Root-N Consistent Semiparametric Estimators of a Dynamic Panel-Sample-Selection Model

George-Levi Gayle
Tepper School of Business, Carnegie Mellon University

Christelle Viauroux
Department of Economics, University of Cincinnati

First Version: September 2004. Current Version : July 2006

Abstract

This paper considers the problem of identification and estimation in panel-data sample-selection models with a binary selection rule when the latent equations contain possibly predetermined variables, lags of the dependent variables, and unobserved individual effects. The selection equation contains lags of the dependent variables from both the latent and the selection equations as well as other possibly predetermined variables relative to the latent equations. We derive a set of conditional moment restrictions that are then exploited to construct a three-step sieve estimator for the parameters of the main equation including a nonparametric estimator of the sample-selection term. In the second step the unknown parameters of the selection equation are consistently estimated using a transformation approach in the spirit of Berkson's minimum chi-square sieve method and a first-step kernel estimator for the selection probability. This second-step estimator is of interest in its own right. It can be used to semiparametrically estimate a panel-data binary response model with a nonparametric individual specific effect without making any other distributional assumptions. We show that both estimators (second and third stage) are $\sqrt{n}$-consistent and asymptotically normal.

1 Introduction

In this paper, we study a panel-data sample-selection model of the form

$$y^*_{it} = \rho y^*_{it-1} + x_{it}\beta + \alpha_i + \varepsilon^*_{it}, \tag{1}$$

$$y_{it} = d_{it} y^*_{it}, \tag{2}$$

and

$$d_{it} = \mathbb{I}\left\{\phi_0 d_{it-1} + \phi_1 y_{it-1} + z_{it}\gamma + \eta_i - u_{it} > 0\right\}, \tag{3}$$

where $\mathbb{I}\{.\}$ denotes the usual indicator function, i indexes individuals ($i = 1, \ldots, n$), t indexes time ($t = 1, \ldots, T$), y^*_{it} is a latent outcome variable, $(y_{it}, d_{it}, x_{it}, z_{it})$ are observed random variables, (α_i, η_i) are unobserved time-invariant individual-specific effects, and $(\varepsilon^*_{it}, u_{it})$ are unobserved random individual time-specific effects assumed to be independent across individuals. The first equation, also called the outcome equation, has an autoregressive structure with x_{it} containing both strictly exogenous variables x^e_{it} and predetermined variables x^p_{it} with respect to ε^*_{it}. The second equation, the selection equation, summarizes the process of observations entering into the sample. It has an autoregressive structure, but also depends explicitly on the lagged outcome of the first equation and on variables z_{it} that may be predetermined with respect to ε^*_{it} and u_{it}.

Although sample-selection models have been studied extensively in the econometrics literature (see Heckman, 1974, 1976, Das, Newey and Vella, 2003, Kyriazidou, 1997, 2001, among others), the model described in (1)–(3) is new in two respects. First, the selection process may depend on variables that are predetermined with respect to the error structure of both the outcome and the selection equations. Second, the outcome equation contains a lagged dependent variable along with other predetermined variables. Therefore our model can be derived from a dynamic utility maximization problem with time-inseparable preferences.

The economic literature contains many examples where panel-sample-selection models that include lags or other predetermined variables would be of interest. Models of this form arise for example in the study of the intertemporal behavior of economic agents (see Hotz, Kydland and Sedlacek, 1988, Altug and Miller, 1998, to name a few). Several applications have shown that the current realization of outcomes or the current decision to participate in the sample is affected by both current and past variables (see Kydland and Prescott, 1982, Altug and Miller, 1998, among others). One illustration is the analysis of a company's investment behavior (Bond and Meghir, 1994). In this case it is reasonable to expect that whether a company decides to invest today will depend on how much it invested in the last period, hence creating a feedback from the continuous outcome (the amount invested last period) to the selection process (the decision to invest today). Variables explaining investment include variables in the agent's information sets that would be correlated with past shocks and hence past values of the dependent variable. These variables are therefore predetermined with respect to the system. Another example of interest is the study of models of life-cycle behavior (Gayle and Miller, 2004) where the dynamic utility maximization of time nonseparable preferences gives rise to feedback effects from the outcome equation to the selection equation. For example, the

decision to participate in the labor force may depend on the wage earned in the past. As the past explains the present, so does the expectation about the future. For example, parents' decision to work in the current period is affected by their expectations about future birth events. Therefore, explanatory variables such as the number of children are predetermined (that is, correlated with lagged values of the error term of the outcome equation, but uncorrelated with its present and future values).[1]

Panel-data dynamic models of sample selection have been studied by Kyriazidou (2001). Her model, however differs from ours in that the selection equation does not have any autoregressive structure, nor does it depend on the lag of the dependent variable from the outcome equation. Consequently, it cannot be directly derived from a dynamic utility maximization problem. Nevertheless, the extensions we consider in this paper are not trivial since Kyriazidou's (2001) estimation and identification strategy critically depends on the assumption that the selection equation only contains strictly exogenous regressors. Hence, we will have to pursue a different identification and estimation strategy from Kyriazidou (2001).

