Phi_{ist}/\Phi_{2,its}\right) \\ q_{it}q_{is}\phi_{2,its}/\Phi_{2,its} \end{bmatrix} \equiv \frac{1}{\sqrt{N}} \sum_{i} \Lambda_i, \tag{I.2}$$

where $I_{\varpi_{ts}}$ is the bivariate probit information matrix for ϖ_{ts}, $\phi_{it} \equiv \phi\left[q_{it}z_i\gamma_t\right]$,

$\phi_{ts} \equiv \phi\left[q_{ts}z_i\gamma_s\right]$,

$\Phi_{its} \equiv \Phi\left[\left(q_{is}z_i\gamma_s - \rho^*_{its}q_{it}z_i\gamma_t\right)/\left(1-\rho^{*2}_{its}\right)^{1/2}\right]$, $\Phi_{ist} \equiv \Phi\left[\left(q_{it}z_i\gamma_t - \rho^*_{its}q_{is}z_i\gamma_s\right)/\left(1-\rho^{*2}_{its}\right)^{1/2}\right]$,

$\Phi_{2,its} \equiv \Phi_2\left\{q_{it}z_i\gamma_t, q_{is}z_i\gamma_s, \rho^*_{its}\right\}$ and $\phi_{2,its} \equiv \phi_2\left\{q_{it}z_i\gamma_t, q_{is}z_i\gamma_s, \rho^*_{its}\right\}$.

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

The so called *delta method* yields[34]

---

[33] The notation $=^{p}$ denotes convergence in probability.

[34] Look at the section for two-stage *extremum estimators* with finite dimensional first-stage nuisance parameters in Lee (1996).

$$\sqrt{N}\left(\hat{\Pi}_{ts} - \Pi_{ts}\right) =^{p} E^{-1}\left(d_t d_s R_{ts} R'_{ts}\right) \cdot \frac{1}{\sqrt{N}} \sum_i \left\{ d_{it} d_{is} e_{its} R_{its} + A \cdot \Lambda_i \right\} \tag{I.3}$$

where

$$A \equiv E\left\{ d_t d_s \left[\left(-\ell_{ts} \frac{\partial \lambda_{ts}}{\partial z \gamma_t} - \ell_{st} \frac{\partial \lambda_{st}}{\partial z \gamma_t} \right) R_{ts} z \quad \left(-\ell_{ts} \frac{\partial \lambda_{ts}}{\partial z \gamma_s} - \ell_{st} \frac{\partial \lambda_{st}}{\partial z \gamma_s} \right) R_{ts} z \quad \left(-\ell_{ts} \frac{\partial \lambda_{ts}}{\partial \rho_{ts}} - \ell_{st} \frac{\partial \lambda_{st}}{\partial \rho_{ts}} \right) R_{ts} \right] \right\}. \tag{I.4}$$

Then

$$\sqrt{N}\left(\hat{\Pi}_{ts} - \Pi_{ts}\right) =^{d} N(0, \Gamma), \tag{I.5}$$

$$\Gamma = E^{-1}\left(d_t d_s R_{ts} R'_{ts}\right) \cdot E\left\{ \left(d_t d_s e_{ts} R_{ts} + A \cdot \Lambda\right)\left(d_t d_s e_{ts} R_{ts} + A \cdot \Lambda\right)' \right\} \cdot E^{-1}\left(d_t d_s R_{ts} R'_{ts}\right).$$

The term $A \cdot \Lambda$ is the effect of the first stage on the second. An estimate 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:

$$\hat{I}^{-1}_{\varpi_{ts}} = \left\{ \frac{1}{N} \sum_i \begin{bmatrix} z_i' \left(q_{it} \hat{\phi}_{it} \hat{\Phi}_{its} / \hat{\Phi}_{2,its} \right) \\ z_i' \left(q_{is} \hat{\phi}_{is} \hat{\Phi}_{ist} / \hat{\Phi}_{2,its} \right) \\ q_{it} q_{is} \hat{\phi}_{2,its} / \hat{\Phi}_{2,its} \end{bmatrix} \cdot \begin{bmatrix} z_i' \left(q_{it} \hat{\phi}_{it} \hat{\Phi}_{its} / \hat{\Phi}_{2,its} \right) \\ z_i' \left(q_{is} \hat{\phi}_{is} \hat{\Phi}_{ist} / \hat{\Phi}_{2,its} \right) \\ q_{it} q_{is} \hat{\phi}_{2,its} / \hat{\Phi}_{2,its} \end{bmatrix}' \right\}^{-1} . \quad \text{(I.6)}$$

3.8 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 $\hat{h}_\tau(z_i)$ using kernel estimators with observations on $(d_{i\tau}, z_i)$. Second, we use the probability estimates to estimate ρ_{ts} by $\hat{\rho}_{ts}$ using a bivariate probit with observations on $\left(d_{it}, d_{is}, \Phi^{-1}\left[\hat{h}_t(z_i)\right], \Phi^{-1}\left[\hat{h}_s(z_i)\right]\right)$. 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 parameter of interest, β, and the coefficients accompanying the sample selection terms $(\ell_{ts}, \ell_{st})'$. The sample moment condition for $\hat{\beta}$ and $(\hat{\ell}_{ts}, \hat{\ell}_{st})'$ in the third stage is

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

where

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

Once the first and the second step estimators have been consistently estimated, the third step estimator can be seen as another two stage semiparametric *extremum estimator* where the first stage estimators are given by the vector of infinite dimensional nuisance parameters[35] $\hat{h}_{ts} = \left(\hat{h}_t, \hat{h}_s\right)' \left(=^p h_{ts} = \left(h_t, h_s\right)'\right)$ and the finite parameter $\hat{\rho}_{ts} \left(=^p \rho_{ts}\right)$. 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

$$\sqrt{N}\left(\hat{\rho}_{ts} - \rho_{ts}\right) =^p \frac{1}{\sqrt{N}} \sum_i I_{\rho_{ts}}^{-1} \cdot \left\{ \left(q_{it} q_{is} \phi_{2,its} / \Phi_{2,its}\right) + E\left[\partial \left(\frac{q_t q_s \phi_{2,ts}}{\Phi_{2,ts}} \middle| z_t \right) \middle/ \partial h'_{ts} \right] \cdot \left[d_{its} - E\left(d_{ts} | z_t\right) \right] \right\}$$
$$\equiv \frac{1}{\sqrt{N}} \sum_i \Lambda_i ,$$

(II.2)

where $I_{\rho_{ts}}$ is the bivariate probit information matrix for ρ_{ts}, $\Phi_{2,its} \equiv \Phi_2 \left\{ q_{it} \Phi^{-1}\left[\hat{h}_t(z_i)\right], q_{is} \Phi^{-1}\left[\hat{h}_s(z_i)\right], \hat{\rho}^*_{its} \right\}$, $\phi_{2,its} \equiv \phi_2 \left\{ q_{it} \Phi^{-1}\left[\hat{h}_t(z_i)\right], q_{is} \Phi^{-1}\left[\hat{h}_s(z_i)\right], \hat{\rho}^*_{its} \right\}$, and $d_{its} = \left(d_{it}, d_{is}\right)'$.

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

[35] 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.

$$\sqrt{N}\left(\hat{\Pi}_{ts}-\Pi_{ts}\right)=^{p} E^{-1}\left(d_t d_s R_{ts} R'_{ts}\right)\cdot\frac{1}{\sqrt{N}}\sum_i\left\{d_{it}d_{ts}e_{its}R_{its}+A\cdot\Lambda_i+\right.$$
$$\left.E\left[\partial\left\{d_t d_s\left\{\Delta y_{ts}-\Delta x_{ts}\beta-\ell_{ts}\cdot\lambda_{ts}-\ell_{st}\cdot\lambda_{st}\right\}R_{ts}\middle|z_i\right\}\middle/\partial h'_{ts}\right]\cdot\left[d_{its}-E\left(d_{ts}\middle|z_i\right)\right]\right\} \quad \text{(II.3)}$$

where

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

Then[36]

$$\sqrt{N}\left(\hat{\Pi}_{ts}-\Pi_{ts}\right)=^{d} N(0,\Gamma), \quad \text{(II.5)}$$

$$\Gamma=E^{-1}\left(d_t d_s R_{ts}R'_{ts}\right)\cdot E\left\{\left(d_t d_s e_{ts}R_{ts}+A\cdot\Lambda+E\left[\partial\left\{d_t d_s\left\{\Delta y_{ts}-\Delta x_{ts}\beta-\ell_{ts}\cdot\lambda_{ts}-\ell_{st}\cdot\lambda_{st}\right\}R_{ts}\middle|z_t\right\}\middle/\partial h'_{ts}\right]\cdot\left[d_{its}-E\left(d_{ts}\middle|z_t\right)\right]\right)\right.$$
$$\left.\left(d_t d_s e_{ts}R_{ts}+A\cdot\Lambda+E\left[\partial\left\{d_t d_s\left\{\Delta y_{ts}-\Delta x_{ts}\beta-\ell_{ts}\cdot\lambda_{ts}-\ell_{st}\cdot\lambda_{st}\right\}R_{ts}\middle|z_t\right\}\middle/\partial h'_{ts}\right]\cdot\left[d_{its}-E\left(d_{ts}\middle|z_t\right)\right]\right)'\right\}\cdot E^{-1}\left(d_t d_s R_{ts}R'_{ts}\right)$$

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 $\hat{\rho}_{ts}$ 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

[36] 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 minimising 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.

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 $E\left[\partial\left\{d_t d_s\left\{\Delta y_{ts} - \Delta x_{ts}\beta - \ell_{ts}\cdot\lambda_{ts} - \ell_{st}\cdot\lambda_{st}\right\}R_{ts}|z_i\right\}/\partial h'_{ts}\right]\cdot\left[d_{its} - E\left(d_{ts}|z_i\right)\right]$ in (II.3) and $E\left[\partial\left(\frac{q_t q_s \phi_{2,ts}}{\Phi_{2,ts}}\middle| z_i\right)\middle/\partial h'_{ts}\right]\cdot\left[d_{its} - E\left(d_{ts}|z_i\right)\right]$ in (II.2) are complex and difficult to calculate, making it hard to form their estimators. There is an alternative estimator, developed in Newey (1992), that does not have these problems[37]. It uses only the form of the functions to derive and the kernels[38] to calculate the estimator. For a scalar ζ the estimator for $E\left[\partial\left(\frac{q_t q_s \phi_{2,ts}}{\Phi_{2,ts}}\middle| z_i\right)\middle/\partial h'_{ts}\right]\cdot\left[d_{its} - E\left(d_{ts}|z_i\right)\right]$ is given by[39]

$$\hat{\vartheta}_i = \left.\frac{\partial}{\partial\zeta}\right|_{\zeta=0}\left\{\frac{1}{N}\sum_j \frac{q_{jt}q_{js}\phi_2\left\{q_{jt}\Phi^{-1}\left[\frac{\hat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\hat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], q_{js}\Phi^{-1}\left[\frac{\hat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\hat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], \hat{\rho}^*_{jts}\right\}}{\Phi_2\left\{q_{jt}\Phi^{-1}\left[\frac{\hat{g}_t(z_j)+\zeta d_{it}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\hat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], q_{js}\Phi^{-1}\left[\frac{\hat{g}_s(z_j)+\zeta d_{is}\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}{\hat{f}(z_j)+\zeta\frac{1}{c_N^{T\cdot f}}K\left(\frac{z_j-z_i}{c_N}\right)}\right], \hat{\rho}^*_{jts}\right\}}\right\} \quad \text{(II.6)}$$

[37] See also Newey (1994b) and Newey and McFadden (1994c).

[38] 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 $(1/N)\sum_i y_i\left(1/c_N^{T\cdot f}\right)K\left(\left(z_j - z_i\right)/c_N\right)$. A kernel estimator of the density of z_j will be a component of the expression before, where y is identically equal to 1.

[39] In practice, we include $\hat{\vartheta}_i - \sum_j \frac{\hat{\vartheta}_j}{N}$.

This estimator can be thought of as the influence of the ith 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).

Newey's (1992) estimator for

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

is given by[40]

$$\hat{\chi}_i = \left.\frac{\partial}{\partial\zeta}\right|_{\zeta=0}\frac{1}{N}\sum_j d_{ji}d_{js}R_{jts}\left\{\Delta y_{jts} - \Delta x_{jts}\beta\right. \tag{II.7}$$

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

[40] In practice, we include $\hat{\chi}_i - \sum_j \frac{\hat{\chi}_j}{N}$.

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

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:

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

Chapter 4

Selection Correction in Panel Data Models: An Application to Labour Supply and Wages*

4.1 Introduction

In many problems of applied econometrics, the equation of interest is only defined for a subset of individuals from the overall population, while the parameters of interest are the parameters that refer to the whole population. Examples are the estimation of wage equations, or hours of work equations, where the dependent variable can only be measured when the individual participates in the labour market. If the sub-population is non-randomly drawn from the overall population, straightforward regression analysis leads to inconsistent parameter estimates. This problem is well known as sample selection bias, and a number of estimators are available which correct for this (see Heckman (1979), or Powell (1994) for an overview).

Another problem is the presence of unobserved heterogeneity in the equation of interest. Economic theory often suggests estimation equations that contain an

* Useful comments and suggestions from Richard Blundell, Christian Dustmann, Arthur van Soest and Frank Windmeijer are gratefully acknowledged. Thanks are also owed to participants at the Primer Encuentro de Economía Aplicada, June 1998, Barcelona, Spain; at the 8th-International Conference on Panel Data, June 1998, Göteborg, Sweden; at the German Socio-Economic Panel Users (GSOEP) Conference, July 1998, Berlin, Germany; at the European Meeting of the Econometrics Society (ESEM), August 1998, Berlin, Germany; at the Royal Economic Society (RES) Annual Conference, March-April 1999, Nottingham, United Kingdom; and at the European Society for Population Economics (ESPE) Conference, June 1999, Torino, Italy.

individual specific effect, which is unobserved, but correlated with the model regressors. Examples are unobserved ability components in wage equations, correlated with wages and education (see Card (1994) for details), or the estimation of Frisch demand functions in the consumption and labour supply literature (see, for instance, Browning, Deaton, and Irish (1985), Blundell and MaCurdy (1999) and MaCurdy (1981)). If unobserved individual specific (and time constant) effects affect the outcome variable, and are correlated with the model regressors, simple regression analysis does not identify the parameters of interest. For the estimation of coefficients on variables which vary over time, panel data provide a solution to this latter problem, and a number of straightforward estimators are available (see Chamberlain (1984), and Hsiao (1986) for overviews).

In many applications, both problems occur simultaneously. If the selection process is time constant, panel estimators solve both problems. But often this is not the case. Recently, some estimators have been proposed which deal with both sources of estimation bias. These estimators require panel data, and produce consistent parameter estimates under various sets of assumptions. We consider three estimators which allow for additive individual specific effects in both the (binary) selection equation and the equation of interest, and, at the same time, allow for the equation of interest being defined for a non-random sub population. These estimators impose different consistency requirements, some of which may be restrictive in particular applications.

The first estimator we consider has been proposed by Wooldridge (1995). It relies on a full parameterisation of the sample selection mechanism, and requires

specifying the functional form of the conditional mean of the individual effects in the equation of interest. It does not impose distributional assumptions about the error terms and the fixed effects in the equation of interest. The second estimator we discuss has been proposed by Kyriazidou (1997). The basic idea of this estimator is to match observations within individuals, which have the same selection effect in two time periods, and to difference out both the individual heterogeneity term, and the selection term. The third estimator has been developed in chapter 3. This method adds a distributional assumption for the error term in the equation of interest.

In the first part of the chapter, we describe the main features of the three estimators, and point out the conditions under which each of them produces consistent estimates of the parameters of interest. Not many applications of these estimators exist in the literature. In the second part of the chapter, we apply the three methods to a typical problem in labour economics. We estimate wage equations for female labour market participants, and try to identify the effect of actual labour market experience on wages. In this application, all the before mentioned problems arise. Female labour market participants are non-randomly drawn from the overall population. Their participation propensity depends on unobservables, which are likely to be correlated with the model regressors. And their productivity depends on unobservables, which are likely to be correlated with the regressors in the main equation.

All three estimators impose the assumption of strict exogeneity of the explanatory variables. In many typical applications, like the one we use as an illustration, this assumption is likely to be violated. We show how all three estimators can be extended to relax this assumption in the main equation, maintaining only the

strict exogeneity of the regressors in the selection equation. We apply the extensions of the estimators to our particular problem, and compare the emerging estimates.

Another problem which frequently occurs with panel data is measurement error in some of the explanatory variables. With most panel surveys, the construction of work history variables needs to be based on retrospective information, which is likely to suffer from measurement error. If the affected variables enter the equation of interest in a non-linear manner, IV estimation may not solve the problem. We show how to address this problem within the methods discussed.

The data for our empirical application is drawn from the German Socio-Economic Panel (GSOEP). The dataset used for estimation is based on the first 12 waves of the panel.

The chapter is organised as follows. In the next section we describe briefly the three estimators and their underlying assumptions. Section 3 compares the estimators. Section 4 discusses problems of implementation, and describes extensions to the case where strict exogeneity of some of the model regressors in the main equation is violated. Section 5 describes the data and the model we estimate. Section 6 presents the results, and section 7 concludes.

4.2 The Model and Estimators

4.2.1 The Model

The model we consider in the following consists of a binary selection rule, which depends on a linear index, and an unobserved (time constant) additive individual effect, which may be correlated with the model regressors. The selection rule assigns individuals in the overall sample population to two different regimes. For one regime, a linear regression equation is defined, which again has an additive unobserved individual component, correlated with the model regressors. The slope parameters of this equation are the parameters of interest.

This model can be written as:

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

$$d_{it}^{*} = z_{it}\gamma - \eta_t - u_{it}; \quad d_{it} = 1\left[d_{it}^{*} \geq 0\right], \tag{2.2}$$

where 1[.] is an indicator function, which is equal to one if its argument is true, and zero otherwise. Furthermore, β and γ are unknown parameter vectors, and x_{it}, z_{it} are vectors of explanatory variables with possibly common elements[41], including both time variant and time invariant variables, and time effects. The α_i and η_t are

[41] For some estimators exclusion restrictions are not required because distributional assumptions (like normality of the error terms) identify the model. We assume throughout that there are exclusion restrictions in (1).

unobservable and time invariant individual specific effects, which are possibly correlated with x_{it} and z_{it}. The ε_{it} and u_{it} are unobserved disturbances. The variable w_{it} is only observable if $d_{it} = 1$. The parameter vector we seek to estimate is β.

We assume that panel data is available. Equation (2.1) could be estimated in levels by pooled ordinary least squares (OLS). This will lead to consistent estimates of β under the following condition:

$$E(\alpha_i + \varepsilon_{it} | x_{it}, d_{it} = 1) = E(\alpha_i | x_{it}, d_{it} = 1) + E(\varepsilon_{it} | x_{it}, d_{it} = 1) = 0, \qquad \forall t. \tag{2.3}$$

Accordingly, OLS estimates on the selected subsample are inconsistent if selection is non-random, and/or if correlated individual heterogeneity is present. In both cases, the conditional expectation in (2.3) is unequal to zero.

One way to eliminate the fixed effects α_i is to use some type of difference estimator. Given identification[42], the consistency condition for an estimator using differences across time instead of level equations is given by the following expression: [43]

$$E(\varepsilon_{it} - \varepsilon_{is} | x_{it}, x_{is}, d_{it} = d_{is} = 1) = 0, \qquad s \neq t, \tag{2.4}$$

---

[42] For identification we require the matrix $E\left[(x_t - x_s)'(x_t - x_s) | x_t, x_s, d_t = d_s = 1\right]$ to be finite and non-singular.

[43] If $s = t - 1$, the data is transformed by applying first differencing over time. Other transformations include mean deviation operators.

where s and t are time periods.

Since condition (2.4) puts no restrictions on how the selection mechanism or the regressors relate to α_i, differencing equation (2.1) across time not only eliminates the problem of correlated individual heterogeneity but also any potential selection problem which operates through α_i.

If conditions (2.3) or (2.4) are satisfied, the OLS estimator or the difference estimator respectively lead to consistent estimates. No specification of the selection process is necessary. If conditions (2.3) and (2.4) are violated, consistent estimation requires to model the selection process. The estimators we describe in the next section take these consistency requirements (2.3) or (2.4) as a starting point. The idea of the estimator by Wooldridge (1995) is to derive an expression for the expected value in (2.3), and to add it as an additional regressor to the equation of interest. The estimator in chapter 3 derives an expression for the expected value in (2.4), which is then added as an additional regressor to the differenced equation. The estimator by Kyriazidou (1997) matches pairs of observations for a given individual for whom the conditional expectation in (2.4) is equal to zero.

4.2.2 Estimation in Levels: Wooldridge's Estimator

The estimation method developed by Wooldridge (1995) relies on level equations. The basic idea is to parameterise the conditional expectations in (2.3) and to add these expressions as additional regressors to the main equation. The method is semiparametric with respect to the main equation, in the sense that it does not require joint normality of the errors in both equations. Similar to Heckman's (1979) two-stage

estimator, only marginal normality of the errors in the selection equation and a linear conditional mean assumption of the errors in the main equation is required. The time dimension allows controlling for individual effects in addition, which requires further assumptions for the conditional means of the individual effects in both equations. Wooldridge (1995) imposes two assumptions on the selection equation (***W1*** and ***W2*** below), and two assumptions about the relationship between $\alpha_i, \varepsilon_{it}$ and the resulting error term in the selection equation (***W3*** and ***W4***).

• ***W1:*** *The regression function of* η_i *on* z_i *is linear.*

Following Chamberlain (1984), Wooldridge (1995) 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: $\eta_i = z_{i1}\delta_1 + ... + z_{iT}\delta_T + c_i$, where c_i is a random component.

• ***W2:*** *The errors in the selection equation,* $v_{it} = u_{it} + c_i$, *are independent of* $\tilde{z}_i$ *and normal* $\left(0, \sigma_t^2\right)$, *where* $\tilde{z}_i = \left(x_i, z_i\right)'$ *with* $x_i = \left(x_{i1}, ..., x_{iT}\right)$ *and* $z_i = \left(z_{i1}, ..., z_{iT}\right)$.[44]

• ***W3:*** *The regression function of* α_i *on* x_i *and* v_{it} *is linear.*[45]

---

[44] v_{it} is heteroskedastic over time whenever u_{it} is.

[45] An alternative assumption is (see Mundlack (1978), Nijman and Veerbeck (1992), and Zabel (1992)) that α_i depends only on the time average of x_{it}.

Accordingly, $E(\alpha_i|\tilde{z}_i, v_{it}) = x_{i1}\psi_1 + ... + x_{iT}\psi_T + \phi_t v_{it}$.[46] The conditional distribution of α_i on x_i, v_{it} is linear, but otherwise unrestricted. We do not observe v_{it}, however, but only the binary selection indicator d_{it}. Therefore, $E(\alpha_i|\tilde{z}_i, v_{it})$ has to be replaced by the expectation of α_i given $(\tilde{z}_i, d_{it} = 1)$, which is obtained by integrating $E(\alpha_i|\tilde{z}_i, v_{it}) = x_{i1}\psi_1 + ... + x_{iT}\psi_T + \phi_t v_{it}$ over $v_{it} \leq z_{i1}\gamma_{t1} + ... + z_{iT}\gamma_{tT}$.[47] This yields $E(\alpha_i|\tilde{z}_i, d_{it} = 1) = x_{i1}\psi_1 + ... + x_{iT}\psi_T + \phi_t E[v_{it}|\tilde{z}_i, d_{it} = 1]$.

• ***W4:*** *ε_{it} is mean independent of $\tilde{z}_i$ conditional on v_{it} and its conditional mean is linear on v_{it}.*

Accordingly, $E(\varepsilon_{it}|\tilde{z}_i, v_{it}) = E(\varepsilon_{it}|v_{it}) = \rho_t v_{it}$. The first equality states that ε_{it} is mean independent of $\tilde{z}_i$ conditional on v_{it}, and the second equality that $E(\varepsilon_{it}|v_{it})$ is linear. No restrictions are imposed on the temporal dependence of ε_{it}, or on $\text{Corr}(\varepsilon_{it}, v_{is})$, for $s \neq t$. Again, as we do not observe v_{it} but the binary selection

---

[46] The key point for identifying the vector β is that, under v_{it} being independent of $\tilde{z}_i$, and the conditional expectation $E(\alpha_i|\tilde{z}_i, v_{it})$ being linear, the coefficients on the x_{ir}, $r = 1,...,T$, are the same regardless of which v_{it} is in the conditioning set. This is crucial to the approach, and follows from the law of iterated expectations. For any t,

$$\begin{aligned} E(\alpha_i|\tilde{z}_i) &= x_{i1}\psi_{t1} + ... + x_{iT}\psi_{tT} + \phi_t E[v_{it}|\tilde{z}_i] \\ &= x_{i1}\psi_{t1} + ... + x_{iT}\psi_{tT} \\ &= x_{i1}\psi_1 + ... + x_{iT}\psi_T. \end{aligned}$$

The second equality follows because $E[v_{it}|\tilde{z}_i] = 0$ under ***W2***, and the third follows from the coefficients in the linear projection of α_i onto x_i being necessarily time-invariant.

[47] $z_{i1}\gamma_{t1} + ... + z_{iT}\gamma_{tT}$ is the reduced form index for the selection equation in (2.2), once the time-constant unobserved effect η_i is specified as in ***W1***.

indicator d_{it}, we must find the expectation of ε_{it} given $(\tilde{z}_i, d_{it} = 1)$. This is obtained by integrating $E(\varepsilon_{it}|\tilde{z}_i, v_{it}) = \rho_t v_{it}$ over $v_{it} \leq z_{i1}\gamma_{t1} + \ldots + z_{iT}\gamma_{tT}$, resulting in $E(\varepsilon_{it}|\tilde{z}_i, d_{it} = 1) = \rho_t E[v_{it}|\tilde{z}_i, d_{it} = 1]$.