The method that we adopt in this paper allows us to identify the structural parameters of the outcome and of the selection equation without placing any parametric assumption on the distribution of the error terms. We will still be able to obtain a $\sqrt{n}$-consistent estimator for the structural parameters that is asymptotically normally distributed. The $\sqrt{n}$- consistency of our estimator is an important improvement over that of Kyriazidou (2001). However this generalization and improvement comes at a cost of two additional restrictions both of which enable us to obtain identification of our model. The first major restriction is to impose the individual specific effect in the selection equation to be a nonparametric function of strictly exogenous individual specific variables. The second restriction requires that the distribution of the predetermined and lagged dependent variables in the selection equation conditional on the instruments in the outcome equation forms a complete family of distribution.

The contribution of this paper is twofold. First it develops a new semiparametric estimator for a more general class of panel sample selection models than is found in the current literature. Second, it develops a new semiparametric estimator for dynamic panel binary choice model. This estimation technique does not rely on specifying the distribution of the errors nor does it rely on the distribution of the initial condition.

The remainder of this paper is organized as follows. In Section 2, the identification of the sample-selection model is discussed. In Section 3, our proposed estimator is defined. Consistency and asymptotic normality of our estimator are established in Section 4. A Monte Carlo study of our estimator is conducted in Section 5 to compare the small-sample performance of our estimator with that of Kyriazidou (2001). An appendix

[1] A predetermined variable is a regressor that is Granger caused by past values of the dependent variable of the system.

contains proofs of the technical results while proofs that are critical to the flow of the paper are left in the text.

2 Identification

Five major issues need to be addressed in order to achieve identification of the model specified in equations (1)–(3). These issues include the presence in equation (1) of the unobserved time-invariant–individual-specific component, α_i, the presence of the lagged dependent variable, y^*_{it-1}, the presence of predetermined variables in x_{it}, the sample-selection mechanism specified by Equations (2) and (3) and the identification of the binary selection model with lagged dependent variables. In the absence of sample selection, the first three issues have been resolved in the panel-data literature (see Anderson and Hsiao, 1982, Arellano and Bond, 1991, Ahn and Schmidt, 1995, to name a few). The presence of sample selection, however, complicates the identification issue. Another issue is the identification of the selection equation as it contains lagged dependent variables and an individual specific effect.

A number of papers have looked at this problem before (Honore, 1993; Arellano and Carrasco, 2003, Hahn, 1997 and Honore and Kyriazidou, 2000). However, these papers have either studied the case where the distribution of the error term is assumed to be parametric or the case where the distribution of the initial condition is parametrically specified. Moreover, they take one of the following approaches for the individual specific effect: either they assume that it is random with a parametric distribution or that it is fixed and can be eliminated in some ways. In this paper we do not assume that the distribution of the error is parametric. We take an alternative approach on the individual specific effects. The following assumptions summarize that approach.

Assumption 2.1 $(\forall i = 1, ..., n,\ t = 1, ..., T)$**:**

1. $\eta_i = \eta(z_i^1)$ *where* z_i^1 *are the strictly exogenous time-invariant components of* z_{it}, $\eta(z_i^1)$ *is an unknown function of* z_i^1.

2. $(\varepsilon^*_{it},\ u_{it})$ *are jointly independent of* z_i^1 and z_{it}.

This restriction is still mild since the distribution of the error terms $(\varepsilon^*_{it},\ u_{it})$ is left unspecified. This is a nonparametric extension of the approach in MaCurdy (1981), which was formalized in Altug and Miller (1998). A version of this restriction can be found in Chamberlain (1986), Nijman and Verbeek (1992), and Zabel (1992), where $\eta(.)$ is specified as a linear function of strictly exogenous variables and the distribution of the error terms are assumed to be normal. Newey (1994) relaxed the linear functional assumption on $\eta(.)$ while retaining the normality assumption on the errors. In what

follows, we use an approach similar to Chen (1998) by relaxing the assumptions on both the functional form and the distribution of the error term. One viable alternative would be to use the fixed-effect specification of Honore and Lewbel (2002). Their fixed-effect model relies, however, on the existence of a special regressor conditionally independent of the individual-specific effect and of the error term in the model, whereas our original formulation does not. We now state additional conditions under which the selection equation is semiparametrically identified. Let

$$P_{it0} \equiv E[d_{it}|d_{it-1}, y_{it-1}, z_{it}, z_i^1],$$

be the true conditional expectation. Let $w_{it} = [d_{it-1}, y_{it-1}, z_{it}']', \omega_{it} = [w_{it}', z_i^{1\prime}]$ and $\widetilde{\gamma} = [\phi_0, \phi_1, \gamma']'$.

Assumption 2.2 ($\forall i = 1, ..., n$, $t = 1, ..., T$)**:**

1. $\|\widetilde{\gamma}\| = 1$.
2. T*he random vector* w_{it} *contains at least one continuous regressor.*
3. $E[\triangle w_{it} \triangle w_{it}']$ *is invertible.*
4. $E[\eta(z_i^1)] = 0$.
5. Let $F_u(.)$ be the distribution function of u_{it}, conditional on ω_{it}. $F_u(.)$ *is strictly increasing, differentiable and non constant on its support.*

Assumptions 2.2.1 and 2.2.2 are standard in the semiparametric literature on the identification of binary choice models (See Manski, 1987 among others). A well known alternative to Assumption 2.2.2 is that a component of w_{it}', say w_{itk} (of associated parameter $\widetilde{\gamma}_k$) has a probability distribution conditional on the remaining components that are absolutely continuous with respect to the Lebesgue measure, and then assume that $|\widetilde{\gamma}_k| = 1$. We do not need to make this assumption here since under Assumption 2.2.5 we are able to estimate the signs of all coefficients. We could however assume that $\widetilde{\gamma}_k = 1$ (or $\widetilde{\gamma}_k = -1$) which along with Assumption 2.2.2 would allow us to estimate the remaining parameters relative to $\widetilde{\gamma}_k$. Assumption 2.2.3 is the traditional full rank condition used for identification in the linear panel data literature. Assumption 2.2.4 is a version of the traditional zero mean assumption in fixed effect models, it serves here as the location normalization. It could be relaxed but then all the nonparametric functions would be identified up to an additive constant. Assumption 2.2.5 is the critical assumption which allows us to identify and estimate our model. Since the identification strategy is very important in understanding the estimation, we will prove it in the text.

Proposition 1 *Under Assumptions 2.1 and 2.2* $(\widetilde{\gamma}_0, F_{0u}(.), \eta_0(z_i^1))$ *are identified.*

Proof. Let $\pi_0 = (\widetilde{\gamma}_0, F_{0u}(.), \eta_0(z_i^1))$ denote the true parameters of our model and let $\pi_1 = (\widetilde{\gamma}_1, F_{1u}(.), \eta_1(z_i^1))$ be another set of parameters that are observationally equivalent to π_0 . Then by equation (3) and Assumption 2.1 we obtain that

$$F_{0u}(w_{it}'\widetilde{\gamma}_0 + \eta_0(z_i^1)) = F_{1u}(w_{it}'\widetilde{\gamma}_1 + \eta_1(z_i^1)), \tag{4}$$

because by observational equivalence both sides are equal to P_{it0}. Assumption 2.2.5 implies that

$$w_{it}'\widetilde{\gamma}_0 + \eta_0(z_i^1) = F_{0u}^{-1}(F_{1u}(w_{it}'\widetilde{\gamma}_1 + \eta_1(z_i^1))). \tag{5}$$

Since $F_u(.)$ is strictly increasing it is differentiable almost everywhere. Differentiating (5) with respect to the continuous regressor w_{itk} gives

$$\begin{aligned} \widetilde{\gamma}_{0k} &= \frac{\partial_{w_{itk}} F_{0u}^{-1}(F_{1u}(w_{it}'\widetilde{\gamma}_1 + \eta_1(z_i^1)))}{\partial_{w_{itk}} F_{1u}^{-1}(F_{1u}(w_{it}'\widetilde{\gamma}_1 + \eta_1(z_i^1)))}\widetilde{\gamma}_{1k}, \\ &= \frac{\partial_{w_{itk}} F_{0u}^{-1}(P_{it0})}{\partial_{w_{itk}} F_{1u}^{-1}(P_{it0})}\widetilde{\gamma}_{1k}. \end{aligned} \tag{6}$$

Let $\xi_k = \frac{\widetilde{\gamma}_{0k}}{\widetilde{\gamma}_{1k}}$, note that $\xi_k > 0$ which follows directly from Assumption 2.2.5. Hence we can rewrite equation (6) as

$$\partial_{w_{itk}} F_{0u}^{-1}(P_{it0}) = \xi_k \partial_{w_{itk}} F_{1u}^{-1}(P_{it0}). \tag{7}$$

Integrating equation (7) over the range $[0, P_{it0}]$ gives us

$$F_{0u}^{-1}(P_{it0}) = \xi_k F_{1u}^{-1}(P_{it0}) + K, \tag{8}$$

where $K = F_{0u}^{-1}(0) - \xi_k F_{1u}^{-1}(0)$. Note that by equation (3), Assumptions 2.1 and 2.2.5, we obtain

$$F_{0u}^{-1}(P_{it0}) = w_{it}'\widetilde{\gamma}_0 + \eta_0(z_i^1), \tag{9}$$

$$F_{1u}^{-1}(P_{it0}) = w_{it}'\widetilde{\gamma}_1 + \eta_1(z_i^1). \tag{10}$$

Taking the first difference of equations (8), (9) and (10) gives us

$$\triangle F_{0u}^{-1}(P_{it0}) = \xi_k \triangle F_{1u}^{-1}(P_{it0}), \tag{11}$$

$$\triangle F_{0u}^{-1}(P_{it0}) = \triangle w_{it}'\widetilde{\gamma}_0, \tag{12}$$

$$\triangle F_{1u}^{-1}(P_{it0}) = \triangle w_{it}'\widetilde{\gamma}_1. \tag{13}$$