Under assumptions ***W1***– ***W4***, Wooldridge (1995) derives an explicit expression for

$$E(\alpha_t + \varepsilon_{it}|\tilde{z}_t, d_{it} = 1) = E(\alpha_t|\tilde{z}_i, d_{it} = 1) + E(\varepsilon_{it}|\tilde{z}_i, d_{it} = 1) = x_{i1}\psi_1 + \ldots + x_{iT}\psi_T + (\phi_t + \rho_t) \cdot E[v_{it}|\tilde{z}_i, d_{it} = 1] \quad (2.3')$$

which results in the following model:

$$w_{it} = x_{i1}\psi_1 + \ldots + x_{iT}\psi_T + x_{it}\beta + \ell_t \lambda(H_{it}/\sigma_t) + e_{it}, \quad (2.5)$$

where $\ell_t = \phi_t + \rho_t$, $H_{it} = z_{i1}\gamma_{t1} + \ldots + z_{iT}\gamma_{tT}$ is the reduced form index in the selection equation for period t, and $\lambda(H_{it}/\sigma_t) = E[v_{it}|\tilde{z}_i, d_{it} = 1]$.

Notice that, since $d_{ir} = 1$ for $r \neq t$ is not included in the conditioning sets of $E(\alpha_t|\tilde{z}_i, d_{it} = 1)$ and $E(\varepsilon_{it}|\tilde{z}_i, d_{it} = 1)$, the selection term $E[v_{it}|\tilde{z}_i, d_{it} = 1]$ is not strictly exogenous in (2.5). The condition, which holds for the new error term in

(2.5), is $E(e_{it}|\tilde{z}_i, d_{it}=1) = E(e_{it}|\tilde{z}_i, v_{it} \leq H_{it}) = 0$. We call this a "contemporaneous exogeneity" of the selection term $E[v_{it}|\tilde{z}_i, d_{it}=1]$ with respect to e_{it} in (2.5).

To obtain estimates for $\lambda(\cdot)$, a probit on $H_{it} = z_{i1}\gamma_{t1} + ... + z_{iT}\gamma_{tT}$ is estimated for each t in the first step. In the second step, equation (2.5) is estimated either by minimum distance or pooled OLS regression. Under the assumptions ***W1-W4***, the estimator for β is consistent. Since dependence between the unobservables in the selection equation, v_{it}, and the unobservables in the main equation, $(\varepsilon_{it}, \alpha_i)$, is allowed for, selection may depend not only on the error ε_{it}, but also on the unobserved individual effect α_i. For time invariant variables or variables that vary systematically over time, β is not separable from ψ. For time varying variables we can identify β given that the coefficients $\psi_1, ..., \psi_T$ are constant for different time periods (assumption ***W3***).

4.2.3 Estimation in Differences I: Kyriazidou's Estimator

The estimator developed by Kyriazidou (1997) relies on pairwise differences over time applied to model (2.1) for individuals satisfying $d_{it} = d_{is} = 1, s \neq t$. The idea of the estimator is as follows. Re-consider first the expression in (2.4):

$$\begin{aligned} &E(\varepsilon_{it} - \varepsilon_{is}|\tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1) = \\ &\quad E(\varepsilon_{it}|\tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1) - E(\varepsilon_{is}|\tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1) \equiv \\ &\quad \lambda_{its} - \lambda_{ist} \end{aligned} \qquad (2.4')$$

where $\tilde{z}_{it} = (x_{it}, z_{it})'$, $\tilde{z}_{is} = (x_{is}, z_{is})'$, and for each time period the selection terms are

$$\begin{aligned}\lambda_{its} &= \mathrm{E}\left(\varepsilon_{it} \middle| \tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i, u_{it} \le z_{it}\gamma - \eta_i, u_{is} \le z_{is}\gamma - \eta_i\right) \\ &= \Lambda\left(z_{it}\gamma - \eta_i, z_{is}\gamma - \eta_i; F\left(\varepsilon_{it}, u_{it}, u_{is} \middle| \tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i\right)\right)\end{aligned}$$

$$\begin{aligned}\lambda_{ist} &= \mathrm{E}\left(\varepsilon_{is} \middle| \tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i, u_{is} \le z_{is}\gamma - \eta_i, u_{it} \le z_{it}\gamma - \eta_i\right) \\ &= \Lambda\left(z_{is}\gamma - \eta_i, z_{it}\gamma - \eta_i; F\left(\varepsilon_{is}, u_{is}, u_{it} \middle| \tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i\right)\right)\end{aligned}$$

where $\Lambda(\cdot)$ is an unknown function and $F(\cdot)$ is an unknown joint conditional distribution function of the errors. The additional variables in the conditioning set in (2.4'), compared to the conditioning set in expression (2.4), follow from the fact that the sample selection mechanism has to be specified in this model. The individual effects in both equations are allowed to depend on the explanatory variables in an arbitrary way, and are not subject to any distributional assumption. Different to Wooldridge (1995), the individual effects are now included in the conditioning set.

Under the assumption that for individuals for whom $z_{it}\gamma = z_{is}\gamma$ and $d_{it} = d_{is} = 1$, the sample selection effect is equal in t and s (that is, $\lambda_{its} = \lambda_{ist}$ in (2.4')), differencing between periods s and t will entirely remove the sample selection problem and, at the same time, the time constant individual heterogeneity component.

In general however there is no reason to expect that $\lambda_{its} = \lambda_{ist}$ holds even for individuals satisfying the conditions above. In particular, the selection terms depend not only on the conditioning vector $(\tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i)$, but also on the joint conditional

distribution of the error terms for the two time periods, which may differ across individuals, as well as over time for the same individual. To ensure that $\lambda_{its} = \lambda_{ist}$ holds, Kyriazidou (1997) imposes a "conditional exchangeability" assumption. The resulting estimator is semiparametric with respect to both the error distribution and the distribution of the fixed effects.

To implement this estimator, Kyriazidou (1997) imposes the following conditions:

• ***K1:*** $(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is})$ *and* $(\varepsilon_{is}, \varepsilon_{it}, u_{is}, u_{it})$ *are identically distributed conditional on* $\tilde{z}_{it}, \tilde{z}_{is}, \alpha_t, \eta_i$. *That is,* $F(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is} | \tilde{z}_{it}, \tilde{z}_{is}, \alpha_t, \eta_i) = F(\varepsilon_{is}, \varepsilon_{it}, u_{is}, u_{it} | \tilde{z}_{it}, \tilde{z}_{is}, \alpha_t, \eta_i)$.

This "conditional exchangeability" assumption implies that the idiosyncratic errors are homoscedastic over time for a given individual. Under this assumption, any time effects are absorbed into the conditional mean.

• ***K2:*** *An appropriate smoothness condition*[48] *is imposed on the selection correction function* $\Lambda(\cdot)$.

This smoothness condition ensures that once ***K1*** holds, $z_{it}\gamma = z_{is}\gamma$ implies $\lambda_{its} = \lambda_{ist}$.

Under assumptions ***K1-K2*** and provided identification is met,[49] the OLS estimator applied to

---

[48] Kyriazidou (1997) imposes a Lipschitz continuity property on the selection correction function $\Lambda(\cdot)$.

$$w_{it} - w_{is} = (x_{it} - x_{is})\beta + e_{its}, \tag{2.6}$$

for individuals satisfying $d_{it} = d_{is} = 1, s \neq t$ and $z_{it}\gamma = z_{is}\gamma$, is consistent. The resulting error $e_{its} \equiv (\varepsilon_{it} - \varepsilon_{is}) - (\lambda_{its} - \lambda_{ist})$ has a conditional expectation that satisfies $E(e_{its}|\tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1) = 0$.

The estimator requires that there are individuals with $z_{it}\gamma = z_{is}\gamma$ with probability one, which is not the case if z_{it} contains a continuous variable. To implement the estimator, Kyriazidou (1997) constructs kernel weights, which are a declining function of the distance $|z_{it}\gamma - z_{is}\gamma|$, and estimates pairwise differenced equations by weighted OLS[50].

The procedure requires estimates of γ, which can be obtained either by smoothed conditional maximum score estimation (see, for instance, Charlier, Melenberg and van Soest (1997) and Kyriazidou (1997))[51] or conditional logit (Chamberlain (1980)) estimation.

---

[49] In this model identification of β requires $E[(x_t - x_s)'(x_t - x_s)d_t d_s | (z_t - z_s)\gamma = 0]$ to be finite and non-singular. Given that we require support of $(z_t - z_s)\gamma$ at zero, nonsingularity requires an exclusion restriction on the set of regressors, namely that at least one of the variables z_{it} is not contained in x_{it}.

[50] The estimator is arbitrarily close to root n-consistency depending on the degree of smoothness one is willing to assume for the kernel function.

[51] Estimating γ by the smoothed conditional maximum score estimator requires additional assumptions (see Manski (1987), Horowitz (1992), Kyriazidou (1994) and Charlier, Melenberg and van Soest (1997) for details).

4.2.4 Estimation in Differences II: Chapter's 3 Estimator

This estimator is also based on pairwise differencing equation (2.1) for individuals satisfying $d_{it} = d_{is} = 1, s \neq t$. Different from Kyriazidou's (1997) estimator, chapter's 3 estimator relies on a parameterisation of the conditional expectation in (2.4). On the other hand, it does not impose the "conditional exchangeability" assumption.

To implement the estimator, the following assumptions are made:

- ***CH31:*** *The regression function of* η_i *on* z_i *is linear*[52].
- ***CH32:*** *The errors in the selection equation,* $v_{it} = u_{it} + c_i$, *are normal* $\left(0, \sigma_t^2\right)$.
- ***CH33:*** *The errors* $\left[\left(\varepsilon_{it} - \varepsilon_{is}\right), v_{it}, v_{is}\right]$ *are trivariate normally distributed conditional on* $\tilde{z}_i$.

The first two assumptions refer to the selection equation and are equivalent to assumptions ***W1*** and ***W2*** above. The third assumption imposes restrictions on the joint conditional distribution of the error terms in the two equations. The method is non-parametric with respect to the individual effects in the main equation and allows, under its semi-parametric version, for a non-parametric conditional mean of the individual effects in the selection equation on the leads and lags of the explanatory variables in that equation.

Under assumptions ***CH31-CH33,*** the resulting estimation equation is given by

---

[52] This assumption corresponds to the "more parametric new estimator" in chapter 3. There, a non-parametric specification of the conditional mean of η_i is also proposed. In that case, $E\left(\eta_i | z_i\right)$ is left unrestricted.

$$w_{it} - w_{is} = (x_{it} - x_{is})\beta + \ell_{ts}\lambda\left(\frac{H_{it}}{\sigma_t}, \frac{H_{is}}{\sigma_s}, \rho_{ts}\right) + \ell_{st}\lambda\left(\frac{H_{is}}{\sigma_s}, \frac{H_{it}}{\sigma_t}, \rho_{ts}\right) + e_{its}, \qquad (2.7)$$

where $H_{i\tau} = z_{i1}\gamma_{\tau 1} + \ldots + z_{iT}\gamma_{\tau T}$, $\tau = t, s$, are the resulting reduced form indices in the selection equation for periods t and s, and $\rho_{ts} = \rho_{(\nu_t/\sigma_t)(\nu_s/\sigma_s)}$ is the correlation coefficient between the errors in the selection equation. Furthermore, $\ell_{ts}\lambda\left(\frac{H_{it}}{\sigma_t}, \frac{H_{is}}{\sigma_s}, \rho_{ts}\right) + \ell_{st}\lambda\left(\frac{H_{is}}{\sigma_s}, \frac{H_{it}}{\sigma_t}, \rho_{ts}\right)$ is the conditional mean $E(\varepsilon_{it} - \varepsilon_{is} | \tilde{z}_i, d_{it} = d_{is} = 1)$ derived from the three-dimensional normal distribution assumption in ***CH33***.[53] The new error term $e_{its} \equiv (\varepsilon_{it} - \varepsilon_{is}) - [\ell_{ts}\lambda_{its} + \ell_{st}\lambda_{ist}]$ has a conditional expectation $E(e_{its} | \tilde{z}_i, \nu_{it} \le H_{it}, \nu_{is} \le H_{is}) = 0$. To construct estimates of the $\lambda(\cdot)$ terms the reduced form coefficients (γ_t, γ_s) will be jointly determined with

---

[53] In the case of the errors in the selection equation being uncorrelated ($\rho_{ts} = 0$), then $E(\varepsilon_{it} - \varepsilon_{is} | \tilde{z}_i, d_{it} = d_{is} = 1) = \ell_t\lambda(H_{it}/\sigma_t) + \ell_s\lambda(H_{is}/\sigma_s)$ where $\lambda(H_{it}/\sigma_t) = E[\nu_{it} | \tilde{z}_i, d_{it} = 1]$ and $\lambda(H_{is}/\sigma_s) = E[\nu_{is} | \tilde{z}_i, d_{is} = 1]$. Given that we define both ν_{it}, ν_{is} as $c_i + u_{i\tau}$, $\tau = t, s$, thus $\rho_{ts} \neq 0$ and the conditional expectation is more complex:

$$E(\varepsilon_{it} - \varepsilon_{is} | \tilde{z}_i, d_{it} = d_{is} = 1) = \ell_{ts}\lambda\left(\frac{H_{it}}{\sigma_t}, \frac{H_{is}}{\sigma_s}, \rho_{ts}\right) + \ell_{st}\lambda\left(\frac{H_{is}}{\sigma_s}, \frac{H_{it}}{\sigma_t}, \rho_{ts}\right),$$

where

$$\lambda\left(\frac{H_{it}}{\sigma_t}, \frac{H_{is}}{\sigma_s}, \rho_{ts}\right) = E[\nu_{it} | \tilde{z}_i, d_{it} = d_{is} = 1] \quad \text{and}$$

$$\lambda\left(\frac{H_{is}}{\sigma_s}, \frac{H_{it}}{\sigma_t}, \rho_{ts}\right) = E[\nu_{is} | \tilde{z}_i, d_{it} = d_{is} = 1].$$

See chapter 3 for details on the more complex lambda functions. Expressions there are derived by following the work of Tallis (1961) for moments of a truncated multivariate normal distribution.

ρ_{ts}, using a bivariate probit for each combination of time periods. The second step is carried out by applying OLS to equation (2.7).

4.3 Comparison of Estimators

Table 1 summarises the main features of the three estimators, and the assumptions they impose on the data. Wooldridge's (1995) method is the only one that relies on level equations. This makes it necessary to specify the functional form for the conditional mean of the individual effects in the main equation, α_i, with respect to the explanatory variables (to allow for individual correlated heterogeneity) and with respect to the random error term v_{it} (to allow for selection that depends on the unobserved effect α_i). In the other methods, α_i is differenced out, and selection may therefore depend on α_i in an *arbitrary* fashion.

With respect to the assumptions on the functional form of the sample selection effects, Kyriazidou (1997) treats them as unknown functions, which need not to be estimated. Wooldridge (1995) and the estimator in chapter 3 parameterise these effects, which imposes three assumptions. First, a normality assumption for the random component of the unobservables in the selection equation ($v_{it} = c_i + u_{it}$). Secondly, to explicitly modelling the dependence of η_i on the explanatory variables. Thirdly, an assumption about the relationship between the errors in the main equation and the v_{it} in the selection equation. In Wooldridge (1995) joint normality of

TABLE 1: COMPARATION OF ESTIMATORS

Estimators	Estimation	Sample selection effects	Distributional assumptions				Specification of conditional means		
			α_i	η_i	ε_{it}	u_{it}	α_I	η_i	ε_{it}
Wooldridge	Levels	Parameterized	None	Normal random component c_i	None	Normal	LP[a] on x_t & $v_{it}=c_i+u_{it}$	LP[a] on z_i	LP[a] on v_{it}
Kyriazidou	Time diff.	Unspecified	None	None	None but CE[b]	None but CE[b]	None	None	None
Chapter's 3	Time diff.	Parameterized	None	Normal random component c_i	Normal	Normal	None	LP[a] on z_i/non-parametric	Linearity from joint normality

Estimators	Time series properties				Sample requirements
	Time dummies or time trend	Time Heterosk.	Serial correlation	Corr(ε_{it},u_{is}) $t \neq s$	
Wooldridge	Yes	Yes	Yes	Unspecified	$d_{it} = 1$
Kyriazidou	Yes	No	CE[b]	CE[b]	$d_{it} = d_{is} = 1,$ $z_{it}\gamma \cong z_{is}\gamma$
Chapter's 3	Yes	Yes	Yes	subject to joint normality	$d_{it} = d_{is} = 1$

[a] LP denotes the linear projection operator.

[b] Subject to the "conditional exchangeability" (CE) assumption according to which the vectors of errors $(\varepsilon_{it}, \varepsilon_{is}, u_{it}, u_{is})$ and $(\varepsilon_{is}, \varepsilon_{it}, u_{is}, u_{it})$ are identically distributed conditional on $\tilde{z}_{it}, \tilde{z}_{is}, \alpha_i, \eta_i$.

unobservables in both equations is not needed once a marginal normality assumption for the v_{it} and a linear projection specification for ε_{it} on v_{it} are imposed. In chapter's 3 estimator, joint normality is assumed, and linearity between ε_{it} and v_{it} results from the joint normality assumption.

Kyriazidou (1997) does not impose any parametric assumption about the distribution of the unobservables in the model. However, the conditional exchangeability assumption in Kyriazidou's (1997) estimator imposes restrictions on the time series properties of the model. This assumption is more demanding than joint conditional stationarity for the time-varying errors (see Kyriazidou (1997) for details). While in Wooldridge (1995) and the estimator in chapter 3 not only the conditional means, but also the second moments of the error terms may incorporate time effects, Kyriazidou's (1997) estimator allows only for time effects in the conditional mean.

All these methods do not impose explicitly restrictions on the pattern of serial-correlation in the error processes. However, in Kyriazidou (1997) serial correlation is allowed as far as this does not invalidate the "conditional exchangeability" assumption. Wooldridge's (1995) method imposes no restriction on the way the time-varying error in the main equation (ε_{it}) relates to the time-varying error in the selection equation (v_{is}), for $s \neq t$. Different to Wooldridge (1995), in chapter's 3 estimator the joint normality assumption (**CH33** above) extends linearity to the correlation between ε_{it} and v_{is} for $s \neq t$, since it includes in the conditioning set not only d_{it}, but d_{it}, d_{is}.

The estimators differ in terms of sample requirements. In Wooldridge (1995) the parameters of interest are estimated from those observations that have $d_{it} = 1$. Chapter's 3 estimator uses individuals with $d_{it} = d_{is} = 1$. Kyriazidou (1997) uses those observations that have $d_{it} = d_{is} = 1$, and for which $z_{it}\gamma$ and $z_{is}\gamma$ are "close". Asymptotically, the effective sample size is smaller for the latter method.

At the stage of implementation, problems may arise with Kyriazidou's (1997) method if there are strong time effects in the selection equation. In this case, it may be difficult to find observations for which $z_{it}\gamma$ and $z_{is}\gamma$ are "close". Furthermore, identification problems arise if for individuals for whom $z_{it}\gamma$ and $z_{is}\gamma$ are "close", also x_{it} is "close" to x_{is}. In this case, a higher weight is given to observations with little time-variation in the explanatory variables in the main equation. A related problem arises if high matching weights are assigned to observations whose x variables change in a systematic manner. In this case it is not possible to separately identify the coefficients of these variables from coefficients on a time trend, or time dummies. These problems are likely to occur in many empirical applications, as we demonstrate below.

4.4 Extensions

4.4.1 Estimation if Regressors are Non-Strictly Exogenous

All the estimators above assume strict exogeneity of the regressors. The variable x_{it} is strictly exogenous relative to ε_{it} if

$$E(\varepsilon_{it}|x_i) = 0, \quad t = 1,\ldots,T. \tag{4.1}$$

A similar statement can be made about z_{it} with respect to u_{it}. If $E(\varepsilon_{it}|x_{it}) = 0$, we call this contemporaneous exogeneity.

In many empirical applications, the strict exogeneity condition (after controlling for both individual heterogeneity and sample selection) is likely to be violated. In the following, we describe how the above three estimators can be extended in this direction. We maintain the strict exogeneity assumption of regressors in the selection equation.

In Wooldridge (1995), the selection correction proposed has been derived under the assumption of strict exogeneity of the regressors conditional on the unobserved effect, that is, $E(\varepsilon_{it}|x_i,\alpha_i) = 0$. The strict exogeneity assumption is, for instance, needed for condition ***W3*** to be valid. To see this, suppose that the variables in the equation of interest are predetermined, and possibly correlated with the individual effects α_i. In this case, the set of valid conditioning variables for the linear

projection of α_i on the regressors differs for different time periods – in period t the conditioning set is the vector $x_i^t \equiv (x_{i1},...,x_{it})$. If however the conditioning set changes over time, the coefficients for the leads and lags of the explanatory variables in the linear projection of α_i will likewise vary over time, thus invalidating ***W3***. Hence, the condition for β to be separately identified from ψ (implying that $\psi_{i1} = \psi_1,...,\psi_{iT} = \psi_T,\ t = 1,...,T$) does not hold.

With pre-determined variables, identification of β requires the assumption that the variables in the main equation are not correlated with the individual effects α_i. Assumption ***W3*** is then substituted by

$$E(\alpha_i | \tilde{z}_i, d_{it} = 1) = q + \phi_t E(v_{it} | \tilde{z}_i, d_{it} = 1). \tag{4.2}$$

For many applications, this assumption is very restrictive.

One way to relax this assumption is to substitute the non-strictly exogenous time-varying correlated regressors by their predictions, and to apply Wooldridge's (1995) estimator. The construction of these predictions is not straightforward, however. For all time periods and for each non-strictly exogenous variable, T unique predictions are required. To identify β, assumption **W3** must hold. Accordingly, predictions for x_i for period t can not be constructed by using the subsample of individuals who participate during that period, where the instruments are both the sample selection term for that period (λ_{it}) and the leads and lags of the explanatory variables in the sample selection equation. This would produce multiple predictions

for the same x_i in different time periods, thus invalidating **W3**. Also, we do not obtain unique predictions for x_i for all periods by including all the sample selection terms in the conditioning set, because the lambda terms are not strictly exogenous in the equation of interest (see discussion above). The way to obtain unique predictions is to predict each component of the vector x_i, using the entire sample of individuals in the participation equation, and all leads and lags of the explanatory variables in that equation as instruments.

The other two estimators rely on difference estimation. Hence pre-determined regressors in the level equation lead to endogenous regressors in the difference equation. In Kyriazidou's (1997) method, a straightforward way to allow for endogenous regressors is an IV type procedure[54]. Let z_i be the set of instrumental variables. Then the difference $(x_{it} - x_{is})$ fitted by z_i is $(\hat{x}_{it} - \hat{x}_{is}) = z_i' \left\{ \sum_j z_j z_j' \right\}^{-1} \sum_j z_j (x_{jt} - x_{js})$, and the IV estimator b_{IV} has the form

$$b_{IV} = \left\{ \sum_i (\hat{x}_{it} - \hat{x}_{is})'(x_{it} - x_{is}) d_{it} d_{is} \varpi_{ts}[(z_{it} - z_{is})\hat{\gamma}] \right\}^{-1} \sum_i (\hat{x}_{it} - \hat{x}_{is})'(w_{it} - w_{is}) d_{it} d_{is} \varpi_{ts}[(z_{it} - z_{is})\hat{\gamma}] \tag{4.3}$$

where $\varpi_{ts}[(z_{it} - z_{is})\hat{\gamma}]$ is the kernel weight for individual i in pair (t,s). This approach allows to maintain the same dimension of $(x_{it} - x_{is})$ in the estimated

[54] The IV version of Kyriazidou's (1997) estimator has been proved to be consistent in Charlier, Melenberg and Van Soest (1997).

instrument set $(\hat{x}_{it} - \hat{x}_{is})$, which is computationally convenient. The pre-estimation of instruments does not affect the second-stage variance.

Given the non-parametric nature of the sample selection terms in this method, identification of the parameters of interest requires some component of z_{it} to be excluded from both the main equation and the instrument set. In practical applications, to find such variables can be hard.

The assumption of strictly exogenous regressors in the main equation for chapter's 3 estimator can be relaxed by applying a generalised method of moments estimator of the form

$$b_{GMM} = \left\{ \sum_i \ddot{x}_{its} \ddot{z}'_{its} \Omega^{-1} \sum_i \ddot{z}_{its} \ddot{x}'_{its} \right\}^{-1} \sum_i \ddot{x}_{its} \ddot{z}'_{its} \Omega^{-1} \sum_i \ddot{z}_{its} (w_{it} - w_{is}), \tag{4.4}$$

where $\ddot{x}_{its} \equiv \left[(x_{it} - x_{is}), \lambda_{its}, \lambda_{ist} \right]'$ and $\ddot{z}_{its} \equiv (z'_i, \lambda_{its}, \lambda_{ist})$. The matrix Ω is given by $\Omega = \sum_i \ddot{z}_{its} \ddot{z}'_{its} r_{its}^2$, where $r_{its} = (w_{it} - w_{is}) - (x_{it} - x_{is}) b^{IV} - \left[\ell_{ts}^{IV} \lambda_{its} + \ell_{st}^{IV} \lambda_{ist} \right]$ are the estimated residuals. The z_i are defined as above, but now the instrument vector for a given pair (t,s), $\ddot{z}_{its}$, also includes the corresponding sample selection terms λ_{its} and λ_{ist}. By setting $\Omega = \sum_i \ddot{z}_{its} \ddot{z}'_{its}$ the GMM estimator becomes a simple IV estimator, and estimates can be used as initial estimates for the GMM estimator.

To summarise, if regressors in the main equation are non-strictly exogenous, the methods of Kyriazidou (1997) and chapter 3 may easily be extended to using IV or

GMM type estimators. For Wooldridge's (1995) estimator, one solution of the problem is to use predicted regressors.

4.4.2 Measurement Error

In typical panel surveys, the construction of work history variables, like tenure and experience, is based on retrospective information, which is likely to suffer from measurement error. An example is labour market experience, which is updated quite precisely during the course of the panel, but where the pre-sample information stems from retrospective data. The measurement error in this case is constant within individuals. If this variable enters the equation of interest in a linear way, differencing eliminates the measurement error. If this variable enters in a non-linear way (for instance, by including squared terms), differencing over time does not eliminate the measurement error, but it eliminates the problem associated to it.

To illustrate this, suppose that the variable x_{it} is measured with error, and we include its level and its square among the regressors in equation (2.1). Let the measured variable x^*_{it} be equal to the true variable x_{it}, plus an individual specific error term:

$$x^*_{it} = x_{it} + e_i, \tag{4.5}$$

where e_t is assumed to be uncorrelated with x_{it}. For Wooldridge's (1995) estimator, writing the true regression equation in (2.5) in terms of the observed variables leads to the following expression:

$$w_{it} = x_{i1}^{*}\psi_1 + \ldots + x_{iT}^{*}\psi_T + x_{i1}^{*2}\Psi_1 + \ldots + x_{iT}^{*2}\Psi_T + x_{it}^{*}\beta_1 + x_{it}^{*2}\beta_2 + \ell_t\lambda\left(H_{it}/\sigma_t\right) + \left[e_{it} - \left(\psi_1 + \ldots + \psi_T + \beta_1\right)e_i + \left(\Psi_1 + \ldots + \Psi_T + \beta_2\right)e_i^2 - 2\left(\Psi_1 x_{i1}^{*} + \ldots + \Psi_T x_{iT}^{*} + \beta_2 x_{it}^{*}\right)e_i\right] \quad (4.6)$$

where the new error term is now given by the expression in brackets.

A common solution to solve the measurement error problem is to use instrumental variable estimation. However, this estimation strategy does not longer lead to consistent estimates in a non-linear error in variables problem, because the error of measurement is no longer additively separable from the regressors (see expression (4.6)). Hence, it is impossible to find instruments which are correlated with the observed regressors, but uncorrelated with the new error term in (4.6).

An alternative solution is to use predicted regressors. In contrast to standard instrumental variables techniques, the use of predicted regressors, once the disturbances of the equation of interest have been purged for correlated heterogeneity and sample selection, allows to estimate the model under some conditions.

Let the true variable x_{it} be determined by a vector of instruments Z_i,

$$x_{it} = Z_i\delta_t + s_{it}. \quad (4.7)$$

Assume that δ_t is known since it is identified from

$$x^*_{it} = Z_i\delta_t + s_{it} + e_i . \tag{4.8}$$

For Wooldridge's (1995) estimator, substitution of (4.7) into equation (2.5) yields the following expression

$$w_{it} = (Z_i\,\delta_1)\psi_1 + \ldots + (Z_i\,\delta_T)\psi_T + (Z_i\,\delta_1)^2\Psi_1 + \ldots + (Z_i\,\delta_T)^2\Psi_T + (Z_i\,\delta_t)\beta_1 + (Z_i\,\delta_t)^2\beta_2 + \ell_t\lambda(H_{it}/\sigma_t) +$$
$$\left[e_{it} + (s_{i1}\psi_1 + \ldots + s_{iT}\psi_T + s_{it}\beta_1) + (s_{i1}^2\Psi_1 + \ldots + s_{iT}^2\Psi_T + s_{it}^2\beta_2) + 2\left((Z_i\,\delta_1)s_{i1}\Psi_1 + \ldots + (Z_i\,\delta_T)s_{iT}\Psi_T + (Z_i\,\delta_t)s_{it}\beta_2\right)\right],$$

where the term in brackets is the new error term. The assumption that s_{it} is independent of Z_i is crucial for consistent estimation, and necessary because of the non-linear specification. Independence guaranties not only that the first conditional moment of s_{it} is equal to zero, but also that the second conditional moment equals zero. Hence, we obtain an expression with linear and quadratic terms in $Z_i\delta_t$, for $t = 1,\ldots,T$, and a new error term that is a function of the original error term, of linear and quadratic terms in s_{it}, and of cross products $s_{it}(Z_i\delta_t)$. To obtain consistent estimates, one needs to assume that $E(\text{new error term}|Z_i\delta_t) = 0$, implying that the Z_i are uncorrelated with the original error term in the equation of interest, and the s_{it} are independent of Z_i.

If estimating the model in differences (as in Kyriazidou (1997) or in chapter's 3 estimator), and writing the true regression equation in (2.6) and (2.7) in terms of the observed variables in (4.5) we obtain:

$$\begin{aligned} w_{it} - w_{is} &= \left(x_{it}^{*} - x_{is}^{*}\right)\beta_1 + \left(x_{it}^{*2} - x_{is}^{*2}\right)\beta_2 + E\left(\varepsilon_{it} - \varepsilon_{is}|\cdot\right) + \left[e_{its} - 2\beta_2\left(x_{it}^{*} - x_{is}^{*}\right)e_i\right] \\ &= \left(x_{it} - x_{is}\right)\beta_1 + \left(x_{it}^{*2} - x_{is}^{*2}\right)\beta_2 + E\left(\varepsilon_{it} - \varepsilon_{is}|\cdot\right) + \left[e_{its} - 2\beta_2\left(x_{it} - x_{is}\right)e_i\right] \end{aligned} \tag{4.9}$$

where $E\left(\varepsilon_{it} - \varepsilon_{is}|\cdot\right)$ is equal to $E\left(\varepsilon_{it} - \varepsilon_{is}|\widetilde{z}_{it}, \widetilde{z}_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1\right)$ for Kyriazidou (1997) and to $E\left(\varepsilon_{it} - \varepsilon_{is}|\widetilde{z}_i, d_{it} = d_{is} = 1\right)$ for chapter's 3 estimator. The new error is given by the term in brackets. Now the measurement error in $\left(x_{it}^{*2} - x_{is}^{*2}\right)$ does not imply a measurement error problem for consistent estimation because e_i is uncorrelated with $\left(x_{it} - x_{is}\right)$.[55] Therefore, differencing eliminates the endogeneity problem due to measurement error, and the IV estimators in section 4.4.1 can be used[56] to address the problem of non-strict exogenous regressors.

[55] Since the error term in (4.9) includes $\left(x_{it} - x_{is}\right)e_i$, and e_i is uncorrelated with $\left(x_{it} - x_{is}\right)$, $E\left[\left(x_{it} - x_{is}\right)e_i\left|\left(x_{it} - x_{is}\right)\right.\right] = E\left[\left(x_{it} - x_{is}\right)e_i\left|\left(x_{it}^{*2} - x_{is}^{*2}\right)\right.\right] = 0$.

[56] Although differencing within individuals does not eliminate the non-linear errors in variables, it does eliminate the problem. The quadratic terms, measured with error, are not longer endogenous to the new error term in the equation. The crucial conditions for this to happen are that the measurement error is time-constant, and uncorrelated with the underlying true variables.

4.5 Empirical Model and Data

4.5.1 Estimation Equation

We apply the estimators discussed in Section 4.2 to analyse wage equations of females, using data from a twelve-year panel. We define the wage equation and the participation equation as:[57]

$$w_{it} = x_{it}\beta + Exp_{it}\xi + Exp_{it}^2\zeta + \alpha_i + \varepsilon_{it}; \quad i = 1,\ldots,N; \quad t = 1,\ldots,T, \tag{5.1}$$

$$d_{it}^* = z_{it}\gamma - \eta_i - u_{it}; \quad d_{it} = 1[d_{it}^* > 0], \tag{5.2}$$

where d_{it} is an indicator variable, being equal to one if the individual participates. The variable d_{it} * is a latent index, measuring the propensity of the individual to participate in the labour market. Our parameter of interest is the effect of actual labour market experience (Exp) on wages. The vector x_{it} is a subset of z_{it} that contains education and time dummies. The vector z_{it} contains in addition age and its square, three variables measuring the number of children in three different age categories, an indicator variable for marital status, an indicator variable for the husband's labour market state, and other household income. We consider the participation equation as a reduced form specification, where labour market experience is reflected by the children indicators, age, and the other regressors. We assume that all regressors in the

[57] See Appendix I for a motivation of this specification.

participation equation are strictly exogenous. The wage variable w_{it} in (5.1) is only observable if $d_{it} = 1$ in (5.2).

Within this model, there are a number of potential sources of bias for the effects of the experience variable. First, unobserved heterogeneity. Unobserved worker characteristics such as motivation and ability or effort may be correlated with actual experience: if high ability workers have a stronger labour market attachment than low ability workers, OLS on equation (5.1) results in upward biased coefficients. Second, sample selection bias. Sample selection occurs if unobservable characteristics affecting the work decision are correlated with the unobservable characteristics affecting the process determining wages. If these unobservable characteristics are correlated with the observables, then failure to control for them will lead to incorrect inference regarding the impact of the observables on wages. Third, experience is likely to be non-strictly exogenous, even after controlling for heterogeneity and sample selection. Labour market experience in any period t is an accumulation of weighted past participation decisions: $Exp_{it} = \sum_{s=1}^{t-1} r_{is} d_{is}$, where r_{is} is the proportion of time individual i allocates in period s to the labour market[58]. In turn, participation depends on wage offers received. Accordingly, any shock to wages in period t affects the level of labour market experience in the future, thus violating condition (4.1). Furthermore, given the above formulation, past shocks to wages affect current experience also by altering the weights r_{is}. A final problem is measurement error.

[58] The process generating experience can be expressed as: $Exp_{it} = Exp_{it-1} + r_{it-1} d_{it-1}$, where by direct substitution we get $Exp_{it} = \sum_{s=1}^{t-1} r_{is} d_{is}$.

As typical in survey data, the experience variable is constructed as the sum of pre-sample retrospective information, and experience accumulated in each year of the survey (see data section for details). Experience updates constructed within the 12 years of the survey should only be marginally affected by miss-measurement, but the pre-sample experience information is likely to suffer quite considerably from measurement error. This results in measurement error in the experience variable, which is constant over time for a given individual.

4.5.2 Data and Sample Retained for Analysis

Our data is drawn from the first 12 waves of the German Socio-Economic Panel (GSOEP) for the years 1984-1995 (see Wagner et al. (1993) for details on the GSOEP). We extract a sample of females between 20 to 64 years old, who have finished their school education, and who have complete data during the sample period on the variables in table 2 (with the exception of wages for females who do not participate in a given period). We exclude individuals who are self-employed in any of the 12 years. We define an individual as participating in the labour market if she reports to have worked for pay in the month preceding the interview. We compute wages by dividing reported gross earnings in the month before the interview by the number of hours worked for pay. We obtain a final sample of 1053 individuals, resulting in 12636 observations. We use both participants and non- participants for the estimation of the selection equation. For estimating the wage equations, we use all females that participate in at least two waves.

Summary statistics and a more detailed description of the variables are given in Table 2. The variable Exp, which reports the total labour market experience of the individual in the year before the interview, is computed in two stages: First, we use information from a biographical scheme, which collects information on various labour market states before entering the panel. This information is provided on a yearly basis, and participation is broken down into part-time and full-time participation. We sum these two labour market states up to generate our total experience variable at entry to the panel. In every succeeding year, this information is updated by using information from a calendar, which lists labour market activities in every month of the year preceding the interview. Again, we sum up part-time and full-time work. Accordingly, after entering the panel, our experience variable is updated on a monthly basis. Furthermore, it relates to the year before the wage information is observed. If wage contracts are re-negotiated at the beginning of each calendar year, this experience information should be the information on which the current contract is based. Participation is defined as being in the state of part-time or full-time employment at the interview time. Non-participation is defined as being in the state of non-employment or unemployment. On average, 54 percent of our sample population participates in the labour market. The average age in the whole sample is 42 years, with individuals in the working sample being slightly younger than in the non-working sample.

We do not restrict our sample to married females. From the 12636 observations, 10680 (84.52 percent) are married females, of whose 51 percent participate in the labour market. We observe a higher percentage of labour market

TABLE 2: DESCRIPTION OF VARIABLES AND SAMPLE STATISTICS (12,636 observations)[a]

Variable	Description	Total Sample	Work=1 (6802 observations)	Work=1 dropping individuals with participation in one year only and observations with missing wages (5861)	Work=0(5834 observations)
Work	dummy variable indicating participation of the female (work=1) or no participation (work=0)	0.538 (0.498)	1 (0)	1 (0)	0 (0)
Lnwage	log gross hourly real wages (1984 West German Marks)	2.681 (0.435)	2.681 (0.435)	2.685 (0.432)	
Exp	years-equivalent worked for money after leaving education	14.373 (9.782)	17.661 (9.407)	17.931 (9.331)	10.541 (8.765)
Exp2	experience squared and divided by 10	30.231 (36.606)	40.040 (38.264)	40.861 (38.122)	18.794 (30.862)
Time	time (year-1900), we also use time dummies for estimation	89.500 (3.452)	89.477 (3.435)	89.457 (3.437)	89.526 (3.472)
Age	age of the female in years	42.263 (9.953)	41.259 (9.356)	41.205 (9.381)	43.434 (10.487)
Age2	age of the female squared and divided by 10	188.527 (84.624)	178.988 (76.917)	178.592 (76.952)	199.650 (91.567)
Ed	education of the female measured as years of schooling	10.847 (1.958)	11.057 (2.129)	11.103 (2.128)	10.602 (1.705)
Hhinc	additional real income per month (in thousands)	2.735 (1.778)	2.439 (1.855)	2.394 (1.897)	3.080 (1.617)
M	dummy variable with value 1 if female married and value 0 if not married	0.845 (0.361)	0.793 (0.404)	0.787 (0.409)	0.905 (0.293)
hwork[b]	dummy variable with value 1 if husband works and value 0 if does not work	0.862 (0.345)	0.877 (0.328)	0.875 (0.331)	0.846 (0.361)
cc1	number of children up to 3 years old in the household	0.117 (0.399)	0.064 (0.301)	0.059 (0.287)	0.179 (0.481)
cc2	number of children between 3 and 6 years old in the household	0.173 (0.442)	0.118 (0.364)	0.110 (0.351)	0.238 (0.511)
cc3	number of children older than 6 years in the household	0.436 (0.739)	0.393 (0.696)	0.366 (0.675)	0.485 (0.784)

[a]Standard errors in parenthesis.
[b]The reported sample statistics for this variable are conditional on the female being married.

participants (72 percent) among the non-married. Of the 1053 females in our sample, 780 are married in each of the 12 periods, 87 are not married in any period, and 186 are married between 1 and 11 years of the sample periods.

Our children variables distinguish between the number of children aged between 1 to 3 years, the number of children aged between 3 and 6 years, and the number of children between 6 and 16 years old. As one should expect, for all three categories, numbers are higher among the non-participants.

To estimate our wage equation conditional on fixed effects, we need repeated wage observations for the same individual. Table 3 reports frequencies of observed wages, as well as the number of state changes between participation and non-participation. 23 percent of our sample individuals participates in none of the 12 years, and about 25 percent in each of the 12 years. More than half of the sample has at least one state change within our observation window and there are no individuals who change state more than 7 times over the 12 years period. In the longitudinal dimension, 767 women (corresponding to 6757 observations) worked for a wage at least in two years during the sample period. Once we drop observations of individuals who do declare participation, but not wages, our number reduces to 5861 observations. The data we use for estimating the wage equation uses all individuals who report wages in at least two periods.

TABLE 3: STATE FREQUENCIES

	Participating Individuals		Number of State Changes		
No. of Years	Frequency	Percent	Changes	Frequency	Percent
0	241	22.89	0	502	47.67
1	45	4.27	1	273	25.93
2	29	2.75	2	131	12.44
3	40	3.80	3	84	7.98
4	53	5.03	4	47	4.46
5	47	4.46	5	10	0.95
6	37	3.51	6	3	0.28
7	49	4.65	7	3	0.28
8	49	4.65			
9	59	5.60			
10	61	5.79			
11	82	7.79			
12	261	24.79			
	1053	100		1053	100

TABLE 4: NUMBER OF OBSERVATIONS WORK=1 VERSUS WORK=0

Years	**Ratios Work=1/0 in participation sample**	**number of Work=1 dropping individuals with participation in one year only and observations with missing wages**
84	565/488	482
85	579/474	500
86	572/481	512
87	561/492	493
88	551/502	479
89	563/490	488
90	576/477	480
91	592/461	496
92	578/475	503
93	576/477	487
94	554/499	482
95	535/518	459
84-95	6802/5834	5861

4.6 Estimation Results

We concentrate most of our discussion on the effect of labour market experience. We use experience and its square as regressors in the wage equation. To facilitate the comparison of results in the various model specifications, we compute the rate of return to work experience

$$\partial w/\partial EXP = \xi + 2\zeta EXP, \tag{6.1}$$

where we evaluate the expression in (6.1) at 14 years of work experience (the sample average).[59] We report estimates in Table 5. The full set of results is given in Table II.1 in the appendix. Rates of return implied by the different methods and for increasing levels of work experience are presented in Table II.2.

Column (1) presents OLS estimates, where we allow for time effects, but no individual effects. The results suggest that, evaluated at 14 years of labour market experience, an additional year increases wages by 1.48 percent. If high ability individuals have a stronger labour market attachment than low ability individuals, then this estimate should be upward biased. Furthermore, sample selection should re-enforce this upward bias if unobservables determining participation are positively correlated with unobservables in the wage equation (either through the α_i or the ε_{it} terms).

[59] Standard errors of this term are easily derived from the variances and covariances of the parameter estimates for ξ and ζ.

TABLE 5: Marginal Experience Effects, WAGE EQUATION[a]

	(1) OLS	(2) FE	(3) DE (OLS)	(4) DE (IV)	(5) DE (GMM)	(6)[b] W (MD)	(7)[c] W (MD) ($Exp\hat{}$)
$\partial w/\partial EXP$ (14 years)	**0.0148*** (0.0007)	**0.0223*** (0.0056)	**0.0200*** (0.0039)	**0.0340*** (0.0054)	**0.0305*** (0.0014)	**0.0148*** (0.0077)	**0.0182*** (0.0038)
Wald Test (Selection)						χ^2_{12} = 17.22 (0.1412)	χ^2_{12} = 17.44 (0.1336)
Wald Test (Fixed Effects)						χ^2_2 = 6.03 (0.049)	χ^2_2 = 5.66 (0.062)
Hausman (Exo-geneity)				χ^2_{14} = 92.84 (0.000)	χ^2_{14} = 55.35 (0.000)		χ^2_{29} = *46.39* *(0.021)*

	(8)[d] K	(9)[d] K (IV)	(10)[c] CH3	(11)[c] CH3 (IV)	(12)[c] CH3 (GMM)
$\partial w/\partial EXP$ (14 years)	**0.0409*** (0.0105)	**0.0116** (0.0637)	**0.0129*** (0.0054)	**0.0122*** (0.0062)	**0.0097*** (0.0017)
Hausman (Selection)	χ^2_2 = *6.6332* (0.036)				
Wald Test (Selection)			χ^2_{132} = *292.60* *(0.000)*	χ^2_{132} = *311.04* *(0.000)*	χ^2_{132} = *3859.11* *(0.000)*
Wald-test (Exo-geneity)				χ^2_{145} = *433.15* *(0.000)*	χ^2_{145} = *1241.19* *(0.000)*

[a] The numbers in parentheses below the coefficient estimates are standard errors. The numbers in parentheses below the test statistics are p-values.

[b] Standard errors **corrected** for the first stage maximum likelihood probit estimates.

[c] Standard errors **corrected** for the first stage maximum likelihood probit estimates and the use of predicted regressors.

[d] Standard errors **corrected** for the prior in the time dummies coefficients.

[e] Standard errors **corrected** for the first stage maximum likelihood bivariate probit estimates.

* Statistically different from zero at the five-percent significance level.

In columns (2) and (3), we present estimators that difference out the fixed effects. Column (2) displays standard fixed-effects (within) estimates (FE), and column (3) difference estimates (DE), where all pair differences within time periods per individual are used.[60] Both estimators allow for individual effects correlated with the explanatory variables. Thus, the upward bias induced by individual fixed effects and any sample selection bias acting through α_i should be eliminated. Interestingly, our estimates increase relative to the simple OLS estimations – point estimates for the fixed effect estimator and the difference estimator are 0.022 and 0.020 respectively.

An explanation for these increases in coefficients is measurement error. As we have shown above, differencing in a quadratic specification eliminates the effect of a time constant measurement error. If the downward bias of the experience coefficient in a level equation, induced by measurement error, is larger than the upward bias due to individual fixed effects, then the coefficient estimates of difference estimators should increase, compared to level estimation.

If past wage shocks affect current experience levels, then experience is not strictly exogenous in the wage level equation. Furthermore, it is endogenous in the difference equation. A common solution to this problem in standard difference estimators is to use instrumental variable techniques. Column (4) and (5) present results when applying IV and GMM techniques to our particular problem. These estimators are obtained by pooled IV and GMM on 66 pairs of combinations of time periods which we can form with a panel of 12 years[61]. As instruments, we use all leads and lags of the variables in the sample selection equation. A Hausman-type test

[60] We estimate pooled OLS on 66 pairs corresponding to 25021 observations.
[61] The IV estimates are used as the first step estimates to obtain the GMM estimates.

comparing the difference IV and GMM estimators with the differenced OLS estimator leads to rejecting exogeneity for the experience variables.

The estimates we obtain for the rate of return to work experience are slightly higher than those obtained with the difference estimators, with point estimates of 0.034 and 0.030 in the IV and GMM estimators respectively. This is consistent with experience being predetermined. If past positive shocks to wages increase the probability of past participation, then the coefficient on the experience variable should be downward biased in a simple difference equation under OLS estimation.

The (IV) difference estimates are consistent under the assumption that selection only works through the fixed effects. If however there is sample selection acting through ε_{it} our instruments are invalid. In this case, the error term will incorporate the extra element:

$$E\left(\varepsilon_{it} - \varepsilon_{is} | d_{it} = d_{is} = 1\right) \neq 0. \tag{6.2}$$

Clearly, a proper instrument set should be uncorrelated with the truncated conditional expectation in (6.2). In most applications, this is unlikely to be the case since the available instruments determine also the selection into the observed regime. For instance, in our case, every variable that affects the participation decision in previous periods should also affect the level of experience in the current period. In this case, $E\left\{\left[\left(\varepsilon_{it} - \varepsilon_{is}\right) + E\left(\varepsilon_{it} - \varepsilon_{is} | d_{it} = d_{is} = 1\right)\right] | z_i \right\} \neq 0$. Accordingly, a time variant selection

process may invalidate instruments in a difference equation, if these instruments are correlated with the selection process.

We now turn to estimation results which take account of a selection process that operates both through ε and α, and we demonstrate how the problems of measurement error and pre-determinedness can be solved within this framework.

4.6.1 Wooldridge's Estimator

Estimation results for Wooldridge's (1995) estimator are presented in columns (6) and (7). We have specified the conditional mean of the individual effects, following Mundlack (1978), as a linear projection on the within individual means of experience and its square. Results in column (6) are based on the assumption that experience is (strictly) exogenous. Results in column (7) allow for endogeneity by using predictions for the experience terms. This procedure takes care of both measurement error, and non-strict exogeneity.

Estimators in columns (6)-(7) are implemented as follows. After obtaining the selection terms by estimating probits for each wave, the wage equation in (2.5) is estimated for each time period. From these estimations, we obtain the unrestricted coefficients for the constant and education, 2 coefficients for the mean of experience and its square, the 2 coefficients of interest for experience and its square, and the coefficient for the selection correction term in a given period. In a second step, we use minimum distance to impose the cross-equation restrictions. To obtain the predictions for the experience variable (results in column (7)), we predict the vector

$\left(Exp_{i1},...,Exp_{i12},\ Exp_{i1}^2,...,Exp_{i12}^2\right)$ using the 1053 individuals in the sample selection equation, as well as all the leads and lags of the explanatory variables in that equation. The components of this vector of predictions are used to obtain the average predicted experience and its average predicted square.

The coefficient estimate for Wooldridge's (1995) estimator is 0.0148 (column 6), which is exactly equal to the OLS result. It is smaller than the fixed effects estimators in columns (2) and (3), which is to be expected if participation is selective and/or there is a measurement error problem in the level equation (which leads to a downward bias). To test for sample selection, we have performed a Wald test on the significance of the selection effects, where *Ho*: $\ell = 0$. This test can be interpreted as a test of selection bias. However, the assumptions under the null hypothesis are stronger than what is required for simple fixed effects estimators, since ***W3*** is maintained under *Ho*[62]. The value for the test statistic is $\chi_{12}^2 = 17.22$, with a p-value of 0.1412. Thus, the null hypothesis can not be rejected. We also performed a Wald test for the joint significance of the ψ coefficients, where *Ho*: $\psi = 0$. The resulting value for the test statistic is larger than the critical value of the χ_2^2, at the five-percent significance level, rejecting the null hypothesis, and indicating the presence of correlated fixed effects.

In column (7) we use predictions for the experience variables. This leads to an increase of the experience coefficient (from 0.014 in column (6) to 0.018 in column (7)). The results indicate that there is endogeneity, induced by non-strict exogeneity of

[62] See Wooldridge (1995) for details on this point.

the experience variable, and/or measurement error. Hausman-type tests, comparing (6) and (7), lead to rejecting exogeneity both after controlling for correlated heterogeneity and sample selection. We perform a Wald tests for the estimates in column (7), testing the null hypotheses that *Ho*: $\ell = 0$ and *Ho*: $\psi = 0$. Again, we cannot reject the null hypothesis *Ho*: $\ell = 0$, but we reject the null hypothesis *Ho*: $\psi = 0$ at a 6.21 percent significance level.