From equations (11), (12) and (13) we obtain that

$$\triangle w_{it}'\widetilde{\gamma}_0 = \xi_k \triangle w_{it}'\widetilde{\gamma}_1. \tag{14}$$

Premultiplying equation (14) by $\triangle w_{it}$ and taking the expectation gives

$$E[\triangle w_{it}\triangle w_{it}']\widetilde{\gamma}_0 = \xi_k E[\triangle w_{it}\triangle w_{it}']\widetilde{\gamma}_1. \tag{15}$$

Assumption 2.2.3 and equation (15) imply that

$$\widetilde{\gamma}_0 = \xi_k \widetilde{\gamma}_1. \tag{16}$$

Assumption 2.2.1 that $\|\widetilde{\gamma}_0\| = \|\widetilde{\gamma}_1\| = 1$ implies that $|\xi_k| = 1$. But since $\xi_k > 0$, this means that $\xi_k = 1$, thus

$$\widetilde{\gamma}_0 = \widetilde{\gamma}_1, \tag{17}$$

$$F_{0u}^{-1}(P_{it0}) = F_{1u}^{-1}(P_{it0}) + K. \tag{18}$$

Equations (9), (10) and (18) imply that

$$w_{it}'\widetilde{\gamma}_0 + \eta_0(z_i^1) = w_{it}'\widetilde{\gamma}_1 + \eta_1(z_i^1) + K. \tag{19}$$

Equation (17) further gives

$$\eta_0(z_i^1) = \eta_1(z_i^1) + K. \tag{20}$$

Finally, Assumption 2.2.4 implies that $K = 0$, hence the model is identified. ■

The literature on sample selection normally takes two different approaches to the identification of the structural parameters, the first is the standard Heckman' s correction and the second is to find a way to eliminate the selection bias indirectly (see Ahn and Powell, 1993 for example). Under the first approach one could either assume the parametric form of the joint error distribution and then obtain the correction term or one could nonparametrically identify the correction term. We will use the nonparametric approach in this paper. Below we state some regularity conditions that allow us to proceed.

Assumption 2.3:

1. ε_{it}^* *is independent of* y_{i0}^* *for all* t *and for each* i.
2. ε_{it}^* *is independent of* α_i *for all* t *and for each* i.
3. *The* $(\varepsilon_{it}^*, u_{it})'s$ *are mutually independent for all* t *and for each* i, *with* $E(\varepsilon_{it}^*) = 0$ *and* $var(\varepsilon_{it}^*) = \sigma_\varepsilon^2$.
4. x_{is}^p *is independent of* ε_{it}^* *for all* $s \leq t$ *and for each* i.

5. x_{is}^e *is independent of* ε_{it}^* *for all* $s, t = 1, \ldots, T$ *and for each* i.

6. z_{is} *is independent of* ε_{it}^* *for all* $s \leq t$ *and for each* i.

7. y_{i0}^* *is i.i.d. with density* $f_{y_0}(.)$ *for each* i.

8. $d_{i0} \in \{0, 1\}$ *i.i.d. with* $\Pr[d_{i0} = 1] = P_o$ *for each* i.

Assumption 2.3 is a strengthening of the standard type of assumptions used in the linear panel-data literature to achieve identification. In the standard linear panel-data literature, these are typically conditional mean independence assumptions, while here we need full conditional independence. This is not unusual in the semiparametric literature, however, and is similar to the assumptions used in Kyriazidou (2001).

Let

$$\zeta_{it} \equiv \{y_{it-1}, y_{it-2}, d_{it-2}, x_{it}, z_{it}, z_{it-1}, z_i^1, \alpha_i, d_{it}d_{it-1} = 1\}.$$

Taking the expectation of y_{it}^* in Equation (1) conditional on ζ_{it} gives for $i = 1, ..., n$, $t = 2, ..., T$,

$$E(y_{it}^* \mid \zeta_{it}) = \rho y_{it-1} + x_{it}\beta + \alpha_i + \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1), \tag{21}$$

where

$$\overline{v}_{it} = \phi_0 d_{it-1} + \phi_1 y_{it-1} + z_{it}\gamma \tag{22}$$

and

$$\overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \equiv E(\varepsilon_{it}^* \mid \zeta_{it}). \tag{23}$$

Equation (21) follows by noting that

$$E(\varepsilon_{it}^* \mid \zeta_{it}) = E\{E(\varepsilon_{it}^* \mid \zeta_{it}, \eta_i) \mid \zeta_{it}\}. \tag{24}$$