4.6.2 Kyriazidou's Estimator

To implement this estimator, we estimate in a first step a conditional logit fixed effects model (see Chamberlain, 1980). The results are displayed in column (4) of Table III.1 in Appendix III. These first step estimates are then used to calculate weights for the pairs of observations in the difference estimator. To construct the weights we use a normal density function for the kernel. We follow the plug-in procedure described by Horowitz (1992) to obtain the optimal kernel bandwidth.[63] Finally, we perform minimum distance to obtain the parameter estimates. The minimum distance estimator is the weighted average of the estimators for each pair, with weights given by the inverse of the corresponding covariance matrix estimate[64].

[63] With this procedure, some initial value for the bandwidth is chosen. Then the parameter estimates, the estimate of the asymptotic bias and the estimate of the covariance matrix are computed. These estimates are used to compute the mean square error minimising bandwidths. We do a search among initial bandwidths, stopping the process when the chosen initial value of the bandwidth is close enough to the optimal one. As we estimate 66 panel wave pairs $t \neq s$, 66 optimal bandwidth are estimated.

[64] In principle, to estimate the optimal weighting matrix for the minimum distance will require estimates for the covariance matrix of the estimators for the different pairs of time periods. However, Charlier, Melenberg and Van Soest (1997) proof that these covariances converge to zero due to the fact that the bandwidth tends to zero as the sample size increases. As a consequence the optimal weighting matrix simplifies to a block diagonal matrix where each block corresponds to the inverse of the covariance matrix for a given panel wave pair.

As discussed above, the estimator relies on a conditional exchangeability assumption that restricts the error terms to be homoscedastic over time. This assumption seems quite restrictive, in particular when estimating wage equations. There is strong evidence that the variance of the wage distribution has increased considerably over the last two decades. The assumption that the error terms in the selection equation are stationary over time is testable. Table III.1 displays results of the selection equation under the assumption of equal variances over time (column (2)), and estimates that relax this assumption (column (3)). A χ^2 test can be used to test for the joint conditional exchangeability assumption. The increment in the distance statistic[65] is 146.8201 with a p-value of 0.0002, which clearly leads to rejecting the null hypothesis (the test statistic is χ^2_{93} distributed)[66]. Therefore, the joint conditional exchangeability assumption is rejected for our application.

When applying this method to our data, we face a further problem: Asymptotically, the method uses only observations for which the index from the sample selection rule is the same in the two time periods. In our application, there are strong time effects in the selection equation. Furthermore, changes in the variable experience are strongly related to changes in our identifying instruments, like, for instance, the number of children. Any systematic increase in experience between two periods can not be distinguished from the time trend; any non-systematic change coincides with a change of variables in the selection equation. By the nature of the

[65] Testing for additional restrictions in minimum distance estimators can be found in Chamberlain (1984).

[66] The degree of freedom is 104 (the number of parameter estimates in the minimum distance for column (3)) minus 11 (the number of additional restrictions imposed in the minimum distance estimator of column (2)).

estimator, however, the latter pairs of observations obtain a small kernel weight, and they therefore contribute very little to identifying the experience effects. Hence, without further assumptions, we can not identify the experience effects. Similar identification problems are likely to occur in any application where unsystematic changes in the variable of interest coincide with differences in the index function used for constructing the weights. For our particular application, a possible solution to this problem is to use information on aggregate wage growth from other sources. To illustrate the estimator, we use here time effects obtained from simple difference estimators.

Estimation results are displayed in columns (8) and (9). In both specifications, we use time effects obtained from the difference estimator in column (3). Column (8) displays results of simple weighted OLS estimation of equation (6). The IV estimates presented in column (9) are obtained by following the procedure described for Kyriazidou's (1997) method in section 4.4.1 above.

As we already pointed out, given the non-parametric nature of the sample selection terms in this method, identification of the IV estimator requires at least one time-varying variable in the selection equation, which is to be excluded not only from the main equation, but also from the instrument set for experience. Such exclusions are difficult to justify in most circumstances. In our particular case, the experience variable measures the total labour market experience of the individual in the year before the interview. Since it is the weighted sum of past participation decisions, it should be explained by variables that influence past participation, like lags of the husband's income, and lagged children variables. Participation in the current period is

affected by current variables (like children(t), hhinc(t), etc.). This should identify the model. We need one exclusion restriction, and we exclude current other household income hhinc(t) from the instrument set for experience.

The estimator in (8) does not correct for possible endogeneity of the experience variable. The coefficient for the experience effect indicates that a year of labour market experience increases wages by 4.1 percent. This estimate is very large. The estimator in (9) corrects for non-strict exogeneity of the experience variable in the level equation. Instrumenting reduces the experience effect to 1.2 percent, but the effect is not statistically significant (which may be due to the smaller effective sample size used for this estimator). Because of the problems discussed above, we do not wish to overemphasise these estimates. Also, the estimates are obviously sensitive to the choice of pre-estimated time effects.

A Hausman-type test comparing the parameter estimates in column (8) with the difference estimator in column (3) indicates that the null hypothesis of no selectivity bias is rejected.

4.6.3 Chapter's 3 Estimator

Columns (10)–(12) present estimates, using the method in chapter 3. Column (10) displays results of simple OLS estimation of equation (2.7). IV-GMM estimates are presented in columns (11) and (12). For estimation, we use each combination of panel waves (t,s), resulting in a total of 66 pairs. To combine these estimates, we use minimum distance.[67] We obtain coefficients for 11 time dummies, the coefficients on

[67] The optimal weighting matrix is obtained from an estimate for the covariance matrix of the

experience and its square, and estimates of 66*2=132 coefficients for the correction terms for all the pairs.[68] The standard errors we present in table 5 are corrected for the first step bivariate probit estimates. The variables used as instruments are the leads and lags of the variables included in the sample selection equation, and the corresponding two sample selection terms of each pair of time periods.

The first step estimator of the parameters $\gamma_t, \gamma_s, \rho_{ts}$, which are used for constructing the correction terms, are obtained by estimating 66 bivariate probits. The parameters we estimate in each bivariate probit are the reduced form parameters of the corresponding indices of the selection rules for the two time periods. We also get an estimate of the correlation coefficient between the errors in the two time periods. The mean value for ρ_{ts} is 0.7862 (se=0.1299) with a minimum at 0.4845 and a maximum at 0.9658. Consequently, we reject on average $Ho: \rho_{ts} = 0$. Correlation appears because of the c_i component in the error term and/or because of serially correlated idiosyncratic errors.

We test whether the 66*2 correction terms are jointly significant, using Wald tests. The resulting values for the test statistics for the estimators in Columns (10) to (12) are clearly larger than the critical values of the χ^2_{132} at any conventional significance level. Hausman-type tests comparing the IV and the GMM estimators with the OLS estimator in Column (10) lead to rejecting exogeneity both after controlling for correlated heterogeneity and sample selection.

---

estimators for the different time periods.

[68] The estimates can be obtained upon request.

The estimated parameters are slightly lower than the OLS estimates, and do not differ very much between specifications. They indicate that, evaluated at 14 years of labour market experience, an additional year increases wages by about 1 percentage point. Compared to Wooldridge's (1995) estimator, estimates are slightly smaller, which may be due to different parametric assumptions imposed by the two estimators. Furthermore, estimates are remarkably similar across specifications. One reason for this similarity is that with chapter's 3 estimator, instrumenting corrects only for the non-strict exogeneity problem. With Wooldridge's (1995) estimator, the use of predicted regressors corrects also for the measurement error bias.

Interesting is also a comparison of wage growth due to aggregate time effects. In the last row of Table II.1, we display average wage growth for the 12 years period due to common time effects. The numbers indicate that the different methods result in different numbers. For instance, chapter's 3 estimator in column (10) assigns about 8 percent more wage growth over the 12 years period to time effects than the simple OLS estimator (column 3). An explanation for these differences is that chapter's 3 method controls for time-varying sample selection (as does Wooldridge's estimator). As pointed out by Moffitt (1984), wages may trend not only because of aggregate wage growth (proxied by the time dummies), but also because of changes in the sample selection over time. If sample selection decreases over time, and if we do not control for selection, the time dummies will pick up this trend, leading to decreasing time effects in standard fixed effects and difference estimators, like the ones displayed in columns (2) to (5). This leads to downward biased time dummies. With chapter's 3

estimator, the time dummies will presumably pick up just the secular productivity growth, since it controls for the decline in sample selection over time.

With this method, the sample selection term is given by a parameterisation of the conditional mean $E(\varepsilon_{it} - \varepsilon_{is} | \tilde{z}_i, d_{it} = d_{is} = 1)$, $s < t$. We obtain for most individuals negative predictions for these expectations. To investigate whether sample selection does indeed decrease over time, we write the estimated values of these conditional means as a function of 11 time dummies in differences (after controlling for the increments in experience and its square). Using minimum distance estimation, we obtain negative and significant coefficients for the time dummies, which increase in absolute value over time. This indicates that sample selection does in fact decline over time.[69]

4.7 Conclusions

In many empirical applications, the equation of interest is defined for a non-random sample of the overall population. Furthermore, at the same time the outcome equation contains an unobserved individual specific component which is correlated with the model regressors. In this chapter we discuss three estimators which may be applied if both problems occur simultaneously: The estimators of Wooldridge (1995), Kyriazidou (1997), and the one in chapter 3. We investigate and compare the

[69] This result is in line with the estimates obtained for the participation equation in Appendix III. Here, the estimates for the time dummies show that female labour force participation increases over the length of the panel. Hence, as participation probabilities increase, sample selection may be reduced.

conditions under which they produce consistent estimates. We show how these estimators can be extended to take account of non-strict exogeneity and/or time constant non-linear errors in variables. We illustrate that, if regressors in the main equation suffer from these problems, the methods of Kyriazidou (1997) and chapter's 3 can be straightforwardly extended to using IV or GMM type estimators. For Wooldridge's (1995) estimator, one solution of the problem is to use predicted regressors.

Not many applications exist for sample selection estimators in panel data models. To learn about the performance of the methods in a practical application, we apply the estimators and their extensions to a typical problem in labour economics: The estimation of wage equations for female workers. The parameter we seek to identify is the effect of actual labour market experience on wages. The problems that arise in this application are non-random selection, and unobserved individual specific heterogeneity which is correlated with the regressors. In addition, actual experience is predetermined, and the experience measure is likely to suffer from measurement error.

A flexible and attractive estimator is that by Kyriazidou (1997). It turns out however that, for our particular application, this estimator is difficult to apply. The estimator is very flexible in that it avoids specifying the sample selection terms, and it requires no parametric assumptions about the unobservables in the model. But it imposes a conditional exchangeability assumption, which is rejected by the data in our particular application. Furthermore, in the case where any non-systematic variation in the variable of interest (experience in our case) coincides with changes in the selection index, this estimator runs into identification problems (between time effects and

experience in our case), that can only be solved by using additional information. We use pre-estimated time dummies from simple difference estimators. To implement the IV estimator (which is producing consistent estimates if experience is pre-determined and/or contemporaneously endogenous), we need a further identification assumption. The estimate we obtain for the effect of labour market experience for the simple Kyriazidou estimator is quite large: Evaluated at 14 years of labour market experience, an additional year increases wages by about 4 percentage points. The estimates are sensitive to the pre-estimated time effects. The IV estimates are smaller, but not precisely estimated.

The results we obtain using Wooldridge's and chapter's 3 estimators indicate that there are correlated fixed effects, and non-random sample selection. With Wooldridge's (1995) estimator, the null hypothesis of no correlated fixed effects is rejected for all specifications. Conditional on fixed effects, the null hypothesis of no sample selection can not be rejected with Wooldridge's (1995) estimator, but it is clearly rejected with chapter's 3 estimator.[70] Using Wooldgridge's (1995) estimator, we reject specifications, which do not allow for predetermined regressors (and contemporaneous endogeneity). Chapter's 3 method rejects strict exogeneity of the experience variable, conditional on taking care of the measurement error problem by first differencing. Accordingly, the use of sample selection models which take care of correlated fixed effects seems to be justified. Furthermore, the extensions we suggest in this chapter seem to be important for our particular application.

[70] For Wooldridge's estimator, however, the assumptions under the null hypothesis are stronger than what is required for simple fixed effects.

The most general estimator using Wooldridge's (1995) method implies an increase in wages by 1.8 percent for one year of labour market experience, evaluated at 14 years of experience. According to this estimator, the return to experience decreases from 3.1 percent for the first year to 2.2 percent after 10 years to 1.2 percent after 20 years (see Table II.2). Estimates of chapter's 3 most general estimator (the GMM) are slightly lower. They range from 2.2 percent after the first year to 1.4 percent after 10 years to 0.4 percent after 20 years. Simple OLS estimates are intermediate. They range from 3.0 percent after 1 year to 1.9 percent after 10 years to 0.8 percent after 20 years of labour market experience.

Our results also indicate that estimates of aggregate wage growth are sensitive to the trend in sample selection. If sample selection decreases over time, simple difference estimators lead to downward biased time effects. In our case, wage growth over the 12 years period due to the aggregate time trend is 14 percent for Wooldridge's most general estimator, and 16 percent for chapter's 3 most general estimator. In contrast, a simple difference estimator assigns only 9 percent of wage growth to aggregate time effect over the 12 years period.

4.8 Appendix I: Econometric Model of Wages

Our econometric model of wages may be motivated as follows. Consider a model where human capital is accumulated in a learning by doing way. The accumulation equation for human capital (measured in monetary units) is then given by:

$$w_{it}^{*} = w_{it-1}^{*} + (r_{it-1}d_{it-1})\xi + \left[(r_{it-1}d_{it-1})^{2} + \left(\sum_{s=1}^{t-1}\sum_{\tau\neq s} r_{is}d_{is}r_{i\tau}d_{i\tau} - \sum_{s=1}^{t-2}\sum_{\tau\neq s} r_{is}d_{is}r_{i\tau}d_{i\tau}\right)\right]\zeta. \quad \text{(I.1)}$$

Here d_{it} is the participation-status variable and r_{is} is the proportion of time individual i allocates in period s to the labour market. Thus, $(r_{it-1}d_{it-1})$ is equivalent to the increase in human capital in a given period. Human capital depreciates while working, which is reflected by the term in brackets. There is no depreciation in periods out of work. The actual market wage is given by $w_{it} = w_{it}^{*} + \overline{\alpha}_{i} + \varepsilon_{it}$, where $\overline{\alpha}_{i}$ is an individual effect and ε_{it} is some idiosyncratic shock. By recursion, we obtain the following wage equation:

$$w_{it} = w_{i1}^{*} + \left(\sum_{s=1}^{t-1} r_{is}d_{is}\right)\xi + \left(\sum_{s=1}^{t-1} r_{is}d_{is}\right)^{2}\zeta + \overline{\alpha}_{i} + \varepsilon_{it}, \quad \text{(I.2)}$$

where the wage in period t depends on the initial wage, w_{i1}^{*}, and cumulative work experience and its square.

We assume that the entry wage, w_{i1}^*, is solely determined by the individual's unobserved ability, and the level of schooling:

$$w_{i1}^* = S\beta_S + \alpha_i^*, \tag{I.3}$$

where S is a measure for years of education, and α_i^* is an error term specific to the individual (e.g. "ability"). Combining (I.2) and (I.3) gives:

$$w_{it} = S\beta_S + \left(\sum_{s=1}^{t-1} r_{is} d_{is}\right)\xi + \left(\sum_{s=1}^{t-1} r_{is} d_{is}\right)^2 \zeta + \alpha_i + \varepsilon_{it}, \tag{I.4}$$

where $\alpha_i \equiv \left(\overline{\alpha}_i + \alpha_i^*\right)$. The specification in (5.1) is obtained by using $\sum_{s=1}^{t-1} r_{is} d_{is} = Exp_{it}$ and by adding to (I.4) time dummies, which reflect aggregate wage growth.

4.9 Appendix II: Tables

TABLE II.1: ESTIMATES FOR THE WAGE EQUATION[a]

Variable	(1) OLS	(2) FE	(3) DE (OLS)	(4) DE (IV)	(5) DE (GMM)	(6)[b] W (MD)	(7)[c] W (MD) ($Ex\hat{p}$)	(8)[d] K	(9)[d] K (IV)	(10)[e] CH3	(11)[e] CH3 (IV)	(12)[e] CH3 (GMM)
CST	0.9990*					1.1111*	1.0162*					
	(0.0310)					(0.0724)	(0.0804)					
D85	0.0047	0.0056	0.0052	-0.0062	-0.0041	0.0247	0.0482*	0.0052	0.0052	0.0275	0.0249	-0.0088
	(0.0204)	(0.0139)	(0.0068)	(0.0074)	(0.0029)	(0.0250)	(0.0236)	(0.0068)	(0.0068)	(0.0203)	(0.0200)	(0.0087)
D86	0.0450*	0.0308	0.0291*	0.0054	0.0168*	0.0466	0.0556*	0.0291*	0.0291*	0.0711*	0.0536*	0.0327*
	(0.0206)	(0.0163)	(0.0095)	(0.0115)	(0.0038)	(0.0295)	(0.0264)	(0.0095)	(0.0095)	(0.0223)	(0.0226)	(0.0080)
D87	0.0773*	0.0588*	0.0597*	0.0238	0.0332*	0.0876*	0.0920*	0.0597*	0.0597*	0.1001*	0.0960*	0.0835*
	(0.0205)	(0.0199)	(0.0125)	(0.0158)	(0.0046)	(0.0332)	(0.0257)	(0.0125)	(0.0125)	(0.0252)	(0.0267)	(0.0108)
D88	0.0826*	0.0492*	0.0602*	0.0131	0.0317*	0.1048*	0.1212*	0.0602*	0.0602*	0.1296*	0.1253*	0.0916*
	(0.0213)	(0.0240)	(0.0159)	(0.0205)	(0.0056)	(0.0409)	(0.0350)	(0.0159)	(0.0159)	(0.0303)	(0.0326)	(0.0121)
D89	0.1051*	0.0614*	0.0715*	0.0131	0.0341*	0.1128*	0.1142*	0.0715*	0.0715*	0.1635*	0.1437*	0.1165*
	(0.0205)	(0.0284)	(0.0194)	(0.0254)	(0.0066)	(0.0468)	(0.0347)	(0.0194)	(0.0194)	(0.0358)	(0.0398)	(0.0137)
D90	0.1399*	0.0941*	0.1048*	0.0355	0.0568*	0.1394*	0.1466*	0.1048*	0.1048*	0.2043*	0.1863*	0.1679*
	(0.0209)	(0.0330)	(0.0230)	(0.0302)	(0.0077)	(0.0543)	(0.0402)	(0.0230)	(0.0230)	(0.0393)	(0.0447)	(0.0149)
D91	0.1453*	0.1142*	0.1254*	0.0454	0.0622*	0.1452*	0.1421*	0.1254*	0.1254*	0.2126*	0.1872*	0.1843*
	(0.0213)	(0.0378)	(0.0268)	(0.0352)	(0.0089)	(0.0582)	(0.0378)	(0.0268)	(0.0268)	(0.0449)	(0.0505)	(0.0164)
D92	0.1684*	0.1274*	0.1434*	0.0523	0.0766*	0.1909*	0.1605*	0.1434*	0.1434*	0.2342*	0.2227*	0.1987*
	(0.0213)	(0.0426)	(0.0304)	(0.0403)	(0.0104)	(0.0660)	(0.0403)	(0.0304)	(0.0304)	(0.0508)	(0.0568)	(0.0174)
D93	0.1683*	0.1258*	0.1439*	0.0422	0.0706*	0.1919*	0.1991*	0.1439*	0.1439*	0.2495*	0.2308*	0.2688*
	(0.0221)	(0.0475)	(0.0342)	(0.0453)	(0.0109)	(0.0719)	(0.0430)	(0.0342)	(0.0342)	(0.0556)	(0.0626)	(0.0195)
D94	0.1724*	0.1276*	0.1461*	0.0335	0.0628*	0.2227*	0.2073*	0.1461*	0.1461*	0.2474*	0.2361*	0.2769*
	(0.0215)	(0.0525)	(0.0380)	(0.0502)	(0.0112)	(0.0790)	(0.0455)	(0.0380)	(0.0380)	(0.0602)	(0.0670)	(0.0218)
D95	0.2159*	0.1398*	0.1594*	0.0367	0.0668*	0.2591*	0.2519*	0.1594*	0.1594*	0.2736*	0.2795*	0.3556*
	(0.0225)	(0.0572)	(0.0415)	(0.0549)	(0.0131)	(0.0856)	(0.0512)	(0.0415)	(0.0415)	(0.0659)	(0.0748)	(0.0238)
ED	0.1133*					0.1065*	0.1086*					
	(0.0020)					(0.0042)	(0.0043)					
EXP	**0.0309***	**0.0349***	**0.0324***	**0.0522***	**0.0473***	**0.0230***	**0.0320***	**0.0525***	**0.0157**	**0.0244***	**0.0248***	**0.0229***
	(0.0019)	(0.0062)	(0.0042)	(0.0058)	(0.0017)	(0.0090)	(0.0060)	(0.0222)	(0.1935)	(0.0060)	(0.0071)	(0.0021)
EXP2	**-0.0058***	**-0.0045***	**-0.0044***	**-0.0065***	**-0.0060***	**-0.0029***	**-0.0049***	**-0.0041**	**-0.0014**	**-0.0041***	**-0.0045***	**-0.0047***
	(0.0005)	(0.0005)	(0.0002)	(0.0003)	(0.0001)	(0.0009)	(0.0012)	(0.0050)	(0.0470)	(0.0005)	(0.0006)	(0.0002)
$\partial w/\partial EXP$	**0.0148***	**0.0223***	**0.0200***	**0.0340***	**0.0305***	**0.0148***	**0.0182***	**0.0409***	**0.0116**	**0.0129***	**0.0122***	**0.0097***
(14 years)	(0.0007)	(0.0056)	(0.0039)	(0.0054)	(0.0014)	(0.0077)	(0.0038)	(0.0105)	(0.0637)	(0.0054)	(0.0062)	(0.0017)
Av. ret. T.	0.1204*	0.0850*	0.0953*	0.0268	0.0461*	0.1387*	0.1399*	0.0953*	0.0953*	0.1739*	0.1624*	0.1607*
dummies	(0.0150)	(0.0336)	(0.0228)	(0.0301)	(0.0075)	(0.0492)	(0.0306)	(0.0228)	(0.0228)	(0.0380)	(0.0428)	(0.0137)

[a] The numbers in parentheses are standard errors.

[b] Standard errors **corrected** for the first stage maximum likelihood probit estimates.

[c] Standard errors **corrected** for the first stage maximum likelihood probit estimates and the use of predicted regressors.

[d] Standard errors **corrected** for the prior in the time dummies coefficients.

[e] Standard errors **corrected** for the first stage maximum likelihood bivariate probit estimates.

*Statistically different from zero at the five-percent significance level.

TABLE II.2: ESTIMATED RATES OF RETURN FOR WORK EXPERIENCE ($\partial w/\partial EXP$)[a]

Years of work experience	(1) OLS	(2) FE	(3) DE (OLS)	(4) DE (IV)	(5) DE (GMM)	(6)[b] W (MD)	(7)[c] W (MD) ($Ex\hat{p}$)	(8)[d] K	(9)[d] K (IV)	(10)[e] CH3	(11)[e] CH3 (IV)	(12)[e] CH3 (GMM)
1	**0.0298***	**0.0340***	**0.0315***	**0.0509***	**0.0461***	**0.0224***	**0.0310***	**0.0516***	**0.0154**	**0.0236***	**0.0239***	**0.0220***
	(0.0018)	(0.0061)	(0.0041)	(0.0057)	(0.0017)	(0.0089)	(0.0058)	(0.0213)	(0.1842)	(0.0059)	(0.0070)	(0.0021)
5	**0.0252***	**0.0304***	**0.0280***	**0.0457***	**0.0413***	**0.0201***	**0.0271***	**0.0483***	**0.0143**	**0.0203***	**0.0203***	**0.0182***
	(0.0015)	(0.0059)	(0.0041)	(0.0056)	(0.0016)	(0.0085)	(0.0051)	(0.0177)	(0.1468)	(0.0057)	(0.0067)	(0.0019)
10	**0.0194***	**0.0259***	**0.0236***	**0.0392***	**0.0353***	**0.0172***	**0.0222***	**0.0442***	**0.0128**	**0.0162***	**0.0158***	**0.0135***
	(0.0010)	(0.0057)	(0.0040)	(0.0055)	(0.0015)	(0.0080)	(0.0043)	(0.0134)	(0.1003)	(0.0055)	(0.0064)	(0.0018)
15	**0.0137***	**0.0214***	**0.0192***	**0.0327***	**0.0293***	**0.0143***	**0.0172***	**0.0400***	**0.0113**	**0.0121***	**0.0113**	**0.0088***
	(0.0006)	(0.0056)	(0.0039)	(0.0054)	(0.0014)	(0.0076)	(0.0038)	(0.0099)	(0.0547)	(0.0053)	(0.0062)	(0.0017)
20	**0.0079***	**0.0170***	**0.0148***	**0.0262***	**0.0233***	**0.0114**	**0.0123***	**0.0359***	**0.0099**	**0.0080**	**0.0068**	**0.0041***
	(0.0004)	(0.0055)	(0.0039)	(0.0053)	(0.0013)	(0.0074)	(0.0035)	(0.0080)	(0.0184)	(0.0052)	(0.0060)	(0.0016)

[a] The numbers in parentheses are standard errors.