The inner expectation $E(\varepsilon_{it}^* \mid \zeta_{it}, \eta_i)$ can be expressed as

$$\begin{aligned} E(\varepsilon_{it}^* \mid \zeta_{it}, \eta_i) &= E[\varepsilon_{it}^* \mid y_{it-1}, y_{it-2}, d_{it-2}, x_{it}, z_{it}, z_{it-1}, z_i^1, \alpha_i, \eta_i, u_{it} < \overline{v}_{it} + \eta_i, u_{it-1} < \overline{v}_{it-1} + \eta_i], \\ &= \frac{\int_{-\infty}^{\overline{v}_{it}+\eta_i} \int_{-\infty}^{\overline{v}_{it-1}+\eta_i} \int_{-\infty}^{+\infty} \varepsilon^* f(\varepsilon^*, u_2, u_1 \mid y_{it-1}, y_{it-2}, d_{it-2}, x_{it}, z_{it}, z_{it-1}, z_i^1, \alpha_i, \eta_i) d\varepsilon^* du_2 du_1}{\int_{-\infty}^{\overline{v}_{it}+\eta_i} \int_{-\infty}^{\overline{v}_{it-1}+\eta_i} \int_{-\infty}^{+\infty} f(\varepsilon^*, u_2, u_1 \mid y_{it-1}, y_{it-2}, d_{it-2}, x_{it}, z_{it}, z_{it-1}, z_i^1, \alpha_i, \eta_i) d\varepsilon^* du_2 du_1}. \end{aligned}$$

Furthermore, by Assumptions 2.1.2, 2.3.2, 2.3.4, and 2.3.5 to 2.3.8, the conditional density $f(\varepsilon^*, u_2, u_1 \mid y_{it-1}, y_{it-2}, d_{it-2}, x_{it}, z_{it}, z_{it-1}, z_i^1, \alpha_i, \eta_i)$ equals the joint density $f(\varepsilon^*, u_2, u_1)$, which implies that

$$\begin{aligned} E(\varepsilon_{it}^* \quad | \quad & \zeta_{it}, \eta_i) \\ = \quad & \frac{\int_{-\infty}^{\overline{v}_{it}+\eta_i} \int_{-\infty}^{\overline{v}_{it-1}+\eta_i} \int_{-\infty}^{+\infty} \varepsilon^* f(\varepsilon^*, u_2, u_1) d\varepsilon^* du_2 du_1}{\int_{-\infty}^{\overline{v}_{it}+\eta_i} \int_{-\infty}^{\overline{v}_{it-1}+\eta_i} \int_{-\infty}^{+\infty} f(\varepsilon^*, u_2, u_1) d\varepsilon^* du_2 du_1}, \\ \equiv \quad & \lambda(\overline{v}_{it} + \eta_i, \overline{v}_{it-1} + \eta_i). \end{aligned} \tag{25}$$

Denote by $f_\eta(.)$ the density of η_i, then we can integrate out the unobserved component η_i over $f_\eta(.)$ and

$$\begin{aligned} E(\varepsilon_{it}^* \quad | \quad & \zeta_{it}) \\ = \quad & \int \lambda(u + \overline{v}_{it}, u + \overline{v}_{it-1}) f_\eta(u) du, \\ \equiv \quad & \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1). \end{aligned} \tag{26}$$

Note that if (23) were equal to zero, then (21) could be estimated as a standard dynamic, linear panel-data model. However, (23) in general is not equal to zero and the model to be estimated is of the form

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \alpha_i + \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) + \varepsilon_{it}, \tag{27}$$

where

$$\varepsilon_{it} = \varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1). \tag{28}$$

The above correction is now similar to that considered by Heckman (1976) except that it is a multi-index specification instead of a single-index specification. In this paper, we allow $\overline{\lambda}(.)$ to have an unknown functional form, which is similar to the formulation used by Das, Newey and Vella (2003) in the cross-section contexts. Equation (27) is an additive semiparametric regression equation similar to that considered by Robinson (1988), except that $\overline{\lambda}(.)$ depends on the unknown parameters of the selection equation. Suppose for the moment that these parameters were known; then, in order to obtain consistent estimates of the remaining parameters in (27), we would need to correct for the presence of y_{it-1} and of the predetermined variables in x_{it}. Following Arellano and Bover (1995), we would look within the system for instruments that will lead to consistent estimates of the parameters of interest.

Unlike in the linear case, we will put restrictions on the latent variables and derive orthogonality and identification conditions implied on observed variables. These conditions are summarized in the proposition below.