[b] Standard errors **corrected** for the first stage maximum likelihood probit estimates.

[c] Standard errors **corrected** for the first stage maximum likelihood probit estimates and the use of predicted regressors.

[d] Standard errors **corrected** for the prior in the time dummies coefficients.

[e] Standard errors **corrected** for the first stage maximum likelihood bivariate probit estimates.

*Statistically different from zero at the five-percent significance level.

4.10 Appendix III: The Participation Equation

Results for the participation equation for a selection of estimators are given in Table III.1. The first model is a pooled probit, not taking account of a possible correlation between the explanatory variables and the individual effects. Columns (2) and (3) report results from a specification where individual effects are written as a linear projection on leads and lags of time-varying regressors (see Chamberlain (1984)).[71] The estimation procedure consists of two steps. In the first step cross-equation restrictions are ignored, and the γ_t are estimated by probit for each time period separately. The second step is a minimum distance step. The results in column (2) impose the restriction that $\sigma_t = \sigma$ for $t = 84,...,95$. In column (3), σ_{84} has been normalised to 1, and the remaining variances are estimated.

Finally, in column (4) we present results from a fixed effect logit model, as proposed by Chamberlain (1980). This is the estimator used for the weights in Kyriazidou's (1997) method. Since the scaling is different, only the sign (and the ratios) of the coefficients can be compared with the other 3 models.

The estimates for the time dummies show that female labour force participation increases over the length of the panel. Participation probabilities increase

---

[71] The individual effect is written as $\eta_i = z_{i1}\delta_1 + ... + z_{iT}\delta_T + c_i$, with $c_i \sim N(0, \sigma_c^2)$ and independent of z_i. The $u_i = (u_{i1},...,u_{iT})'$ are assumed to be i.i.d. $N(0,\Sigma)$. Define $\sigma_t = (\tilde{\sigma}_t^2 + \sigma_c^2)^{1/2}$, were $\tilde{\sigma}_t^2$ is the t^{th} diagonal element of Σ. Then $P[d_{it} = 1|z_i] = \Phi\left[\frac{z_{it}\gamma - (z_{i1}\delta_1 + ... + z_{iT}\delta_T)}{\sigma_t}\right] = \Phi[z_{i1}\gamma_{t1} + ... + z_{iT}\gamma_{tT}]$ where $\gamma_t = \sigma_t^{-1}(\delta_1',...,\delta_{t-1}', \gamma' - \delta_t', \delta_{t+1}',...,\delta_T')'$.

TABLE III.1: SOME ESTIMATES FOR THE PARTICIPATION EQUATION[a]

Variables	(1) Pooled probit	(2) Chamberlain(1984) $\sigma = 1$	(3) Chamberlain (1984, $\sigma_{84} = 1$)	(4) Conditional logit (1980)
CST	-2.1644*	-1.9921*	-1.4615*	
	(0.2309)	(0.2662)	(0.2686)	
D85	-0.1394*	-0.0570	-0.1396*	0.1897
	(0.0575)	(0.0599)	(0.0666)	(0.1548)
D86	0.0633	0.0652	0.0558	1.1056*
	(0.0579)	(0.0592)	(0.0638)	(0.1722)
D87	0.0623	0.0573	0.0001	1.5592*
	(0.0580)	(0.0601)	(0.0560)	(0.1962)
D88	0.0679	0.0770	-0.0220	2.0835*
	(0.0580)	(0.0603)	(0.0538)	(0.2265)
D89	0.1329*	0.1779*	0.0878	2.7388*
	(0.0581)	(0.0605)	(0.0586)	(0.2614)
D90	0.2056*	0.2014*	0.1073*	3.5187*
	(0.0584)	(0.0609)	(0.0595)	(0.3010)
D91	0.6617*	0.6045*	0.9144*	5.4085*
	(0.0623)	(0.0644)	(0.1262)	(0.3570)
D92	0.2786*	0.2454*	0.0988*	4.7497*
	(0.0589)	(0.0621)	(0.0541)	(0.3867)
D93	0.3138*	0.3269*	0.1008*	5.3176*
	(0.0593)	(0.0634)	(0.0537)	(0.4318)
D94	0.2920*	0.2786*	0.1048*	5.6722*
	(0.0595)	(0.0644)	(0.0570)	(0.4770)
D95	0.2649*	0.2492*	0.0889	6.0183*
	(0.0599)	(0.0651)	(0.0581)	(0.5228)
AGE	0.1443*	0.1432*	0.1097*	
	(0.0111)	(0.0124)	(0.0146)	
AGE2	-0.0021*	-0.0022*	-0.0017*	-0.0069*
	(0.0001)	(0.0001)	(0.0002)	(0.0006)
ED	0.0806*	0.0902*	0.0878*	
	(0.0066)	(0.0071)	(0.0090)	
CC1	-0.7635*	-0.5880*	-1.1583*	-1.9587*
	(0.0368)	(0.0419)	(0.0941)	(0.1079)
CC2	-0.5757*	-0.4361*	-0.5092*	-1.3773*
	(0.0298)	(0.0369)	(0.0501)	(0.0907)
CC3	-0.2265*	-0.1027*	-0.2053*	-0.3807*
	(0.0174)	(0.0260)	(0.0302)	(0.0717)
HWORK	0.1032*	0.0094	-0.0281	0.2923*
	(0.0372)	(0.0510)	(0.0439)	(0.1355)
HHINC	-0.1383*	-0.0430*	-0.0506*	-0.3334*
	(0.0070)	(0.0085)	(0.0092)	(0.0375)
M	-0.3171*	-0.5324*	-0.2983*	-1.5269*
	(0.0433)	(0.0766)	(0.0680)	(0.1980)
$sigma_{85}$			1.1650*	
			(0.1425)	
$sigma_{86}$			1.1404*	
			(0.1140)	
$sigma_{87}$			0.8926*	
			(0.0799)	
$sigma_{88}$			0.7948*	
			(0.0699)	
$sigma_{89}$			0.9103*	
			(0.0860)	
$sigma_{90}$			0.8675*	
			(0.0871)	
$sigma_{91}$			1.9506*	
			(0.2130)	
$sigma_{92}$			0.6530*	
			(0.0583)	
$sigma_{93}$			0.6313*	
			(0.0551)	
$sigma_{94}$			0.7508*	
			(0.0677)	
$sigma_{95}$			0.7940*	
			(0.0743)	

[a] The numbers in parentheses are standard errors.
* Statistically different from zero at the five-percent significance level.

until the age of 30-35 (depending on the specification), and decrease thereafter. An increase in other family income (hhinc) has a negative effect on the participation probability, indicating that leisure is a normal good. The dummy for the husband working has a positive effect on the participation probability, but is insignificant in two out of the four specifications. The effect of education is positive, indicating that educational achievements increase participation. The number of children in different age groups has a negative effect, where the effect decreases with the age group of the children.

The specification in column 1 does not control for correlated individual specific effects, while specifications in the other columns do. When we compare the first two columns, we observe that the effect of the children variables, and other household income decreases quite substantially. This is consistent with the notion that unobserved ability components which increase the woman's competitiveness in the labour market (and therefore her participation propensity) are negatively correlated with the number of children. They also seem to be negatively correlated with other household income.

The results in column (3) allow for different variances over time. The coefficient of the constant term is similar in columns (1) and (2) but much smaller (in absolute value) in column (3). To test for the 11 additional restrictions imposed on column (2), relative to column (3), we perform a χ^2 test (see Chamberlain (1984) to test for additional restrictions in minimum distance estimators). The increment in the distance statistic is 146.8201 with a p-value = 0.0002, which clearly leads to rejecting

the null hypothesis (the test statistic is χ^2_{93} distributed)[72]. We conclude that there are different variances over time for the error term in the selection equation.

[72] The degree of freedom is 104 (the number of parameter estimates in the minimum distance for column (3)) minus 11 (the number of additional restrictions imposed in the minimum distance estimator of column (2)).

Chapter 5

New Semiparametric Pairwise Difference Estimators for Panel Data Sample Selection Models*

5.1 Introduction

In a panel data sample selection model, where both the selection and the regression equation may contain individual effects allowed to be correlated with the observable variables, Wooldridge (1995) proposed a method for correcting for selection bias. Kyriazidou (1997) proposes an estimator imposing weaker distributional assumptions. A more parametric approach, getting ride of some assumptions in the previous methods, has been developed in chapter 3.

The method by Wooldridge (1995), based on estimation of a model in levels, requires a linear projection for the individual effects in the equation of interest on the leads and lags of the explanatory variables. The other two methods overcome this problem by estimation of a model in differences over time for a given individual. Time differencing for the same individual will eliminate the individual effects from the regression equation. The work of Kyriazidou (1997) is the less parametric of the

* I am grateful to Bo Honoré, Myoung-jae Lee and Frank Windmeijer for useful comments and suggestions. Thanks are also owed to participants at the Econometric Society European Meeting (ESEM), August/September 1999, Santiago de Compostela, Spain.

three methods, in the sense that the distribution of all unobservables is left unspecified, and it allows for an arbitrary correlation between individual effects and regressors. The price paid is in terms of another assumption, that is, the called *conditional exchangeability* assumption for the errors in the model. This assumption allows for individual heteroskedasticity of unknown form but it imposes homoskedasticity over time. The advantage of the estimator proposed in chapter 3 is that it allows for the variance of the errors to vary over time. It is then relaxed the assumption that the errors for a given individual are homoskedastic. For this we pay the price of assuming a trivariate normal distribution for the errors in the model.

According to the results of the Monte Carlo investigation of the finite-sample properties of Wooldridge (1995) and Kyriazidou's (1997) estimators (chapter 2) we can conclude that important factors of bias or lack in precision in the estimates come from misspecification problems related to the individual effects in the main equation and violations of the *conditional exchangeability* assumption. The estimator in chapter 3 gets ride of both factors as it can be seen in the Monte Carlo experiments presented in that chapter. However, the need to assume a trivariate normal distribution for the errors may question the robustness of the estimator against misspecification of the error distribution. The work in this chapter has been developed with the aim of keeping the properties of the estimator in chapter 3 but allowing for a free joint trivariate distribution.

In this chapter, estimation of the coefficients in a "double-index" selectivity bias model is considered under the assumption that the selection correction function depends only on the conditional means of some observable selection variables. We

will present two alternative methods. The first one follows the familiar two-step approach proposed by Heckman (1976,1979) for selection models. The procedure will first estimate consistently and nonparametrically the conditional means of the selection variables. In the second step we will not only take pair differences for the same individual over time (to eliminate the individual effects as in Kyriazidou (1997) and chapter 3) but also after this we will take pairwise differences across individuals to eliminate the sample selection correction term (the idea of pairwise differencing across individuals in a cross section setting appears in Powell (1987) and Ahn and Powell (1993)). On the resulting model after this double differencing we will apply a weighted least squares regression with decreasing weights to pairs of individuals with larger differences in their "double index" variables, and then larger differences in the selection correction terms. The alternative method will need just pairwise differences over time for the same individual but it will include three steps. The first one will be identical to the corresponding one in the other method, that is, nonparametrically we will estimate the conditional means of the selection variables. In the second step we will estimate by nonparametric regression the conditional means of pairwise differences in explanatory variables and pairwise differences in dependent variables on the selection variables (the "double index") estimated in the first step. The third step will use these nonparametric regression estimators to write a model in the spirit of the semiparametric regression model of Robinson (1988), which will be estimated by OLS.

The chapter is organised as follows. Section 2 describes the model, discusses some related identification issues, and revises assumptions on the sample selection

correction terms in the available difference estimators for panel data sample selection models. Section 3 presents the new estimators. In Section 4 we show the link between them. Section 5 reports results of a small Monte Carlo simulation study of its finite sample performance. Section 6 gives concluding remarks, and the Appendices provide formulae for the asymptotic variance-covariance matrices.

5.2 The Model and the Available Estimators

5.2.1 The Model

Our case of study is a panel data sample selection model. In this model we are interested in the estimation of the regression coefficients β in the equation

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

$$d_{it}^* = f_t(z_i) - c_i - u_{it}; \quad d_{it} = 1[d_{it}^* \geq 0], \tag{2.2}$$

where $z_i = (z_{i1},...,z_{iT})$. x_{it} and z_i are vectors of explanatory variables (which may have components in common), ε_{it} and u_{it} are unobserved disturbances, α_i are individual-specific effects allowed to be correlated to the explanatory variables x_i,

and c_i are individual-specific effects uncorrelated to z_i. Whether or not observations for y_{it} are available is denoted by the dummy variable d_{it}.

In (2.2) there is no need to impose any parametric assumption about the shape of the selection indicator index $f_t(z_i)$. In fact, by assuming that depends on all the leads and lags of an F-dimensional vector of conditioning variables z we allow for an individual effects structure with correlation with the explanatory variables and/or for sample selection indices with a lagged endogenous variable as explanatory variable. 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 which govern the selection probability may be known in advance. Like misspecification of the parametric form of the selection function, misspecification of the parametric form of the index function results in general in inconsistent estimators of the coefficients in the equation of interest, as pointed out by Ahn and Powell (1993).

Time differencing on the observational equation (2.1) for those individuals which have $d_{it} = d_{is} = 1 \ (s \neq t)$ we get

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + (\varepsilon_{it} - \varepsilon_{is}). \tag{2.3}$$

It might be the case that we do not want to specify any selection indicator function but we just want to assume that selection depends on a $T \times F$ - vector z_i. In this case, by assuming that $(\varepsilon_{it} - \varepsilon_{is})$ is mean independent of x_{it}, x_{is}, conditional on z_i and

$d_{it} = d_{is} = 1$, the expectation of $(\varepsilon_{it} - \varepsilon_{is})$ conditional on selection (i.e. $d_{it} = d_{is} = 1$) is a function of only z_i, so that the expectation of $(y_{it} - y_{is})$ conditional on selection takes the form

$$E[y_{it} - y_{is} | x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1] = (x_{it} - x_{is})\beta + E[\varepsilon_{it} - \varepsilon_{is} | x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1]$$
$$= (x_{it} - x_{is})\beta + \theta_{ts}(z_i), \tag{2.4}$$

and consequently, a selection corrected regression equation for $(y_{it} - y_{is})$ is given by

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + \theta_{ts}(z_i) + (e_{it} - e_{is}), \tag{2.5}$$

where we have taken out from the error term $(\varepsilon_{it} - \varepsilon_{is})$ in (2.3) its conditional mean $E[(\varepsilon_{it} - \varepsilon_{is}) | x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1] = \theta_{ts}(z_i)$ driven by sample selection. Thus, $E[(e_{it} - e_{is}) | x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1] = 0$ by construction and $\theta_{ts}(\cdot)$ is an unknown function of the $T \times F -$ vector z_i.

5.2.2 Identification Issues and Available Estimators

Equation (2.5) provides insight concerning identification. Notice that if some linear combination $(x_{it} - x_{is})\mu$ of $(x_{it} - x_{is})$ where equal to any function of z_i, then there would be asymptotically perfect multicollinearity among the variables on the right-

hand side of equation (2.5), and β could not be estimated from a regression of observed $(y_{it} - y_{is})$ on $(x_{it} - x_{is})$ and $\theta_{ts}(\cdot)$. The reason is that any approximation to the unknown function of z_i, $\theta_{ts}(\cdot)$, will also be able to approximate the linear combination of $(x_{it} - x_{is})\mu$, resulting in asymptotic perfect multicollinearity. To guaranty that taking any nontrivial μ there is no measurable function $\Gamma(z_i)$ such that $(x_{it} - x_{is})\mu = \Gamma(z_i)$ we need to impose the following identification assumption:

Assumption 1: $E\left\{d_{it}d_{is}\left[(x_{it} - x_{is}) - E(x_{it} - x_{is}|z_i)\right]\cdot\left[(x_{it} - x_{is}) - E(x_{it} - x_{is}|z_i)\right]'\right\}$ is non-singular, i.e. for any $\mu \neq 0$ there is no measurable function $\Gamma(z_i)$ such that $(x_{it} - x_{is})\mu = \Gamma(z_i)$.

Accordingly, identification of β requires the strong exclusion restriction that none of the components of $(x_{it} - x_{is})$ can be an exact linear combination of components of z_i. This implies that (x_{it}, x_{is}) and z_i cannot have any components in common.

As in sample selection models typically individual components of the vector z_i appear in the vector of regressors x_{it}, x_{is} in the main equation, we are interested in structures for the selection correction component which permit identification under this situation. If we do not want identification to relay on strong exclusion restrictions we should impose more structure on $\theta_{ts}(z_i)$ for the stochastic restriction $E\left[(e_{it} - e_{is})|x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1\right] = 0$ to identify β. In the literature there are

different ways to impose this structure for models with sample selection. The restricted form of the selection correction in (2.5) is typically derived through imposition of restrictions on the behaviour of the indicator variables $d_{i\tau}(\tau = t,s)$ given z_i; that is, the indicator variables $d_{i\tau}$ are assumed to depend upon $f_\tau(z_i)$ through the binary response model in (2.2). In what remains of this section we make a revision of this literature to understand the contribution of the methods proposed in section 3. The following classification obeys to different degrees of distributional assumptions for the unobservables in the model and to whether or not it is imposed a parametric form for the index function in the selection equation.

Case A.

One way of imposing more structure on the form of the selection correction $\theta_{ts}(z_i)$ is as follows

$$\begin{aligned}
\theta_{ts}(z_i) &= E\left[(\varepsilon_{it} - \varepsilon_{is})|x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1\right] \\
&= E\left[(\varepsilon_{it} - \varepsilon_{is})|x_{it}, x_{is}, z_i, c_i + u_{it} \le f_t(z_i), c_i + u_{is} \le f_s(z_i)\right] \\
&= E\left[(\varepsilon_{it} - \varepsilon_{is})|x_{it}, x_{is}, z_i, c_i + u_{it} \le f(z_i,\gamma_t), c_i + u_{is} \le f(z_i,\gamma_s)\right] \\
&= \Lambda\left\{f(z_i,\gamma_t), f(z_i,\gamma_s); F_3\left[(\varepsilon_{it} - \varepsilon_{is}), (c_i + u_{it}), (c_i + u_{is})|x_{it}, x_{is}, z_i\right]\right\} \\
&= \Lambda\left\{f(z_i,\gamma_t), f(z_i,\gamma_s); F_3\left[(\varepsilon_{it} - \varepsilon_{is}), (c_i + u_{it}), (c_i + u_{is})|f(z_i,\gamma_t), f(z_i,\gamma_s)\right]\right\} \\
&= \Lambda\left\{f(z_i,\gamma_t), f(z_i,\gamma_s)\right\},
\end{aligned} \tag{2.6}$$

where the function $\Lambda\{\cdot,\cdot\}$ is unknown and $f(.,.)$ are scalar single index functions of known parametric form (which can be linear but not necessarily). The joint conditional distribution function F_3 of the error terms $(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i$ depends only upon the double index $\{f(z_t,\gamma_t),f(z_t,\gamma_s)\}$. A consequence of ignorance concerning the form of this distribution is that the functional form of $\Lambda\{\cdot,\cdot\}$ is unknown.

The selection correction term $\theta_{ts}(z_i)$ can be written as in (2.6) when $(\varepsilon_{it}-\varepsilon_{is})$ and $(c_i+u_{it}),(c_i+u_{is})$ are independent of x_{it}, x_{is}, z_i, or alternatively, when $(\varepsilon_{it}-\varepsilon_{is})$ is mean independent of x_{it}, x_{is}, z_i conditional on $(c_t+u_{it}),(c_i+u_{is})$, and $(c_t+u_{it}),(c_t+u_{is})$ are independent of x_{it}, x_{is}, z_i. The conditional mean independence assumption always holds if $[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})]$ is independent of x_{it}, x_{is}, z_i, but we do not require $(\varepsilon_{it}-\varepsilon_{is})$ to be independent of x_{it}, x_{is}, z_i. Under any of the two alternative sets of assumptions the expectation of $(\varepsilon_{it}-\varepsilon_{is})$ conditional on selection (i.e. $d_{it}=d_{is}=1$) is a function of only $\{f(z_i,\gamma_t),f(z_i,\gamma_s)\}$, so that the expectation of $(y_{it}-y_{is})$ conditional on selection takes the form

$$E[y_{it}-y_{is}|x_{it},x_{is},z_i,d_{it}=d_{is}=1]=(x_{it}-x_{is})\beta+\Lambda\{f(z_i,\gamma_t),f(z_i,\gamma_s)\}. \qquad (2.7)$$

The selection corrected regression equation for $(y_{it}-y_{is})$ is given by

$$\begin{aligned} &y_{it} - y_{is} = (x_{it} - x_{is})\beta + \Lambda\{f(z_i,\gamma_t), f(z_i,\gamma_s)\} + e_{its}, \\ &E[e_{its} | x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1] = 0. \end{aligned} \tag{2.8}$$

We need the following identification assumption for β to be identified in (2.8):

Assumption 2:

$$E\left\{d_{it}d_{is}\left[(x_{it} - x_{is}) - E\left(x_{it} - x_{is} | f(z_i,\gamma_t), f(z_i,\gamma_s)\right)\right]\cdot\left[(x_{it} - x_{is}) - E\left(x_{it} - x_{is} | f(z_i,\gamma_t), f(z_i,\gamma_s)\right)\right]'\right\}$$

is non-singular, i.e. for any $\mu \neq 0$ there is no measurable function $\Gamma(f(z_i,\gamma_t), f(z_i,\gamma_s))$ such that $(x_{it} - x_{is})\mu = \Gamma(f(z_i,\gamma_t), f(z_i,\gamma_s))$.

When $f(\cdot,\cdot)$ are non-linear functions identification of β is guaranteed without exclusion restrictions. Under the case of $f(\cdot,\cdot)$ being linear we do not have to explicitly impose exclusion restrictions (as in a cross-section model) because in the panel data case time-variant variables in the selection equation appear as natural exclusion restrictions. For instance, for a given pair (t,s) time-varying variables in the remaining periods of the panel ($\tau = 1,...,T$, $\tau \neq t,s$) will act as exclusion restrictions.

In (2.6) we incorporate more structure on $\theta_{ts}(z_i)$ by adding, as extra identifying information, that the distribution of the indicators $d_{i\tau}$ $(\tau = t,s)$ depends on the double index $\{f(z_i,\gamma_t), f(z_i,\gamma_s)\}$. The double index structure of the selection

correction permits identification even when individual components of the conditioning vector z_i appear in the regressors x_{it}, x_{is}.

Case B.

A fully standard parametric approach applied to (2.6) leads to

$$\begin{aligned}
\theta_{ts}(z_i) &= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right] \\
&= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le f_t(z_i),c_i+u_{is}\le f_s(z_i)\right] \\
&= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le f(z_i,\gamma_t),c_i+u_{is}\le f(z_i,\gamma_s)\right] \\
&= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le z_i\gamma_t,c_i+u_{is}\le z_i\gamma_s\right] \\
&= \Lambda\left\{z_i\gamma_t,z_i\gamma_s;F_3\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i\right]\right\} \\
&= \Lambda\left\{z_i\gamma_t,z_i\gamma_s;\Phi_3\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i\right]\right\},
\end{aligned} \tag{2.9}$$

where $f(\cdot)$ are scalar aggregators in the selection equation of a linear parametric form and we have imposed strong stochastic restrictions by specifying the joint conditional distribution function F_3 of the error terms $(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i$ as a trivariate normal distribution function Φ_3. Under these parametric assumptions, the form of the selection term, to be added as an additional regressor to the differenced equation in (2.3), can be worked out (see chapter 3). Under this fully parametric approach the estimation method developed in chapter 3 consists on a two steps estimator. The method eliminates the individual effects from the equation of interest by taking time differences conditioning to observability of the individual in two time periods. Two correction terms, which form depends upon the linear scalar aggregator

function and the joint trivariate normal distribution function assumed for the unobservables in the model, are worked out. Given consistent first step estimates of these terms, simple least squares in the equation of interest can be used to obtain consistent estimates of β in the second step. Because of the linearity assumption for $f(\cdot)$, the estimator under Case B corresponds to the called "More parametric new estimator" in chapter 3.

Case C.

Relaxing in Case B the parametric form for the index functions $f(.,.)$ we get

$$\begin{aligned}
\theta_{ts}(z_i) &= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right] \\
&= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le f_t(z_i),c_i+u_{is}\le f_s(z_i)\right] \\
&= E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le F^{-1}\left[h_t(z_i)\right],c_i+u_{is}\le F^{-1}\left[h_s(z_i)\right]\right] \\
&= \Lambda\left\{F^{-1}\left[h_t(z_i)\right],F^{-1}\left[h_s(z_i)\right];F_3\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i\right]\right\} \\
&= \Lambda\left\{\Phi^{-1}\left[h_t(z_i)\right],\Phi^{-1}\left[h_s(z_i)\right];\Phi_3\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i\right]\right\},
\end{aligned} \tag{2.10}$$

where the selection indicator indices $f_\tau(\cdot)$, $\tau=t,s$ are unknown and of unrestricted form. We have still imposed as in Case B strong stochastic restrictions by specifying the joint conditional distribution of the errors $(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i$ as trivariate normal. The values of these semiparametric indices in the selection equation are recovered by applying the inversion rule $f_t(z_i)=\Phi^{-1}\left[h_t(z_i)\right]$ and

$f_s(z_i) = \Phi^{-1}[h_s(z_i)]$, where the conditional expectations $h_\tau(z_i) = E(d_{i\tau}|z_i)$ for $\tau = t,s$ are replaced with nonparametric estimators $\hat{h}_\tau(z_i) = \hat{E}(d_{i\tau}|z_i)$, such as kernel estimators, and Φ^{-1} is the inverse of a standard normal cumulative distribution function. Given the unrestricted treatment of the functions $f_\tau(\cdot)$ in (2.2) the estimator under Case C corresponds to the three steps estimator called "Less parametric new estimator" in chapter 3.

Both for Case B and Case C, although chapter's 3 estimators are based upon an independence assumption where $[(\varepsilon_{it} - \varepsilon_{is}),(c_i + u_{it}),(c_i + u_{is})]'$ is independent of x_{it}, x_{is}, z_i with a joint normality of the error terms, for chapter's 3 methods to work, it is sufficient to have a) marginal normality for $(c_i + u_{it}),(c_i + u_{is})$ and consequently joint normality of $(c_i + u_{it})$ and $(c_i + u_{is})$; b) independence of x_{it}, x_{is}, z_i for $(c_i + u_{it})$ and $(c_i + u_{is})$; c) a conditional mean independence assumption of $(\varepsilon_{it} - \varepsilon_{is})$ from x_{it}, x_{is}, z_i once conditioning to $(c_i + u_{it})$ and $(c_i + u_{is})$; d) a linear projection of $(\varepsilon_{it} - \varepsilon_{is})$ on $[(c_i + u_{it}),(c_i + u_{is})]$. Furthermore, the normality of $(c_i + u_{it})$ and $(c_i + u_{is})$ could be relaxed under other distributional assumptions, but it can be difficult to give a closed form for the sample selection correction term as in the normal case. Under a, b, c, and d

$$E\left[(\varepsilon_{it} - \varepsilon_{is})\middle| \nu_{its} = \{(c_i + u_{it}),(c_i + u_{is})\}'\right] = \nu_{its}' E^{-1}[\nu_{its}\nu_{its}'] E[\nu_{its}(\varepsilon_{it} - \varepsilon_{is})] = \nu_{its}'\delta, \tag{2.11}$$

where $\delta = (\delta_{ts}, \delta_{st})' = E^{-1}\left[v_{its} v_{its}'\right] E\left[v_{its}(\varepsilon_{it} - \varepsilon_{is})\right]$.

Then, the selection bias is

$$E\left[(\varepsilon_{it} - \varepsilon_{is}) | c_i + u_{it} \leq f_t(z_i), c_i + u_{is} \leq f_s(z_i)\right] = \delta' \cdot E\left(v_{its} | c_i + u_{it} \leq f_t(z_i), c_i + u_{is} \leq f_s(z_i)\right), \tag{2.12}$$

expression which can be worked out with the results for a truncated normal distribution in Tallis (1961) and which leads to the same sample selection correction terms than in chapter 3 under full joint normality.

Chapter's 3 estimators (under Case B and Case C) do not require technically exclusion restrictions. However, in a panel data model they appear naturally with the presence of any time-varying variable in the selection equation.

Case D.

By following a different approach to Case A, B and C we find

$$\begin{aligned}
&\theta_{ts}(z_{it}, z_{is}) = E\left[(\varepsilon_{it} - \varepsilon_{is}) | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1\right] \\
&= E\left[\varepsilon_{it} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1\right] - E\left[\varepsilon_{is} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, d_{it} = d_{is} = 1\right] = \\
&= \mathrm{E}\left(\varepsilon_{it} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, u_{it} \leq z_{it}\gamma - \eta_i, u_{is} \leq z_{is}\gamma - \eta_i\right) \\
&-\mathrm{E}\left(\varepsilon_{is} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i, u_{is} \leq z_{is}\gamma - \eta_i, u_{it} \leq z_{it}\gamma - \eta_i\right) \\
&= \Lambda\left\{z_{it}\gamma - \eta_i, z_{is}\gamma - \eta_i; F_3\left[\varepsilon_{it}, u_{it}, u_{is} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i\right]\right\} \\
&-\Lambda\left\{z_{is}\gamma - \eta_i, z_{it}\gamma - \eta_i; F_3\left[\varepsilon_{is}, u_{is}, u_{it} | x_{it}, x_{is}, z_{it}, z_{is}, \alpha_i, \eta_i\right]\right\} = 0,
\end{aligned} \tag{2.13}$$

where the equality to zero holds if $z_{it}\gamma = z_{is}\gamma$ and $F_4\left[\varepsilon_{it},\varepsilon_{is},u_{it},u_{is}|x_{it},x_{is},z_{it},z_{is},\alpha_i,\eta_i\right] = F_4\left[\varepsilon_{is},\varepsilon_{it},u_{is},u_{it}|x_{it},x_{is},z_{it},z_{is},\alpha_i,\eta_i\right]$ (joint *conditional exchangeability* assumption). There are no prior distributional assumptions on the unobserved error components but they are subject to this joint *conditional exchangeability* assumption. The idea of imposing these conditions, under which first differencing for a given individual not only eliminates the individual effects in the main equation but also the sample selection effects, is exploited by the estimator developed by Kyriazidou (1997). Conditioning to a given individual the estimation method is developed independently of the individual effects in the selection equation. For this reason we do not need to explicitly consider, parametrically or non-parametrically, the correlation between the individual effects in that equation and the explanatory variables. Implicit in (2.13) there is an indicator variable $d_{it} = 1\left[z_{it}\gamma - \eta_i - u_{it} \geq 0\right]$, which implies that part of the flexibility in (2.2) is suppressed by assuming that $f_t(z_i) - c_i = z_{it}\gamma - \eta_t$.

In Kyriazidou's (1997) model identification of β requires $E\left[(x_{it} - x_{is})'(x_{it} - x_{is})d_{it}d_{is}|(z_{it} - z_{is})\gamma = 0\right]$ to be finite and non-singular. Given that we need support of $(z_{it} - z_{is})\gamma$ at zero, nonsingularity requires an exclusion restriction on the set of regressors, namely that at least one of the variables z_{it} is not contained in x_{it}.

Concluding this section, chapter's 3 approach imposes strong stochastic restrictions, by specifying the joint conditional distribution of the error terms

$(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})$ as trivariate normal. Under this assumption "sample selectivity regressors" that asymptotically purge the equation of interest of its selectivity bias can be computed and the corrected model can be estimated by OLS on the selected subsample of individuals observed the two time periods. However, if the joint distribution of the error terms is misspecified, then the estimator of β will be inconsistent in general. The semiparametric method developed by Kyriazidou (1997) relaxes the assumption of a known parametric form of the joint distribution but it imposes a parametric form for the index function $f_t(.)$ and the named joint *conditional exchangeability* assumption for the time varying errors in the model. The two semiparametric methods for panel data sample selection models proposed in this chapter will avoid the mentioned limitations in the available methods. In particular, no distributional assumptions for the error terms are needed compared with chapter's 3 estimators and no exchangeability is required compared with Kyriazidou (1997).

5.3 The Proposed Estimators

When the selection errors (c_i+u_{it},c_i+u_{is}) are assumed independent of the regressors x_{it},x_{is},z_i, and continuously distributed with support on the entire real line, the conditional distribution of $c_i+u_{i\tau}$ $(\tau=t,s)$ given $f_\tau(z_i)$ –the function $F(\cdot)$ in (2.10)– is an invertible cumulative distribution function and the assumption of a known parametric form of the regression function in the selection equation can be

relaxed. Furthermore, when $F(\cdot)$ is invertible we can allow for a free distribution function for the error terms in the model. To see this, we define the probability to be selected into the sample as

$$E\left[d_{i\tau}|x_{it},x_{is},z_i\right]=E\left[d_{i\tau}|z_i\right]=h_\tau(z_i)=\Pr\left[d_{i\tau}=1|z_i\right]=\Pr\left[c_i+u_{i\tau}\le f_\tau(z_i)|z_i\right]=F\left[f_\tau(z_i)\right].$$
$$\tau=t,s \tag{3.1}$$

Thus, the selection correction function can be determined by

$$\begin{aligned}\theta_{ts}(z_i)&=E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right]\\&=E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le f_t(z_i),c_i+u_{is}\le f_s(z_i)\right]\\&=E\left[(\varepsilon_{it}-\varepsilon_{is})|x_{it},x_{is},z_i,c_i+u_{it}\le F^{-1}\left[h_t(z_i)\right],c_i+u_{is}\le F^{-1}\left[h_s(z_i)\right]\right]\\&=\Lambda\left\{F^{-1}\left[h_t(z_i)\right],F^{-1}\left[h_s(z_i)\right];F_3\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|x_{it},x_{is},z_i\right]\right\}\\&=\Lambda\left\{F^{-1}\left[h_t(z_i)\right],F^{-1}\left[h_s(z_i)\right];F_3\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})|F^{-1}\left[h_t(z_i)\right],F^{-1}\left[h_s(z_i)\right]\right]\right\}\\&=\Lambda\left\{F^{-1}\left[h_t(z_i)\right],F^{-1}\left[h_s(z_i)\right]\right\}=\lambda\left\{h_t(z_i),h_s(z_i)\right\}.\end{aligned} \tag{3.2}$$

The expression in (2.10) differs from the one in (3.2) in that the former was assuming a known parametric distribution function for the errors, to be able to invert the probabilities of selection into the sample for recovering the values of the functions $f_\tau(\cdot)$, $\tau=t,s$. Now, conditional to the function $F(\cdot)$ being invertible, we avoid the need to invert by assuming that the sample selection correction term is an unknown function of the probabilities themselves in place of another unknown function of

$f_\tau(\cdot)$, $\tau = t,s$. Expression (3.2) can be summarised by the following mean "double-index" restriction

$$\begin{aligned} E\left[(\varepsilon_{it} - \varepsilon_{is})|x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1\right] &= E\left[(\varepsilon_{it} - \varepsilon_{is})|h_t(z_i), h_s(z_i), d_{it} = d_{is} = 1\right] \\ &= \lambda\left\{h_t(z_i), h_s(z_i)\right\}, \end{aligned} \tag{3.3}$$

where we need the sample selection correction term, to be included in (2.3), to be a continuous function of the probabilities in (3.1). The regressors x_{it}, x_{is} and z_i should not enter separately into the correction term.

Under the "double-index" restriction in (3.3) we consider estimation of the parameter vector β of a "double-index, partially linear" model of the form[73]

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + \lambda[h_t(z_i), h_s(z_i)] + (e_{it} - e_{is}), \tag{3.4}$$

where $\lambda(\cdot,\cdot)$ is an unknown, smooth function of two scalars, unobservable "indices" $h_t(z_i), h_s(z_i)$. By construction the error term in (3.4) has conditional mean zero,

[73] Our estimation methods rely on a mean double index restriction. However, also the model implies the stronger double index restriction that the conditional distribution of the differenced errors in the main equation given selection and x_{it}, x_{is}, z_i depends only on $h_t(z_i), h_s(z_i)$. Thus the conditional expectation of any function of $(\varepsilon_{it} - \varepsilon_{is})$, and not just $(\varepsilon_{it} - \varepsilon_{is})$ itself, will depend only on $h_t(z_i), h_s(z_i)$. Consequently, the $Var\left[(\varepsilon_{it} - \varepsilon_{is})|x_{it}, x_{is}, z_i, d_{it} = d_{is} = 1\right]$ is restricted to $Var\left[(\varepsilon_{it} - \varepsilon_{is})|h_t(z_i), h_s(z_i), d_{it} = d_{is} = 1\right]$. We are unable to allow for heteroskedasticity in a general form. We allow for conditional heteroskedasticity of the errors as long as it is of double index form. The efficiency of the estimators could be improved by using additional moment restrictions exploiting the stronger distributional indices restriction.

$$E\left[\left(e_{it}-e_{is}\right)\middle|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right]=E\left[\left(e_{it}-e_{is}\right)\middle|h_t\left(z_i\right),h_s\left(z_i\right),d_{it}=d_{is}=1\right]=0 \qquad (3.5)$$

It is interesting, at this stage, to show what happens when we try to develop a semiparametric estimator of β in (3.4) relying only on time differences for a given individual. As a result we get that, even conditioning to probabilities, we cannot avoid the *exchangeability* assumption in Kyriazidou's (1997) estimator. For illustration we decompose the conditional mean of the differenced error in (3.3) in two terms

$$\begin{aligned}&E\left[\left(\varepsilon_{it}-\varepsilon_{is}\right)\middle|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right]=E\left[\varepsilon_{it}\middle|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right]-E\left[\varepsilon_{is}\middle|x_{it},x_{is},z_i,d_{it}=d_{is}=1\right]=\\&\Lambda\left\{h_t\left(z_i\right),h_s\left(z_i\right);F_3\left[\varepsilon_{it},c_i+u_{it},c_i+u_{is}\middle|x_{it},x_{is},z_i\right]\right\}-\Lambda\left\{h_s\left(z_i\right),h_t\left(z_i\right);F_3\left[\varepsilon_{is},c_i+u_{is},c_i+u_{it}\middle|x_{it},x_{is},z_i\right]\right\}\\&=\Lambda_{its}-\Lambda_{ist},\end{aligned} \qquad (3.6)$$

where for $\Lambda_{its}=\Lambda_{ist}$ we need $h_t\left(z_i\right)=h_s\left(z_i\right)$ and the *conditional exchangeability* assumption

$$F_4\left[\varepsilon_{it},\varepsilon_{is},c_i+u_{it},c_i+u_{is}\middle|x_{it},x_{is},z_i\right]\equiv F_4\left[\varepsilon_{is},\varepsilon_{it},c_i+u_{is},c_i+u_{it}\middle|x_{it},x_{is},z_i\right]. \qquad (3.7)$$

Notice that this *conditional exchangeability* assumption implies for any potential first step estimator the conditional stationarity assumption

$$F_{(c_i+u_{it})|x_{it},x_{is},z_i} \equiv F_{(c_i+u_{is})|x_{it},x_{is},z_i}. \tag{3.8}$$

First step estimation methods compatible with this condition are the conditional maximum score estimator (Manski, (1987)), the conditional smoothed maximum score estimator (Kyriazidou, (1994); Charlier, Melenberg, and van Soest, (1995)), and the conditional maximum likelihood estimator (Chamberlain, (1980)). All these methods are developed independently of the individual fixed effects in a *structural* sample selection equation, and for this reason (3.8) can be rewritten as

$$F_{u_{it}|x_{it},x_{is},z_i,c_i} \equiv F_{u_{is}|x_{it},x_{is},z_i,c_i} \leftrightarrow F_{u_{it}|x_{it},x_{is},z_{it},z_{is},\eta_i} \equiv F_{u_{is}|x_{it},x_{is},z_{it},z_{is},\eta_i} \tag{3.9}$$

Furthermore, the use of these methods implies a linearity assumption for the index in the selection rule, which means that $f_t(z_i)$ in (2.2) is assumed to be equal to $z_{it}\gamma - \eta_i + c_i$. According to Ahn and Powell (1993) if the *latent* regression function is linear, conditioning on probabilities is equivalent to conditioning on $z_{i\tau}\gamma, \tau = t,s$. Thus, we will end up with the sample selection correction term of Case D in (2.13) in section 2 above, exploited by the estimation procedure in Kyriazidou (1997). There, it was necessary to assume that a root-n-consistent estimator $\hat{\gamma}$ of the true γ was available. The two estimation methods we propose will avoid this requirement.

5.3.1 Weighted Double Pairwise Difference Estimator (WDPDE)

In sample selection models with cross section data pairwise-difference estimators are

constructed with pairs of observations across individuals. Up to date, in panel data sample selection models they are constructed, not across individuals, but over time for the same individual (Kyriazidou (1997)). In our approach the pairs of observations will be constructed across individuals in differences over time. The motivation of the method is both to eliminate the individual effects and to get ride of sample selection problems. The drawback of Kyriazidou's (1997) estimator was given by the fact that elimination of the sample selection effects needed the named joint *conditional exchangeability* assumption. In our method, given a pair of observations characterised by the vector $\left[\left\{\left(y_{it}-y_{is}\right),\left(x_{it}-x_{is}\right)\right\},\left\{\left(y_{jt}-y_{js}\right),\left(x_{jt}-x_{js}\right)\right\}\right]$ with $d_{it}=d_{is}=1, d_{jt}=d_{js}=1$ and $h_{its}\equiv\left(h_{it},h_{is}\right)=\left(h_{jt},h_{js}\right)\equiv h_{jts}$,

$$\begin{aligned}\left(y_{it}-y_{is}\right)-\left(y_{jt}-y_{js}\right)&=\left[\left(x_{it}-x_{is}\right)-\left(x_{jt}-x_{js}\right)\right]\beta+\\&\left[\lambda\left\{h_t\left(z_i\right),h_s\left(z_i\right)\right\}-\lambda\left\{h_t\left(z_j\right),h_s\left(z_j\right)\right\}\right]+\left[\left(e_{it}-e_{is}\right)-\left(e_{jt}-e_{js}\right)\right]\\&=\left[\left(x_{it}-x_{is}\right)-\left(x_{jt}-x_{js}\right)\right]\beta+\left[\left(e_{it}-e_{is}\right)-\left(e_{jt}-e_{js}\right)\right],\end{aligned}$$

(3.10)

because of

$$\begin{aligned}&E\left[\left(\varepsilon_{it}-\varepsilon_{is}\right)-\left(\varepsilon_{jt}-\varepsilon_{js}\right)\middle|\left\{h_t\left(z_i\right),h_s\left(z_i\right)\right\}=\left\{h_t\left(z_j\right),h_s\left(z_j\right)\right\},d_{it}=d_{is}=1,d_{jt}=d_{js}=1\right]\\&=E\left[\left(\Delta y_{its}-\Delta y_{jts}\right)-\left(\Delta x_{its}-\Delta x_{jts}\right)\beta\middle|\left\{h_t\left(z_i\right),h_s\left(z_i\right)\right\}=\left\{h_t\left(z_j\right),h_s\left(z_j\right)\right\},d_{it}=d_{is}=1,d_{jt}=d_{js}=1\right]\\&=\left[\lambda\left\{h_t\left(z_i\right),h_s\left(z_i\right)\right\}-\lambda\left\{h_t\left(z_j\right),h_s\left(z_j\right)\right\}\right]=0,\end{aligned}$$

(3.11)

where Δ denotes the increment of a variable from period s to t. By construction, in (3.10)

$$E\left[\left(e_{it}-e_{is}\right)-\left(e_{jt}-e_{js}\right)\middle|\left\{h_t\left(z_i\right),h_s\left(z_i\right)\right\}=\left\{h_t\left(z_j\right),h_s\left(z_j\right)\right\},d_{it}=d_{is}=1,d_{jt}=d_{js}=1\right]=0. \tag{3.12}$$

How close are the vectors of conditional means $\left(h_{its},h_{ist}\right)$ in the conditioning set will be weighted by the bivariate kernel weighs

$$\hat{\omega}_{ijts}\equiv\frac{1}{g_{2N}^2}k\left(\frac{\hat{h}_{its}-\hat{h}_{jts}}{g_{2N}}\right)d_{it}d_{is}d_{jt}d_{js}, \tag{3.13}$$

with

$$\hat{h}_t\left(z_i\right)=\frac{\sum_{l\neq i}^{N}K_{il}d_{lt}}{\sum_{l\neq i}^{N}K_{il}},\quad \hat{h}_s\left(z_i\right)=\frac{\sum_{l\neq i}^{N}K_{il}d_{ls}}{\sum_{l\neq i}^{N}K_{il}},\quad K_{il}\equiv K\left(\frac{z_i-z_l}{g_{1N}}\right), \tag{3.14}$$

where k and K are the kernel functions and g_{2N} and g_{1N} are the bandwidths. Thus, we estimate the unobservable conditional expectations in (3.1) by nonparametric kernel regression.

The estimator will be of the form

$$\hat{\beta} = \left[\hat{S}_{xx}\right]^{-1} \hat{S}_{xy},$$
$$\hat{S}_{xx} \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\omega}_{ijts} \left[(x_{it} - x_{is}) - (x_{jt} - x_{js})\right]' \left[(x_{it} - x_{is}) - (x_{jt} - x_{js})\right] \tag{3.15}$$
and
$$\hat{S}_{xy} \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\omega}_{ijts} \left[(x_{it} - x_{is}) - (x_{jt} - x_{js})\right]' \left[(y_{it} - y_{is}) - (y_{jt} - y_{js})\right]$$

Then the WDPDE has a closed form solution that comes from a weighted least squares regression of the distinct differences $(y_{it} - y_{is}) - (y_{jt} - y_{js})$ in dependent variables on the distinct differences $(x_{it} - x_{is}) - (x_{jt} - x_{js})$ in regressors, using $\hat{\omega}_{ijts}$ in (3.13) as bivariate kernel weighs. We only have to include pairs of observations for individuals observed two time periods and we have to exclude pairs of individuals for which $h_{its} \neq h_{jts}$.

The advantages of this estimator are as follows. No distributional assumptions for the error terms are needed compared with the estimators in chapter 3 or Wooldridge (1995), and no *conditional exchangeability* assumption is needed compared with Kyriazidou (1997). We do not require conditions for a given individual over time to eliminate the selection terms but conditions among individuals in time differences.

In general, in (3.2) the unknown joint conditional distribution of $\left[(\varepsilon_{it} - \varepsilon_{is}), (c_i + u_{it}), (c_i + u_{is})\right]$ could differ across individuals as well as across time pairs. In particular, since in (3.10) to eliminate the sample selection effects, differences are taken across individuals in time differences, it is not required that

$\left[(\varepsilon_{it}-\varepsilon_{is}),(c_i+u_{it}),(c_i+u_{is})\right]$ be i.i.d over time but i.i.d across individuals. In other words, we may allow the functional form of F_3 and consequently of λ in (3.2) to vary with the pair (t,s) but not across individuals. We need $\left[(\varepsilon_{it}-\varepsilon_{is}),(\varepsilon_{jt}-\varepsilon_{js}),(c_i+u_{it}),(c_i+u_{is}),(c_j+u_{jt}),(c_j+u_{js})\right]$ to be identically distributed that $\left[(\varepsilon_{jt}-\varepsilon_{js}),(\varepsilon_{it}-\varepsilon_{is}),(c_j+u_{jt}),(c_j+u_{js}),(c_i+u_{it}),(c_i+u_{is})\right]$ conditional on $\{h_t(z_i),h_s(z_i)\}=\{h_t(z_j),h_s(z_j)\}$. That is,

$$F\left[(\varepsilon_{it}-\varepsilon_{is}),(\varepsilon_{jt}-\varepsilon_{js}),(c_i+u_{it}),(c_i+u_{is}),(c_j+u_{jt}),(c_j+u_{js})\middle|\{h_t(z_i),h_s(z_i)\}=\{h_t(z_j),h_s(z_j)\}\right]\equiv$$
$$F\left[(\varepsilon_{jt}-\varepsilon_{js}),(\varepsilon_{it}-\varepsilon_{is}),(c_j+u_{jt}),(c_j+u_{js}),(c_i+u_{it}),(c_i+u_{is})\middle|\{h_t(z_i),h_s(z_i)\}=\{h_t(z_j),h_s(z_j)\}\right] \tag{3.16}$$

This is crucial to our method for eliminating the sample selection effects.

In this model identification of β requires $E\left[\left\{(x_{it}-x_{is})-(x_{jt}-x_{js})\right\}'\left\{(x_{it}-x_{is})-(x_{jt}-x_{js})\right\}d_{it}d_{is}d_{jt}d_{js}\middle|h_{its}-h_{jts}=0\right]$ to be finite and non-singular. Then, if the extra identifying information in (3.1) is exploited, the stochastic restriction (3.5) is sufficient for identification provided the regressors $(x_{it}-x_{is})$ have sufficient variability given the indices h_{its}. This condition rules out any deterministic function of h_{its} as a component of the regression vector $(x_{it}-x_{is})$. Moreover, nonsingularity imposes some restrictions on the form of the selection equation. Given that we require support of $h_{its}-h_{jts}$ at zero, if, for example,

the latent regression function is linear then conditioning on h_{its} is equivalent to conditioning on the linear indices, and nonsingularity requires an exclusion restriction on the set of regressors, namely that at least one of the variables in z_i is not contained in x_{it}. As we already discussed, with panel data these exclusion restrictions appear naturally with the presence of time-varying variables in the selection equation. However, if the true *latent* regression function is non-linear in z we have identification even without exclusion restrictions because these non-linear terms are implicitly excluded from the regression function of interest.