Proposition 2 *For $i = 1, ..., n$, $t = 4, ..., T$, let*
$F_{it} \equiv \left\{ y_{i0}, \ldots, y_{it-3}, d_{i0}, \ldots, d_{it-3}, x^p_{i1}, \ldots, x^p_{it-2}, x^e_{i1}, \ldots, x^e_{iT}, z_{i1}, \ldots, z_{it-2}, z^1_i, d_{it}d_{it-1} = 1 \right\}$.
Under Assumptions 2.1 and 2.3, the following holds:

$$E[(\triangle y_{it} - \rho_0 \triangle y_{it-1} - \triangle x_{it}\beta_0 - \triangle \overline{\lambda}_0(\overline{v}_{it}, \overline{v}_{it-1}, \overline{v}_{it-2}, z^1_i) \mid F_{it}] = 0.$$

Identification of the parameters $(\rho_0, \beta_0, \triangle\overline{\lambda}_0)$ conditional on (ϕ_0, ϕ_1, γ) is not as straightforward as in the standard linear model, because $\triangle y_{it-1}$ and $\overline{v}_{it}$ are endogenous with respect to $\triangle\varepsilon_{it}$. This endogeneity arises as both $\triangle y_{it-1}$ and $\overline{v}_{it}$ are functions of y^*_{it-1}. Moreover, y^*_{it-1} is a function of ε^*_{it-1}, of which $\triangle\varepsilon_{it}$ is a function. Also in the presence of predetermined variables, x^p_{it} and z^p_{it} may be functions of ε^*_{it-1}. If $\overline{v}_{it}$ was not a function y^*_{it-1} and hence $\triangle y_{it-1}$ and $\triangle x^p_{it}$ were the only endogenous variables in the model, we could use the standard instrument variable conditions for identification. However, since $\triangle\overline{\lambda}_0$ is a nonparametric function of endogenous variables it becomes a problem of semiparametric identification with endogeneity (see Darolles, Florens and Renault, 2002 and Newey and Powell, 2003 for a discussion of the problem). In order to obtain identificaton we will impose the following conditions:

Assumption 2.4 ($\forall\ i = 1, ..., n,\ t = 3, ..., T$)**:**

1. *The distribution of* $(\triangle y_{it-1}, x^p_{it}, \overline{v}_{it})$ *conditional on*

$$\widetilde{F}_{it} \equiv \left\{ \begin{array}{c} y_{i0}, \ldots, y_{it-3}, x^p_{i1}, \ldots, x^p_{it-2}, x^e_{i1}, \ldots, x^e_{iT}, z^p_{i1}, \ldots, z^p_{it-2}, \\ x^p_{i1}, \ldots, x^p_{iT}, z^p_{i1}, \ldots, z^p_{iT}, z^1_i, d_{i0} \times \ldots \times d_{it} = 1 \end{array} \right\}$$

   forms a complete family of distribution in the sense of Newey and Powell (2003).

2. $\triangle\overline{\lambda}_0(.) \in \Lambda^{p_2}_{c_2}$ *with* $p_2 > 1$*;* $E[\triangle\overline{\lambda}_0(\overline{v}_{it}, \overline{v}_{it-1}, \overline{v}_{it-2}, z^1_i)|\widetilde{F}_{it}] \notin$ *linear span* $(\triangle y_{it-1}, \triangle x_{it})$ *and* $E[\widetilde{F}_{it}\widetilde{F}'_{it}]$ *is finite positive definite.*[2]

Proposition 3 *Under Assumptions 2.1, 2.3, 2.4 and for* $\overline{v}_{it}$, $\overline{v}_{it-1}$ *and* $\overline{v}_{it-2}$ *known,* $(\rho_0, \beta_0, \overline{\lambda}_0(.))$ *is identified.*

Assumption 2.4.1 states that the conditional distribution of the endogenous variables in the model is complete conditional on the instruments. This assumption places restrictions on the joint distribution of the errors terms and the predetermined variables.

[2] Λ^p_c is a Hölder ball, this controls the smoothness of the functional space to which $\triangle\overline{\lambda}_0$ may belong. We will formally define a Hölder ball in the next section .

Consider the simple case where x_{it} and z_{it} contains only strictly exogenous variables with respect both (ε^*_{it}, u_{it}) so that the only predeterminedness comes from y^*_{it-1}. Let $f_{\varepsilon u}(y^*_{it-1}, u)$ be the density of (ε^*_{it}, u_{it}). Then a sufficient condition for Assumption 2.4.1 to hold is that the cumulative distribution function associated with the density

$$f(y^*_{it-1}|\widetilde{F}_{it}) \equiv \frac{\int_{-\infty}^{\phi_0+\phi_1 y^*_{it-1}+\gamma z_{it}+\eta_i} f_{\varepsilon u}(y^*_{it-1}, u)du}{\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{\phi_0+\phi_1 y^*_{it-1}+\gamma z_{it}+\eta_i} f_{\varepsilon u}(y^*_{it-1}, u)du\right] dy^*_{it-1}}, \tag{29}$$

forms a complete family of distribution. Assumption 2.4.2 is the same as those imposed by Cosslett (1991) and Newey (1999) in the cross-section sample-selection context and Robinson (1988) in the selection version of the additive semiparametric regression. The proof of Proposition 3 follows from checking appropriate conditions of Theorem 4.3 of Newey and Powell (2003). Note that we could relax the completeness assumption to a bounded completeness assumption. This would restrict the nonparametric function to be bounded and at the same time would broaden the class of distributions (see Blundell, Chen and Kristensen, 2004 for details).