5.3.2 Single Pairwise Difference Estimator (SPDE)

We also consider estimation of the parameter vector β of the "double-index, partially linear" model of (3.4). However, under this alternative estimation procedure we generalise Robinson's (1988) "partially linear" model to the case of panel data sample selection models. In the model in (3.4)

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + \lambda[h_t(z_i), h_s(z_i)] + (e_{it} - e_{is}), \tag{3.17}$$

we have already eliminated the individual effects in the main regression equation by taking time differences for a given individual. If we take conditional expectations in (3.17) we get

$$E\left(y_{it}-y_{is}\left|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right.\right)= \\ E\left(x_{it}-x_{is}\left|h_t(z_i),h_s(z_i)\right.\right)\beta+\lambda[h_t(z_i),h_s(z_i)] \tag{3.18}$$

To get ride of the selection bias in (3.17) we take out from that expression its conditional expectation in (3.18), and then we get the "centred" equation

$$\left(y_{it}-y_{is}\right)-E\left(\left(y_{it}-y_{is}\right)\left|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right.\right)= \\ \left\{\left(x_{it}-x_{is}\right)-E\left(\left(x_{it}-x_{is}\right)\left|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right.\right)\right\}\beta+\left(e_{it}-e_{is}\right). \tag{3.19}$$

To proceed with our estimation strategy, first, we estimate the two indices $h_t(z_i),h_s(z_i)$, which correspond to the probabilities defined in (3.1), with the same nonparametric kernel estimators of (3.14). Second, we insert in (3.19) the nonparametric regression kernel estimators of $E\left(\left(y_{it}-y_{is}\right)\left|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right.\right)$ and $E\left(\left(x_{it}-x_{is}\right)\left|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right.\right)$. By using the same kernel as in (3.14) above, these estimated conditional means are of the form

$$E_N\left(\left(y_{it}-y_{is}\right)\left|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right.\right)=\frac{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\left(y_{jt}-y_{js}\right)}{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}}, \\ E_N\left(\left(x_{it}-x_{is}\right)\left|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right.\right)=\frac{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\left(x_{jt}-x_{js}\right)}{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}}, \tag{3.20}$$

$$\hat{\omega}_{ijts}\equiv\frac{1}{g_{2N}^2}k\left(\frac{\hat{h}_{its}-\hat{h}_{jts}}{g_{2N}}\right)d_{it}d_{is}d_{jt}d_{js}.$$

Finally, in a third step, we apply least squares regression of the differences $(y_{it} - y_{is}) - E_N\left((y_{it} - y_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it} = d_{is} = 1\right)$ on the differences in regressors $(x_{it} - x_{is}) - E_N\left((x_{it} - x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it} = d_{is} = 1\right)$ to get

$$\begin{aligned}
&\hat{\beta} = \left[\hat{S}_{xx}\right]^{-1}\hat{S}_{xy},\\
&\hat{S}_{xx} \equiv \sum_{i=1}^{N} d_{it}d_{is}\left\{(x_{it} - x_{is}) - E_N\left((x_{it} - x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it} = d_{is} = 1\right)\right\}'\\
&\qquad\cdot\left\{(x_{it} - x_{is}) - E_N\left((x_{it} - x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it} = d_{is} = 1\right)\right\}
\end{aligned} \tag{3.21}$$

and

$$\begin{aligned}
\hat{S}_{xy} \equiv &\sum_{i=1}^{N} d_{it}d_{is}\left\{(x_{it} - x_{is}) - E_N\left((x_{it} - x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it} = d_{is} = 1\right)\right\}'\\
&\cdot\left\{(y_{it} - y_{is}) - E_N\left((y_{it} - y_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it} = d_{is} = 1\right)\right\}.
\end{aligned}$$

Identification of β requires that none of the components of $(x_{it} - x_{is})$ can be exact linear combinations of components of $\left[h_t(z_i),h_s(z_i)\right]$. Then, as for the WDPDE, we have identification provided the regressors $(x_{it} - x_{is})$ have sufficient variability given the indices h_{its}.

5.4 Relationship Between the WDPDE and the SPDE

We can extend the WDPDE and the SPDE to allow for endogeneity of some

components of the regressors in the main equation, using an instrumental variables version of both estimators. The exogenous variables z_i can be used to construct a k-dimensional vector (dimension of x_{it}) of "instrumental variables" for $(x_{it} - x_{is})$. In particular, if we let the instruments be suitable functions of the conditioning variables z_i and z_j, algebraically these instruments are defined as $Z_{its} \equiv Z_{ts}(z_i)$ for some function Z_{ts}: $\Re^{F*T} \to \Re^k$. The WDPDE in (3.15) rewritten as a weighted instrumental variables estimator is given by the following expression:

$$\hat{\beta} = \left[\hat{S}_{Zx}\right]^{-1} \hat{S}_{Zy},$$
$$\hat{S}_{Zx} \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\omega}_{ijts} \left[Z_{its} - Z_{jts}\right]\left[(x_{it} - x_{is}) - (x_{jt} - x_{js})\right] \tag{4.1}$$

and

$$\hat{S}_{Zy} \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\omega}_{ijts} \left[Z_{its} - Z_{jts}\right]\left[(y_{it} - y_{is}) - (y_{jt} - y_{js})\right]$$

For the SPDE in (3.21) we can also present an instrumental variables version. As in some other applications of kernel regression estimators, $E_N\left((y_{it} - y_{is}) \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)$ and $E_N\left((x_{it} - x_{is}) \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)$ cause technical difficulties associated with its random denominators, which can be small (which need not be bounded away from zero). To avoid this problem a convenient choice of instrumental variables is the product of the original instruments Z_{its} with the sum in the denominators of $E_N\left((y_{it} - y_{is}) \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)$ and

$E_N\left((x_{it}-x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)$; that is, the instruments $\hat{Z}_{its}$ are defined as

$$\hat{Z}_{its} \equiv Z_{its}\cdot\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\,. \tag{4.2}$$

Instrumental variables regression, using as instruments $\hat{Z}_{its}$ in (4.2), of

$(y_{it}-y_{is})-E_N\left((y_{it}-y_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)$ on

$(x_{it}-x_{is})-E_N\left((x_{it}-x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)$ gives

$$\begin{aligned}
&\hat{\beta}=\left[\hat{S}_{Zx}\right]^{-1}\hat{S}_{Zy},\\
&\hat{S}_{Zx}\equiv\sum_{i=1}^{N}d_{it}d_{is}\left(Z_{its}\cdot\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\right)\left\{(x_{it}-x_{is})-E_N\left((x_{it}-x_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\right\}\\
&\quad=\sum_{i=1}^{N}d_{it}d_{is}\left(Z_{its}\cdot\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\right)\left\{(x_{it}-x_{is})-\frac{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}(x_{jt}-x_{js})}{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}}\right\}\\
&\hat{S}_{Zy}\equiv\sum_{i=1}^{N}d_{it}d_{is}\left(Z_{its}\cdot\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\right)\left\{(y_{it}-y_{is})-E_N\left((y_{it}-y_{is})\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\right\}\\
&\quad=\sum_{i=1}^{N}d_{it}d_{is}\left(Z_{its}\cdot\sum_{j\neq i}^{N}\hat{\omega}_{ijts}\right)\left\{(y_{it}-y_{is})-\frac{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}(y_{jt}-y_{js})}{\sum_{j\neq i}^{N}\hat{\omega}_{ijts}}\right\}.
\end{aligned} \tag{4.3}$$

The estimator in (4.3) can be shown to be algebraically equivalent to the $\hat{\beta}$ defined in

(4.1) above[74]. In both cases identification of β requires $E\left[\left[Z_{its}-Z_{jts}\right]\left\{\left(x_{it}-x_{is}\right)-\left(x_{jt}-x_{js}\right)\right\}d_{it}d_{is}d_{jt}d_{js}\middle|h_{its}-h_{jts}=0\right]$ to be finite and non-singular. With instrumental variables we require the covariance matrix of the instruments with the explanatory variables to be of full rank. Although the difference in the indices $h_{its}-h_{jts}$ is zero, the distribution of the corresponding differences $\left[Z_{its}-Z_{jts}\right]$ of instrumental variables and the regressors $\left(x_{it}-x_{is}\right)-\left(x_{jt}-x_{js}\right)$ can still be of full dimension provided the Z_{ts} and the $\left(x_t-x_s\right)$ have sufficient variability given the indices h_{ts}. This condition rules out any deterministic function of h_{ts} as a component of the regression vector $\left(x_t-x_s\right)$. If the latent regression function is linear nonsingularity requires some component of z_i to be excluded from both $\left(x_{it}-x_{is}\right)$ and Z_{its}. With panel data exclusion restrictions appear naturally for $\left(x_{it}-x_{is}\right)$, but for Z_{its} we need an exclusion restriction, which may be difficult to justify. In general, though, if the true *latent* regression function is non-linear in z we have identification even without exclusion restrictions, because these non-linear terms are implicitly excluded from the regression function of interest.

[74] To show that they are equivalent we use the symmetry property of the kernel function, that is, $K_{ij}=K_{ji}$.

5.5 Monte Carlo Results

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

$$\begin{aligned} y_{it} &= d_{it}\left[x_{it}\beta + \alpha_i + \varepsilon_{it}\right]; \quad i = 1,\ldots,N; \quad t = 1,2, \\ d_{it}^{*} &= z_{1it}\gamma_1 + z_{2it}\gamma_2 + \left(z_{1it}\cdot z_{2it}\right)\gamma_3 - \eta_i - u_{it}; \quad d_{it} = 1\left[d_{it}^{*} \geq 0\right], \end{aligned} \tag{5.1}$$

where y_{it} is observed if $d_{it} = 1$. The true value of β, γ_1, γ_2 and γ_3 is 1; the regressors in both equations are exogenous variables where z_{1it} and z_{2it} follow a N(0,1) and x_{it} is equal to the variable z_{2it}. The individual effects are generated as $\eta_i = -\left[\left(z_{1i1} + z_{1i2}\right)/2 + \left(z_{2i1} + z_{2i2}\right)/2 + \left(z_{1i1}\cdot z_{2i1} + z_{1i2}\cdot z_{2i2}\right)/2 + \chi_2^2(0,1) + 0.07\right]$ and $\alpha_i = \left(x_{i1} + x_{i2}\right)/2 + \sqrt{2}\cdot\chi_2^2(0,1) + 1$. The function $f_t(z_i)$ in (2.2) corresponds to $0.07 + z_{1it}\gamma_1 + z_{2it}\gamma_2 + \left(z_{1it}\cdot z_{2it}\right)\gamma_3 + \left(z_{1i1} + z_{1i2}\right)/2 + \left(z_{2i1} + z_{2i2}\right)/2 + \left(z_{1i1}\cdot z_{2i1} + z_{1i2}\cdot z_{2i2}\right)/2$, and the random term c_i to $\chi_2^2(0,1)$. The particular design for $f_t(z_i)$ is driven by the fact that we do not need to restrict the index function in the selection equation to be linear. The time varying errors in the model are $u_{it} = \chi_2^2(0,1)$ and $\varepsilon_{it} = 0.8\cdot u_{it} + 0.6\cdot\chi_2^2(0,1)$. The errors in the main equation are generated as a linear function of a random component and the errors in the selection equation, what guarantees the existence of non-random selection into the sample. We report results

when normalised and central χ^2 distributions with 2 degrees of freedom are considered. Our estimators are distributionally free methods and therefore they are robust to any distributional assumption.

The results with 100 replications and different sample sizes (N= 250, 500, 750 and 1000) are presented in Table 1 (WDPDE) and Table 2 (SPDE). It is fair to say that we will probably need bigger sample sizes than the ones included in the experiments to exploit the properties of these estimators. The tables report the estimated mean bias for the estimators, the small sample standard errors (SE), and as not all the moments of the estimators may exist in finite samples some measures based on quantiles, as the median bias, and the median absolute deviation (MAD) are also reported. In Panel A we report the finite sample properties of the estimator that ignores sample selection. The purpose in presenting these results is to make explicit the importance of the sample selection problem in our experiment design. In Table 1, this estimator is obtained by applying least squares to the model in double differences where correction for sample selection has been ignored, and for the sample of individuals who are observed in both time periods, i.e. those for whom $d_{i1} = d_{i2} = 1$. In Table 2, it is obtained by applying least squares to the model in single differences over time for a given individual observed the two time periods.

In Panels B and C we implement second (R=1), fourth (R=3), and sixth (R=5) higher order bias reducing kernels of Bierens (1987). They correspond to a normal, to a mixture of two normals and to a mixture of three normals, respectively. The

TABLE 1: Weighted Double Pairwise Difference Estimator (WDPDE)

$$u_{it} = \chi_2^2(0,1)$$

$$\varepsilon_{it} = 0.8 \cdot u_{it} + 0.6 \cdot \chi_2^2(0,1)$$

$$\alpha_i = (x_{i1} + x_{i2})/2 + \sqrt{2} \cdot \chi_2^2(0,1) + 1$$

$$\eta_i = -\left[(z_{1i1} + z_{1i2})/2 + (z_{2i1} + z_{2i2})/2 + (z_{1i1} \cdot z_{2i1} + z_{1i2} \cdot z_{2i2})/2 + \chi_2^2(0,1) + 0.07\right]$$

PANEL A

Ignoring Correction For Sample Selection				
N	**Mean Bias**	**Median Bias**	**SE**	**MAD**
250	0.1099	0.1186	0.1416	0.1194
500	0.0937	0.1005	0.1239	0.1024
750	0.0933	0.0911	0.1075	0.0911
1000	0.0912	0.0887	0.1015	0.0887

PANEL B

	R=1 & g=1				R=1 & g=0.5				R=1 & g=3			
N	**Mean Bias**	**Median Bias**	**SE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**MAD**
250	0.0650	0.0735	0.1444	0.1143	0.0969	0.0842	0.1811	0.1194	0.0917	0.0913	0.1365	0.0966
500	0.0426	0.0496	0.1139	0.0747	0.0679	0.0806	0.1419	0.1101	0.0762	0.0816	0.1137	0.0861
750	0.0258	0.0248	0.0827	0.0580	0.0499	0.0398	0.1025	0.0655	0.0707	0.0777	0.0935	0.0777
1000	0.0282	0.0244	0.0782	0.0580	0.0570	0.0629	0.0991	0.0744	0.0700	0.0654	0.0882	0.0654

PANEL C

	R=3 & g=1				R=5 & g=1			
N	**Mean Bias**	**Median Bias**	**SE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**MAD**
250	0.0713	0.0754	0.1465	0.1055	0.0852	0.0879	0.1439	0.1051
500	0.0576	0.0605	0.1112	0.0888	0.0669	0.0722	0.1099	0.0808
750	0.0464	0.0550	0.0864	0.0709	0.0700	0.0612	0.0949	0.0625
1000	0.0531	0.0589	0.0844	0.0662	0.0735	0.0749	0.0907	0.0751

TABLE 2: Single Pairwise Difference Estimator (SPDE)

$$u_{it} = \chi_2^2(0,1)$$

$$\varepsilon_{it} = 0.8 \cdot u_{it} + 0.6 \cdot \chi_2^2(0,1)$$

$$\alpha_i = (x_{i1} + x_{i2})/2 + \sqrt{2} \cdot \chi_2^2(0,1) + 1$$

$$\eta_i = -\left[(z_{1i1} + z_{1i2})/2 + (z_{2i1} + z_{2i2})/2 + (z_{1i1} \cdot z_{2i1} + z_{1i2} \cdot z_{2i2})/2 + \chi_2^2(0,1) + 0.07\right]$$

PANEL A

Ignoring Correction For Sample Selection				
N	**Mean Bias**	**Median Bias**	**SE**	**MAD**
250	0.1090	0.1156	0.1412	0.1182
500	0.0940	0.1001	0.1242	0.1011
750	0.0930	0.0906	0.1074	0.0906
1000	0.0911	0.0889	0.1014	0.0889

PANEL B

	R=1 & g=1				R=1 & g=0.5				R=1 & g=3			
N	**Mean Bias**	**Median Bias**	**SE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**MAD**
250	0.0448	0.0431	0.1373	0.1029	0.0910	0.1039	0.1497	0.1114	0.0705	0.0670	0.1255	0.0861
500	0.0165	0.0134	0.0942	0.0583	0.0494	0.0554	0.1028	0.0773	0.0616	0.0684	0.1037	0.0818
750	0.0074	0.0167	0.0704	0.0510	0.0441	0.0431	0.0792	0.0547	0.0443	0.0505	0.0753	0.0550
1000	0.0063	0.0053	0.0641	0.0431	0.0432	0.0470	0.0718	0.0597	0.0472	0.0409	0.0708	0.0495

PANEL C

	R=3 & g=1				R=5 & g=1			
N	**Mean Bias**	**Median Bias**	**SE**	**MAD**	**Mean Bias**	**Median Bias**	**SE**	**MAD**
250	0.0749	0.0680	0.1550	0.1054	0.0764	0.0672	0.1354	0.0989
500	0.0459	0.0526	0.1277	0.0781	0.0552	0.0703	0.2379	0.0844
750	0.0471	0.0356	0.1277	0.0562	0.0704	0.0525	0.1934	0.0706
1000	0.0370	0.0379	0.0837	0.0578	0.0729	0.0690	0.1600	0.0746

bandwidth sequence for the first step is[75] $g_{1N} = g_1 \cdot N^{-1/[2(R+1)+2T\cdot F]}$, where T=2 is the number of time periods and F=2 is the dimension of $z_{i\tau}$. The first step probabilities $h_1(z_i)$ and $h_2(z_i)$ are estimated by *leave-one-out* kernel estimators constructed as in (3.14). The bandwidth sequence for the weights in the second step of the WDPDE and the SPDE is $g_{2N} = g_2 \cdot N^{-1/[2(R+1)+2q]}$, where q=2 is the dimension of the vectors h_{ts}. The constant part of the bandwidths was chosen equal to 1, 0.5 or 3 in both steps. There was no serious attempt at optimal choice.

From both tables we see that in Panels B and C the estimators are less biased than the estimator ignoring correction for sample selection (Panel A). The bias are all positive, they increase as the kernel order increases and they diminish with sample size. The best behaviour is found with R=1 (a second order kernel) and a constant part of the bandwidth $g_1 = g_2 = g = 1$. Some anomalous results for sample size 1000 may be claiming the use of some trimming to ensure that all the kernel estimators are well behaved. The SPDE performs slightly better than the WDPDE, which can have its origin on the extra differencing present in the latter method.

---

[75] 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 $g_{1N} = g_1 \cdot N^{-1/[2(R+1)+T\cdot F]}$.

5.6 Concluding Remarks

In this chapter, estimation of the coefficients in a "double-index" selectivity bias model is considered under the assumption that the selection correction function depends only on the conditional means of some observable selection variables. We present two alternative methods. The first is a "weighted double pairwise difference estimator" because it is based in the comparison of individuals in time differences. The second is a "single pairwise difference estimator" because only differences over time for each individual are required. Their advantages are the following. They are distributionally free estimators compared with our earlier estimator in chapter 3 and Wooldridge's (1995) estimator. Furthermore, no *conditional exchangeability* assumption or parametric sample selection index is required compared with Kyriazidou (1997). The methods do not require strict exogeneity for the variables in the main equation, and they are shown to be equivalent under the proper choice of instruments for each estimator.

The finite sample properties of the estimators are investigated by Monte Carlo experiments. The results of our small Monte Carlo simulation study show the following. Both estimators are less biased than the estimator ignoring correction for sample selection. The bias are all positive, they increase as the kernel order increases and they diminish with sample size. The best behaviour is found with R=1 (a second order kernel) and a constant part of the bandwidth g=1. The SPDE performs slightly better than the WDPDE, which can have its origin on the extra differencing present in the latter method.

5.7 Appendix I: The Variance-Covariance Matrix for the WDPDE

One variation inside the termed "semiparametric M-estimators" by Horowitz (1988) defines the WDPDE of β as a minimaser of a second-order (bivariate) U-statistic,

$$\hat{\beta} = \underset{\beta \in \mathrm{B}}{\arg\min} \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ \left[\left(\Delta y_{its} - \Delta y_{jts} \right) - \left(\Delta x_{its} - \Delta x_{jts} \right) \beta \right] \sqrt{\hat{\omega}_{ijts}} \right\}^2 \equiv \underset{\beta \in \mathrm{B}}{\arg\min} \; U_{0N}(\beta), \tag{I.1}$$

that will solve an approximate first order condition

$$\binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\Delta x_{its} - \Delta x_{jts} \right)' \left[\left(\Delta y_{its} - \Delta y_{jts} \right) - \left(\Delta x_{its} - \Delta x_{jts} \right) \hat{\beta} \right] \hat{\omega}_{ijts} = 0, \tag{I.2}$$

where $\Delta y_{its} = y_{it} - y_{is}$, $\Delta x_{its} = x_{it} - x_{is}$, and $\hat{\omega}_{ijts}$ is defined by expression (3.13) in the main text above. The empirical loss-function in (I.1) and the estimating equations in (I.2) also depend upon an estimator of the nonparametric components h_{its} and h_{ist} defined in (3.1) and (3.14). To derive the influence function for an estimator satisfying (I.2), we first do an expansion around β of (I.2) and subsequently a functional mean-value expansion around $\left(h_{its} - h_{jts} \right)$ to determine the effect on $\hat{\beta}$ of estimation of $\left(\hat{h}_{its} - \hat{h}_{jts} \right)$.

Expanding (I.2) around β we get

$$
\begin{aligned}
0 = &\binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\Delta x_{its} - \Delta x_{jts}\right)' \left[\left(\Delta y_{its} - \Delta y_{jts}\right) - \left(\Delta x_{its} - \Delta x_{jts}\right)\beta\right] \hat{\omega}_{ijts} \\
&- \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\Delta x_{its} - \Delta x_{jts}\right)' \left(\Delta x_{its} - \Delta x_{jts}\right) \hat{\omega}_{ijts} \left(\hat{\beta} - \beta\right),
\end{aligned}
\tag{I.3}
$$

from where

$$
\begin{aligned}
\sqrt{N}\left(\hat{\beta} - \beta\right) = &\left\{ \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\Delta x_{its} - \Delta x_{jts}\right)' \left(\Delta x_{its} - \Delta x_{jts}\right) \hat{\omega}_{ijts} \right\}^{-1} \cdot \\
&\sqrt{N} \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\Delta x_{its} - \Delta x_{jts}\right)' \left(v_{its} - v_{jts}\right) \hat{\omega}_{ijts} \equiv \\
&\left\{\hat{S}_{xx}\right\}^{-1} \sqrt{N} \hat{S}_{xv},
\end{aligned}
\tag{I.4}
$$

being $\left(v_{its} - v_{jts}\right) \equiv \left[\left(\varepsilon_{it} - \varepsilon_{is}\right) - \left(\varepsilon_{jt} - \varepsilon_{js}\right)\right] \equiv \left[\left(\Delta y_{its} - \Delta y_{jts}\right) - \left(\Delta x_{its} - \Delta x_{jts}\right)\beta\right]$.

If we analyse the components of (I.4):

$$
1) \quad \hat{S}_{xx} \equiv 2 \cdot \frac{1}{N} \sum_{i=1}^{N-1} \frac{1}{N-1} \sum_{j=i+1}^{N} \left(\Delta x_{its} - \Delta x_{jts}\right)' \left(\Delta x_{its} - \Delta x_{jts}\right) \hat{\omega}_{ijts} =^{p} S_{xx} =^{p} 2 \cdot \sum_{xx} \quad . \tag{I.5}
$$

As $S_{xx} = U_{1N}$, that is a bivariate U-statistic, by using U-statistics asymptotic theory we know $\sqrt{N} U_{1N} =^{p} 2 \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} E\left[\left(\Delta x_{its} - \Delta x_{jts}\right)' \left(\Delta x_{its} - \Delta x_{jts}\right) \omega_{ijts} \middle| \Delta x_{its}, h_{its}, d_{it}, d_{is}\right]$

and then

$$U_{1N} =^{p} 2\cdot\frac{1}{N}\sum_{i=1}^{N} E\left[\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(\Delta x_{its}-\Delta x_{jts}\right)\omega_{ijts}\left|\Delta x_{its},h_{its},d_{it},d_{is}\right.\right]= \\ 2\cdot E\left\{E\left[\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(\Delta x_{its}-\Delta x_{jts}\right)\omega_{ijts}\left|\Delta x_{its},h_{its},d_{it},d_{is}\right.\right]\right\}=2\cdot\sum_{xx}\ . \tag{I.6}$$

The matrix $\sum_{xx}$ is easily handled, since $\frac{1}{2}\hat{S}_{xx}$ consistently estimates it.

2) $\sqrt{N}\hat{S}_{xv}$ expanded around $\left(h_{its}-h_{jts}\right)$,

$$\sqrt{N}\hat{S}_{xv}\equiv\sqrt{N}\binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(v_{its}-v_{jts}\right)\frac{1}{g_{2N}^{2}}k\left(\frac{\hat{h}_{its}-\hat{h}_{jts}}{g_{2N}}\right)d_{its}d_{jts}= \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N-1}\frac{2}{N-1}\sum_{j=i+1}^{N}\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(v_{its}-v_{jts}\right)\frac{1}{g_{2N}^{2}}k\left(\frac{h_{its}-h_{jts}}{g_{2N}}\right)d_{its}d_{jts}+ \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N-1}\frac{2}{N-1}\sum_{j=i+1}^{N}\sum_{l=1}^{N}\left(\Delta x_{jts}-\Delta x_{lts}\right)'\left(v_{jts}-v_{lts}\right)\frac{1}{g_{2N}^{3}}k'\left(\frac{h^{*}_{its}-h^{*}_{jts}}{g_{2N}}\right)d_{jts}d_{lts}K_{jl}\left[\sum_{l=1}^{N}K_{jl}\right]^{-1}\left(\binom{d_{it}}{d_{is}}-\hat{h}_{its}\right) \tag{I.7}$$

where $d_{its}\equiv d_{it}d_{is}$, $k'(.)$ is the derivative of the second-stage kernel $k(.)$ and K_{jl} is defined as in (3.14). The expression (I.7) includes derivatives of the weights with respect to $\left(h_{its}-h_{jts}\right)$ (kernel derivatives).

For the first term on the right hand side of (I.7),

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N-1}\frac{2}{N-1}\sum_{j=i+1}^{N}\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(v_{its}-v_{jts}\right)\frac{1}{g_{2N}^{2}}k\left(\frac{h_{its}-h_{jts}}{g_{2N}}\right)d_{its}d_{jts}=^{p}$$
$$\frac{2}{\sqrt{N}}\sum_{i=1}^{N}E\left[\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(v_{its}-v_{jts}\right)\frac{1}{g_{2N}^{2}}k\left(\frac{h_{its}-h_{jts}}{g_{2N}}\right)d_{its}d_{jts}\left|\Delta x_{its},v_{its},h_{its},d_{its}\right.\right], \tag{I.8}$$

and for the second,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N-1}\frac{2}{N-1}\sum_{j=i+1}^{N}\sum_{l=1}^{N}\left(\Delta x_{jts}-\Delta x_{lts}\right)'\left(v_{jts}-v_{lts}\right)\frac{1}{g_{2N}^{3}}k'\left(\frac{h^{*}_{its}-h^{*}_{jts}}{g_{2N}}\right)d_{jts}d_{lts}K_{jl}\left[\sum_{l=1}^{N}K_{jl}\right]^{-1}\left(\begin{pmatrix}d_{it}\\ d_{is}\end{pmatrix}-\hat{h}_{its}\right)=^{p}$$
$$\frac{2}{\sqrt{N}}\sum_{i=1}^{N}E\left\{\sum_{l=1}^{N}\left(\Delta x_{jts}-\Delta x_{lts}\right)'\left(v_{jts}-v_{lts}\right)\frac{1}{g_{2N}^{3}}k'\left(\frac{h_{its}-h_{jts}}{g_{2N}}\right)d_{jts}d_{lts}K_{jl}\left[\sum_{l=1}^{N}K_{jl}\right]^{-1}\middle|h_{its}\right\}\cdot\left(\begin{pmatrix}d_{it}\\ d_{is}\end{pmatrix}-\hat{h}_{its}\right) \tag{I.9}$$

Substituting (I.5), (I.8) and (I.9) in (I.4) we get

$$\sqrt{N}\left(\hat{\beta}-\beta\right)=^{p}\left\{\sum\nolimits_{xx}\right\}^{-1}\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{E\left[\left(\Delta x_{its}-\Delta x_{jts}\right)'\left(v_{its}-v_{jts}\right)\frac{1}{g_{2N}^{2}}k\left(\frac{h_{its}-h_{jts}}{g_{2N}}\right)d_{its}d_{jts}\left|\Delta x_{its},v_{its},h_{its},d_{its}\right.\right]+\right.$$
$$\left.E\left[\sum_{l=1}^{N}\left(\Delta x_{jts}-\Delta x_{lts}\right)'\left(v_{jts}-v_{lts}\right)\frac{1}{g_{2N}^{3}}k'\left(\frac{h_{its}-h_{jts}}{g_{2N}}\right)d_{jts}d_{lts}K_{jl}\left[\sum_{l=1}^{N}K_{jl}\right]^{-1}\middle|h_{its}\right]\cdot\left(\begin{pmatrix}d_{it}\\ d_{is}\end{pmatrix}-\hat{h}_{its}\right)\right\} \tag{I.10}$$

that is asymptotically normal,

$$\sqrt{N}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,\ \sum\nolimits_{xx}^{-1}\Omega_{xx}\left[\sum\nolimits_{xx}^{-1}\right]'\right), \tag{I.11}$$

where $\sum_{xx}$ is estimated by $\frac{1}{2}\hat{S}_{xx}$ by (I.5), and Ω_{xx} can be estimated by

$$\hat{\Omega}_{xx} \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\psi}_{its} + \hat{\zeta}_{its}\left(\begin{pmatrix} d_{it} \\ d_{is} \end{pmatrix} - \hat{h}_{its}\right)\right]\left[\hat{\psi}_{its} + \hat{\zeta}_{its}\left(\begin{pmatrix} d_{it} \\ d_{is} \end{pmatrix} - \hat{h}_{its}\right)\right]' \tag{I.12}$$

where

$$\hat{\psi}_{its} \equiv \frac{1}{N-1}\sum_{j=1}^{N}\left(\Delta x_{its} - \Delta x_{jts}\right)'\left(\hat{v}_{its} - \hat{v}_{jts}\right)\frac{1}{g_{2N}^{2}}k\left(\frac{\hat{h}_{its} - \hat{h}_{jts}}{g_{2N}}\right)d_{its}d_{jts},$$

$$\hat{\zeta}_{its} \equiv \frac{1}{N-1}\sum_{j=1}^{N}\sum_{l=1}^{N}\left[\left(\Delta x_{jts} - \Delta x_{lts}\right)'\left(\hat{v}_{jts} - \hat{v}_{lts}\right)\frac{1}{g_{2N}^{3}}k'\left(\frac{\hat{h}_{its} - \hat{h}_{jts}}{g_{2N}}\right)d_{jts}d_{lts}K_{jl}\left[\sum_{l=1}^{N}K_{jl}\right]^{-1}\right] \tag{I.13}$$

The general theory derived for minimisers of m^{th}-order U-statistics can be applied to show $\sqrt{N}$ – consistency and to obtain the large sample distribution of the WDPDE for panel data sample selection models. The variance-covariance matrix for this estimator depends upon the conditional variability of the errors in the regression equation and the deviations of the selection indicators from their conditional means, $\begin{pmatrix} d_{it} \\ d_{is} \end{pmatrix} - \hat{h}_{its}$.

5.8 Appendix II: The Variance-Covariance Matrix for the SPDE

We can define the SPDE of β as a minimaser of

$$\hat{\beta} = \underset{\beta \in \mathrm{B}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} \left\langle \left\{ \left[\Delta y_{its} - E_N\left(\Delta y_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] - \left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] \cdot \beta \right\} d_{it} d_{is} \right\rangle^2 \tag{II.1}$$

that will solve an approximate first order condition

$$-\frac{1}{N}\sum_{i=1}^{N}\left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right]' \cdot \left\{ \left[\Delta y_{its} - E_N\left(\Delta y_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] - \left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] \cdot \hat{\beta} \right\} d_{it} d_{is} = 0, \tag{II.2}$$

Expanding (II.2) around β we get

$$\begin{aligned} 0 = &-\frac{1}{N}\sum_{i=1}^{N}\left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right]' \cdot \\ &\left\{ \left[\Delta y_{its} - E_N\left(\Delta y_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] - \left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] \cdot \beta \right\} d_{it} d_{is} + \\ &\frac{1}{N}\sum_{i=1}^{N}\left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right]' \left[\Delta x_{its} - E_N\left(\Delta x_{ts} \middle| \hat{h}_t(z_i), \hat{h}_s(z_i), d_{it} = d_{is} = 1\right)\right] d_{it} d_{is} \cdot \\ &\left(\hat{\beta} - \beta\right) \end{aligned} \tag{II.3}$$

from where

$$\sqrt{N}\left(\hat{\beta}-\beta\right)=$$
$$\left\{\frac{1}{N}\sum_{i=1}^{N}\left[\Delta x_{its}-E_N\left(\Delta x_{ts}\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\right]'\left[\Delta x_{its}-E_N\left(\Delta x_{ts}\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\right]d_{it}d_{is}\right\}^{-1}$$
$$\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\Delta x_{its}-E_N\left(\Delta x_{ts}\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\right]'\left[\left(\varepsilon_{it}-\varepsilon_{is}\right)-E_N\left(\varepsilon_t-\varepsilon_s\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\right]$$
$$\cdot d_{it}d_{is}$$

(II.4)

where $\left(\varepsilon_{it}-\varepsilon_{is}\right)=\Delta y_{its}-\Delta x_{its}\beta$, and

$$-E_N\left(\varepsilon_t-\varepsilon_s\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)=-E_N\left(\Delta y_{its}\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)+$$
$$E_N\left(\Delta x_{its}\middle|\hat{h}_t(z_i),\hat{h}_s(z_i),d_{it}=d_{is}=1\right)\beta.$$

(II.5)

It can be shown that the inverted matrix in (II.4) is consistent for

$$A=$$
$$E\left\{\left[\Delta x_{ts}-E\left(\Delta x_{ts}\middle|h_t(z),h_s(z),d_t=d_s=1\right)\right]'\left[\Delta x_{ts}-E\left(\Delta x_{ts}\middle|h_t(z),h_s(z),d_t=d_s=1\right)\right]d_td_s\right\}.$$

(II.6)

We shall analyse now the term $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}$ in (II.4). We have to work out the

effect of estimating four infinite dimensional conditional means $\left(h_t(z),\ h_s(z),\ E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right),\ E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right)$ on the asymptotic variance of our parameter of interest β. The moment condition for the summand in $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}$ can be written as

$$E\left\{m\left[h_t(z),\ h_s(z),\ E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right),\ E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right]\right\}=0 \tag{II.7}$$

where

$$m\left[h_t(z),\ h_s(z),\ E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right),\ E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right]=$$
$$\left[\Delta x_{ts}-E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right)\right]'\left[\left(\varepsilon_t-\varepsilon_s\right)-E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right]d_td_s. \tag{II.8}$$

The following four derivatives are of interest:

$$\frac{\partial m}{\partial E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right)}=-\left[\left(\varepsilon_t-\varepsilon_s\right)-E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right]d_td_s, \tag{II.9}$$

$$\frac{\partial m}{\partial E\left(\varepsilon_t - \varepsilon_s \middle| h_t(z), h_s(z), d_t = d_s = 1\right)} = -\left[\Delta x_{ts} - E\left(\Delta x_{ts} \middle| h_t(z), h_s(z), d_t = d_s = 1\right)\right] d_t d_s ,$$

(II.10)

$$\frac{\partial m}{\partial h_t(z)} = -\nabla_t E\left(\Delta x_{ts} \middle| h_t(z), h_s(z), d_t = d_s = 1\right)' \left[\left(\varepsilon_t - \varepsilon_s\right) - E\left(\varepsilon_t - \varepsilon_s \middle| h_t(z), h_s(z), d_t = d_s = 1\right)\right] d_t d_s$$
$$-\left[\Delta x_{ts} - E\left(\Delta x_{ts} \middle| h_t(z), h_s(z), d_t = d_s = 1\right)\right]' \nabla_t E\left(\varepsilon_t - \varepsilon_s \middle| h_t(z), h_s(z), d_t = d_s = 1\right) d_t d_s ,$$

(II.11)

$$\frac{\partial m}{\partial h_s(z)} = -\nabla_s E\left(\Delta x_{ts} \middle| h_t(z), h_s(z), d_t = d_s = 1\right)' \left[\left(\varepsilon_t - \varepsilon_s\right) - E\left(\varepsilon_t - \varepsilon_s \middle| h_t(z), h_s(z), d_t = d_s = 1\right)\right] d_t d_s$$
$$-\left[\Delta x_{ts} - E\left(\Delta x_{ts} \middle| h_t(z), h_s(z), d_t = d_s = 1\right)\right]' \nabla_s E\left(\varepsilon_t - \varepsilon_s \middle| h_t(z), h_s(z), d_t = d_s = 1\right) d_t d_s ,$$

(II.12)

where $\nabla_t E\left(\cdot \middle| h_t(z), h_s(z), d_t = d_s = 1\right)$ and $\nabla_s E\left(\cdot \middle| h_t(z), h_s(z), d_t = d_s = 1\right)$ are the derivatives of $E\left(\cdot \middle| h_t(z), h_s(z), d_t = d_s = 1\right)$ with respect to $h_t(z)$ and $h_s(z)$, respectively.

For the moment condition in (II.7) a functional expansion around $h_t(z)$, $h_s(z)$, $E\left(\Delta x_{ts} \middle| h_t(z), h_s(z), d_t = d_s = 1\right)$ and $E\left(\varepsilon_t - \varepsilon_s \middle| h_t(z), h_s(z), d_t = d_s = 1\right)$ gives

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m\left[\hat{h}_t(z_i),\ \hat{h}_s(z_i),\ E_N\left(\Delta x_{ts}|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right),\ E_N\left(\varepsilon_t-\varepsilon_s|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right)\right]$$
$$=^{p}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{m\left[h_t(z_i),\ h_s(z_i),\ E_N\left(\Delta x_{ts}|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right),\ E_N\left(\varepsilon_t-\varepsilon_s|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right)\right]\right.$$
$$+\left\{E\left[\frac{\partial m}{\partial E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right)}\middle|h_t(z),h_s(z),d_t=d_s=1\right]\left[\Delta x_{ts}-E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right)\right]\right.$$
$$+E\left[\frac{\partial m}{\partial E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)}\middle|h_t(z),h_s(z),d_t=d_s=1\right]\left[\left(\varepsilon_t-\varepsilon_s\right)-E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right]$$
$$\left.\left.+E\left[\frac{\partial m}{\partial h_t(z)}\middle|h_t(z),h_s(z),d_t=d_s=1\right]\left[d_t-h_t(z)\right]+E\left[\frac{\partial m}{\partial h_s(z)}\middle|h_t(z),h_s(z),d_t=d_s=1\right]\left[d_s-h_s(z)\right]\right\}\right\} \tag{II.13}$$

For our estimator, the two means of $\partial m/\partial E\left(\cdot|h_t(z),h_s(z),d_t=d_s=1\right)$ conditional on $h_t(z),h_s(z)$, and $d_t=d_s=1$ are zero (see (II.9) and (II.10), above). Furthermore, the corresponding two terms for $E\left[\partial m/\partial h_t(z)\middle|h_t(z),h_s(z),d_t=d_s=1\right]$ and $E\left[\partial m/\partial h_s(z)\middle|h_t(z),h_s(z),d_t=d_s=1\right]$, according to (II.11) and (II.12), are also zero because of

$$E\left[\left(\varepsilon_t-\varepsilon_s\right)-E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\middle|h_t(z),h_s(z),d_t=d_s=1\right]=0$$

and

$$E\left[\Delta x_{ts}-E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right)\middle|h_t(z),h_s(z),d_t=d_s=1\right]=0\,.$$

Hence, there is no effect of estimating the four infinite dimensional nuisance

parameters on the asymptotic variance of β given that the correction term in $\{\cdot\}$ in (II.13) is equal to zero. Therefore, we get

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}-\beta\right) &=^{p} A^{-1}\cdot \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N}&\left[\Delta x_{its}-E\left(\Delta x_{ts}\middle|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right)\right]'\left[\left(\varepsilon_{it}-\varepsilon_{is}\right)-E\left(\varepsilon_t-\varepsilon_s\middle|h_t(z_i),h_s(z_t),d_{it}=d_{is}=1\right)\right] \\ &\cdot d_{it}d_{is} \\ &= A^{-1}\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i\end{aligned} \tag{II.14}$$

that is asymptotically normal,

$$\sqrt{N}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,\ A^{-1}E\left(\xi\xi'\right)A^{-1}\right), \tag{II.15}$$

where

$$\xi_i\equiv\left[\Delta x_{its}-E\left(\Delta x_{ts}\middle|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right)\right]'\left[\left(\varepsilon_{it}-\varepsilon_{is}\right)-E\left(\varepsilon_t-\varepsilon_s\middle|h_t(z_i),h_s(z_i),d_{it}=d_{is}=1\right)\right]d_{it}d_{is}. \tag{II.16}$$

A can be estimated as in (II.4), while $E\left(\xi\xi'\right)$ is estimated by replacing all the conditional means involved, that is

$$\left(h_t(z),\ h_s(z),\ E\left(\Delta x_{ts}|h_t(z),h_s(z),d_t=d_s=1\right),\ E\left(\varepsilon_t-\varepsilon_s|h_t(z),h_s(z),d_t=d_s=1\right)\right),$$

with nonparametric estimates.

Chapter 6

Summary and Conclusions

This last chapter summarises the results of the Monte Carlo experiments in chapters 2, 3 and 5, and the main empirical results in chapter 4.

Chapter 2 is concerned about the finite sample performance of Wooldridge (1995) and Kyriazidou's (1997) estimators for "fixed-effects" panel data sample selection models. The results of a small Monte Carlo simulation study show the following. First, Wooldridge's (1995) estimator is less biased than Kyriazidou's (1997) estimator and it reaches faster its asymptotic behaviour. Second, Wooldridge's (1995) suffers from an important misspecification bias problem when the linear projection functional form for the individual effects in the main equation is invalidated. However, breaking down the linearity assumption for the individual effects in the selection equation hardly influences the bias for the second step estimates. In contrast to Wooldridge's (1995) estimator, Kyriazidou's (1997) method is free from misspecification problems affecting the individual effects in both equations. Third, both estimators are not robust to the violation of the underlying strict exogeneity assumption. Finally, Wooldridge's (1995) estimator is robust to violations of the *conditional exchangeability* assumption. When this condition breaks down the main effect on Kyriazidou's (1997) estimator is in terms of precision in the estimates. Furthermore, we get larger finite-sample bias than in Wooldridge's (1995) estimator.

In chapter 3 we introduce a new estimator for panel data sample selection models with "fixed-effects". The estimator relaxes some of the assumptions in the methods in chapter 2. We present two versions of the estimator depending on the treatment of the individual effects in the selection equation. If they are explicitly allowed to depend on the explanatory variables in a linear way (as in Wooldridge (1995)) we have a version of the estimator referred to as the "more parametric new estimator". However, if they are explicitly allowed to depend on the explanatory variables in a fully unrestricted way we call the estimator "less parametric new estimator". 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) estimator. 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 do not seem to affect too badly the proposed estimator.

In chapter 4 to learn about the performance of the methods in chapters 2 and 3 in a practical application, we apply the estimators and their extensions (taking account of non-strict exogeneity and/or time constant non-linear errors in variables) to a typical problem in labour economics: The estimation of wage equations for female workers. The parameter we seek to identify is the effect of actual labour market

experience on wages. The estimator by Kyriazidou (1997) turns out, for our particular application, difficult to apply. It imposes a *conditional exchangeability* assumption, which is rejected by the data. Furthermore, in the case where any non-systematic variation in the variable of interest (experience in our case) coincides with changes in the selection index, this estimator runs into identification problems (between time effects and experience in our case). The results we obtain using Wooldridge's and chapter's 3 estimators indicate that there are correlated fixed effects, and non-random sample selection. Using Wooldgridge's (1995) estimator, we reject specifications, which do not allow for predetermined regressors (and contemporaneous endogeneity). Chapter's 3 method rejects strict exogeneity of the experience variable, conditional on taking care of the measurement error problem by first differencing. The most general estimator using Wooldridge's (1995) method implies an increase in wages by 1.8 percent for one year of labour market experience, evaluated at 14 years of experience. Estimates of chapter's 3 most general estimator (the extension to GMM) are slightly lower. Our results also indicate that estimates of aggregate wage growth are sensitive to the trend in sample selection.

In chapter 5, estimation of the coefficients in a "double-index" selectivity bias model is considered under the assumption that the selection correction function depends only on the conditional means of some observable selection variables. We present two alternative methods. The first is referred to as a "weighted double pairwise difference estimator" (WDPDE) because of being based in the comparison of individuals in time differences. We call the second method a "single pairwise difference estimator" (SPDE) because only differences over time for a given

individual are required. Their advantages are the following. They are distributionally free estimators compared with our earlier estimator in chapter 3 and Wooldridge's (1995) estimator. Furthermore, no *conditional exchangeability* assumption or parametric sample selection index is required compared with Kyriazidou (1997). The methods do not require strict exogeneity for the variables in the main equation, and they are shown to be equivalent under the proper choice of instruments for each estimator. The results of our small Monte Carlo simulation study show the following. Both estimators are less biased than the estimator ignoring correction for sample selection. The bias are all positive, they increase as the kernel order increases and they diminish with sample size. The SPDE performs slightly better than the WDPDE, which can have its origin on the extra differencing present in the latter method.

Bibliography

- AHN, H. AND J. K. POWELL (1993), "Semiparametric estimation of censored selection models with a nonparametric selection mechanism", Journal of Econometrics, 58, 3-29.

- ARELLANO, M. AND R. CARRASCO (1996), "Binary choice panel data models with predetermined variables", CEMFI Working Paper No. 9618.

- BERNDT, E., B. HALL, R. HALL, AND J. HAUSMAN (1974), "Estimation and inference in non-linear structural models", Annals of Economic and Social Measurement, 3/4, 653-665.

- BIERENS, H. J. (1987), " Kernel estimators of regression functions ", in Advances in Econometrics, Fifth World Congress, Volume I, Econometric Society Monographs, No. 13, ED. T. F. BEWLEY, Cambridge University Press.

- BLUNDELL, R. and T. MACURDY (1999), "Labour Supply: A Review of Alternative Approaches", in: O. Ashenfelter and D. Card (eds), Handbook of Labor Economics, Vol. III, North-Holland, Amsterdam.

- BROWNING, M, DEATON, A., and M. IRISH (1985), "A Profitable Approach to Labour Supply and Commodity Demand over the Life Cycle, Econometrica, 53, 503-543.

- CARD, D. (1994), "Earnings, schooling and ability revisited", National Bureau of Economic Research, Working Paper 4832.

- CHAMBERLAIN, G. (1980), " Analysis of covariance with qualitative data ", Review of Economic Studies, XLVII, 225-238.

- CHAMBERLAIN, G. (1982), " Multivariate regression models for panel data ", Journal of Econometrics, 18, 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., B. MELENBERG, AND A. H. O. VAN SOEST (1995), " A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation ", Statistica Nederlandica, 49, 324-342.

- CHARLIER, E., B. MELENBERG, AND A. H. O. VAN SOEST (1997), "An analysis of housing expenditure using semiparametric models and panel data", CentER Discussion Paper, no. 9714, Tilburg University, The Netherlands.

- HAM, J. C. (1982), " Estimation of a labour supply model with censoring due to unemployment and underemployment ", Review of Economic Studies, XLIX, 335-354.

- HÄRDLE, W. (1990), Applied nonparametric regression, Cambridge University Press.

- HÄRDLE, W. AND O. LINTON (1994), " Applied nonparametric methods ", in R. F. ENGLE AND D. L. McFADDEN, EDS., Handbook of Econometrics, Volume IV, Elsevier Science.

- HÄRDLE, W. AND R. CHEN (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 a simple estimates for such models ", Annals of Economics and Social Measurement, 15, 475-492.

- HECKMAN, J. (1979), " Sample selection bias as a specification error ", Econometrica, 47, 153-161.

- HOROWITZ, J. (1992), "A smoothed maximum score estimator for the binary response model", Econometrica, 60, 505-531.

- HOROWITZ, J. (1998), Semiparametric methods in econometrics, lecture notes in statistics-131, Springer.

- HOROWITZ, J. L. (1988), "Semiparametric M-estimation of censored linear regression models", Advances in Econometrics, 7, 45-83.

- HSIAO, C. (1986), Analysis of panel data, Cambridge: Cambridge University Press.

- 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, No. 6, 1335-1364.

- LEE, L. F. (1982), "Some approaches to the correction of selectivity bias ", Review of Economic Studies, XLIX, 355-372.

- LEE, M. J. (1996), Methods of moments and semiparametric econometrics for limited dependent variable models, Springer.

- MaCURDY, T. E. (1981), "An empirical model of labor supply in a life cycle setting", Journal of Political Economy, 89, 1059-1085.

- MANSKI, C. (1975), "Maximum score estimation of the stochastic utility model of choice", Journal of Econometrics, 3, 205-228.

- MANSKI, C. (1985), "Semiparametric analysis of discrete response: asymptotic properties of maximum score estimation", Journal of Econometrics, 27, 313-334.

- MANSKI, C. (1987), " Semiparametric analysis of random effects linear models from binary panel data ", Econometrica, 55, 357-362.

- MARRON, J. S. AND M. P. WAND (1992), " Exact mean integrated squared error", Annals of Statistics, 20, 712-736.

- MOFFITT, R. (1984), "Profiles of fertility, labour supply and wages of married women: a complete life-cycle model", Review of Economic Studies, 51, 263-278.

- MÜLLER, H.G. (1984), "Smooth optimum kernel estimators of densities, regression curves and modes", Annals of Statistics, 12,766-774.

- MUNDLAK, Y. (1978), " On the pooling of time series and cross section data ", Econometrica, 46, 69-85.

- NADARAYA, E. A. (1964), " On estimating regression ", Theory Prob. Appl., 10, 186-190.

- NEWEY, W. K (1994b), " Kernel estimation of partial means and a general variance estimator ", Econometric Theory, 10, 233-253.

- NEWEY, W. K . AND D. McFADDEN (1994c), "Large sample estimation and hypothesis testing ", in R. F. ENGLE AND D. L. McFADDEN, EDS., Handbook of Econometrics, Volume IV, Elsevier Science.

- NEWEY, W. K. (1992), " Partial means, kernel estimation, and a general asymptotic variance estimator ", mimeo, MIT.

- NEWEY, W. K. (1994a), " The asymptotic variance of semiparametric estimators ", Econometrica, Vol. 62, No. 6, 1349-1382.

- NIJMAN, T. AND M. VERBEEK (1992), " Nonresponse in panel data: the impact on estimates of a life cycle consumption function ", Journal of Applied Econometrics, 7, 243-257.

- POIRIER, D. J. (1980), "Partial observability in bivariate probit models ", Journal of Econometrics, 12, 209-217.

- POWELL, J. L. (1987), "Semiparametric estimation of bivariate latent variable models", Working paper number 8704, Revised April 1989 (Social Systems Research Institute, University of Wisconsin, Madison, WI).

- POWELL, J. L. (1994), "Estimation of semiparametric models", Handbook of Econometrics, Vol. 4, 2444-2521.

- ROBINSON, P. M. (1988), "Root-N-consistent semiparametric regression", Econometrica, Vol. 56, No. 4, 931-954.

- ROCHINA-BARRACHINA, M. E. (1999), "A new estimator for panel data sample selection models", Annales d'Économie et de Statistique, 55/56, 153-181.

- SILVERMAN, B. W. (1986), Density estimation for statistics and data analysis, Chapman and Hall.

- SZU, H. and HARTLEY, R. (1987), "Fast simulated annealing", Physics Letters A, vol. 122, #3, 4, 157-162.

- TALLIS, G. M. (1961), " The moment generating function of the truncated multi-normal distribution ", Journal of the Royal Statistical Society, 23, Series b, 223-229.

- VERBEEK, M. (1990), " On the estimation of a fixed effects model with selectivity bias ", Economics Letters, 34, 267-270.

- VERBEEK, M. AND T. NIJMAN (1992), " Testing for selectivity bias in panel data models ", International Economic Review, 33, 681-703.

- WAGNER, G., R. BURKHAUSER AND F. BEHRINGER (1993), "The English language public use file of the German Socio-Economic Panel", Journal of Human Resources, 27, 429-433.

- WATSON, G. S. (1964), " Smooth regression analysis ", Sankhya, Series A, 26, 359-372.

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

- ZABEL, J.E. (1992), " Estimating fixed effects and random effects with selectivity ", Economics Letters, 40, 269-272.