3 Estimation

We consider a three-step estimator, where the first step is a nonparametric estimator of the individual conditional probability of being observed in the sample each period, the second and third step are semiparametric minimum-distance estimators. In a sense, the estimators are analogous to Heckman's (1976) two-step procedure for the cross-sectional Gaussian disturbance model. The difference is that the selection equation is estimated by a distribution-free method that depends on a preliminary nonparametric estimator rather than by a Probit, and a nonparametric approximation of the selection correction function, $\triangle\overline{\lambda}_0(.)$, is used instead of the inverse Mills ratio. In this regard, our estimator is similar to Newey's (1999) two-step procedure for the cross-sectional case, except that we have a preliminary nonparametric estimate of the conditional selection probability. Our estimator is closest to Chen's (1998) three-step procedure for the static panel-sample-selection model. However, he did not introduce lagged-dependent and other predetermined variables, he used a least-square regression in the second and third steps, whereas we have conditional moment restrictions, which lead to different estimators.

We will use sieve-extremum estimation methods to estimate our model. There is a growing literature on this topic with important theoretical contributions by Shen and

Wong (1994), Shen (1997), Chen and Shen (1998), Chen, Linton, Van Keilegom (2003), Ai and Chen (2003), among many others. See also Chen (2005) for an extensive survey on the sieve estimation literature. The most important consideration in choosing a sieve's space for an approximation is how well it approximates a given class of functions. Restricting our attention to functions that belong to a Hölder space, we first introduce a measure of the approximation error that will play an important role in the large-sample properties of our estimator.

Suppose that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$ is the Cartesian product of compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_d$. Let $0 < \tau \leq 1$. A real-valued function $\hbar$ on $\mathcal{X}$ is said to satisfy a Hölder condition with exponent τ if there is a positive number c such that $|\hbar(x) - \hbar(y)| \leq c \|x - y\|_E^\tau$ for all $x, y \in \mathcal{X}$, where $\|x\|_E = \left(\sum_{l=1}^d x_l^2\right)^{\frac{1}{2}}$ is the Euclidean norm of $x = (x_1, \ldots, x_d) \in \mathcal{X}$. Given a d-tuple $\delta = (\delta_1, \ldots, \delta_d)$ of nonnegative integers, set $[\delta] = \delta_1 + \cdots + \delta_d$ and let $\mathcal{D}^\delta$ denote the differential operator defined by

$$\mathcal{D}^\delta = \frac{\partial^{[\delta]}}{\partial x_1^{\delta_1} \ldots \partial x_d^{\delta_d}}. \tag{30}$$

Let φ be a nonnegative integer and set $p = \varphi + \tau$. A real-valued function h on $\mathcal{X}$ is said to be p-smooth if it is φ times continuously differentiable on $\mathcal{X}$ and $\mathcal{D}^\delta$ satisfies a Hölder condition with exponent τ for all δ with $[\delta] = \varphi$. More generally, a real valued function $h(x, \theta)$ is said to be Hölder continuous in $\theta \in \Theta$ if there exists a constant $\kappa \in (0, 1]$ and a measurable function $c(x)$ with $\mathrm{E}(c(x)^2)$ bounded and such that $|h(x, \theta_1) - h(x, \theta_2)| \leq c(x) \|\theta_1 - \theta_2\|_s^\kappa$ for all $x \in \mathcal{X}$, θ_1, $\theta_2 \in \Theta$ where $\|.\|_s$ is a norm such as the sup or the L_2-norm. Denote the class of all p-smooth real-valued functions on $\mathcal{X}$ by $\Lambda^p(\mathcal{X})$ and the space of all φ-fold continuously differentiable real-valued functions on $\mathcal{X}$ by $C^\varphi(\mathcal{X})$. Define a Hölder ball with smoothness $p = \varphi + \tau$ as

$$\Lambda_c^p(\mathcal{X}) = \left\{ \hbar \in C^\varphi(\mathcal{X}) : \sup_{x \in \mathcal{X}.} |\hbar(x)| \leq c, \sup_{[\delta]=\varphi} \sup_{x,y \in \mathcal{X},\ x \neq y} \frac{\left|\mathcal{D}^\delta \hbar(x) - \mathcal{D}^\delta \hbar(y)\right|}{\|x - y\|_E^\tau} \leq c \right\}. \tag{31}$$

We restrict all our nonparametric functions to belong to a Hölder ball because these functions are well approximated by linear sieves,[3] which we choose in this paper.

Let us denote by θ a generic real-valued function with bounded domain $\mathcal{X} \subset \mathcal{R}^d$, let $\|\theta\|_\infty \equiv \sup_{x \in \mathcal{X}} |\theta(x)|$ be the L_∞ norm, and $\|\theta\|_{2,leb} \equiv \left\{\int_{\mathcal{X}} [\theta(x)]^2 dx / vol(\mathcal{X})\right\}^{\frac{1}{2}}$ be the scaled L_2 norm relative to the Lebesgue measure on $\mathcal{X}$. The sieve approximation errors to $\theta_0 \in \Lambda_c^p(\mathcal{X})$ in $L_\infty(\mathcal{X}, leb)$-norm and $L_2(\mathcal{X}, leb)$-norm are defined as

$$e_{\infty n} \equiv \inf_{\hbar \in \Theta_n} \|\hbar - \theta_0\|_\infty \text{ and } e_{2n} \equiv \inf_{\hbar \in \Theta_n} \|\hbar - \theta_0\|_{2,leb}.$$

[3] A sieve is called a linear sieve if it is a linear span of finitely many known basis functions.

3.1 Selection Equation Estimation

Let $\omega_{it} \equiv (d_{it-1}, y_{it-1}, z'_{it}, z_i^1)'$ be the vector of observed variables that affect the probability of selection. Let (ϕ_{00}, γ_0) denote the true value of (ϕ_0, γ). In order to obtain a $\sqrt{n}$-consistent estimator of the finite-dimensional parameters in the outcome equation, a sufficient condition is to require that any estimator of $(\phi_0, \phi_1, \gamma)'$ be asymptotically equivalent to the sample average that depends only on ω_{it}. In particular, let there exist a function $\Psi(\omega)$ such that $\sqrt{n}((\widehat{\phi}_0, \widehat{\phi}_1, \widehat{\gamma})'\text{-}(\phi_{00}, \phi_{10}, \gamma_0)') = \sum_{i=1}^n \Psi(\omega_i)/\sqrt{n} + o_p(1)$, where $\omega_i = (\omega_{i1}, \ldots, \omega_{iT})'$, $E(\Psi(\omega_i)) = 0$ then $E(\Psi(\omega_i)\Psi(\omega_i)')$ exists and is nonsingular. This is the same requirement as in Newey (1999) in the cross-section context. However, while in the cross-section context there are a number of distribution-free estimators that have this property, in the context of our specific selection equation, there are none to the best of our knowledge. The closest distribution-free estimator that could be used in our context is in Honore and Lewbel (2002), but as pointed out in Section 2 above, this would require us to fundamentally change our identification assumptions. What we will do instead is to develop an example of how one such an estimator can be constructed.

The estimator of the selection parameters $(\phi_0, \phi_1, \gamma)'$ is derived from the relationship between the conditional probability of selection and the selection parameters themselves over the full sample of observations. Conditional on ω_{it}, the probability of selection is $P(\omega_{it}) \equiv E[\mathrm{d}_{it} \mid \omega_{it}]$. Equation (3) implies that $\forall i = 1, ..., n$, $t = 1, ..., T$,

$$P_0(\omega_{it}) = \Pr\left[u_{it} < \phi_0 d_{it-1} + \phi_1 y_{it-1} + z_{it}\gamma + \eta(z_i^1) \mid \omega_{it}\right]. \tag{32}$$

Therefore, an alternative representation of the conditional probability of selection is

$$P_0(\omega_{it}) = F_u(\phi_0 d_{it-1} + \phi_1 y_{it-1} + z_{it}\gamma + \eta(z_i^1)). \tag{33}$$

Under Assumption 2.2.5, $F_u(.)$ is a strictly monotone increasing function, therefore its inverse exists and we obtain the following relationship between the conditional selection probability $P(\omega_{it})$ and the parameters in the selection equation $(\phi_0, \phi_1, \gamma)'$:

$$F_u^{-1}(P_0(\omega_{it})) = \phi_0 d_{it-1} + \phi_1 y_{it-1} + z_{it}\gamma + \eta(z_i^1). \tag{34}$$

Suppose there exists a consistent estimator of $P_0(\omega_{it})$, say $\widehat{P}(\omega_{it})$. By taking a mean value expansion of $F_u^{-1}(\widehat{P}(\omega_{it}))$ around the true selection probability, $P_0(\omega_{it})$, we obtain

$$F_u^{-1}(\widehat{P}(\omega_{it})) = \phi_0 d_{it-1} + \phi_1 y_{it-1} + z_{it}\gamma + \eta(z_i^1) + \upsilon_{it} + \xi_{it}, \tag{35}$$

where $P_*(\omega_{it})$ lies between $\widehat{P}(\omega_{it})$ and $P_0(\omega_{it})$. Let $f_u(.)$ be the density of $F_u(.)$, then $\upsilon_{it} = \frac{1}{f_u[F_u^{-1}(P_0(\omega_{it}))]}(\widehat{P}(\omega_{it}) - P_0(\omega_{it}))$ and $\xi_{it} = \left(\frac{1}{f_u[P_*(\omega_{it})]} - \frac{1}{f_u[F_u^{-1}(P_0(\omega_{it}))]}\right)(\widehat{P}(\omega_{it}) - P_0(\omega_{it}))$. The terms υ_{it} and ξ_{it} are weighted discrepancy measures between the true

conditional selection probability and the estimated one. If $\widehat{P}(\omega_{it})$ is a consistent estimator, then υ_{it} and ξ_{it} are asymptotically zero. Equation (35) then approximately defines a heteroscedastic transformation model of which we can take first differences to eliminate the unknown nuisance function, $\eta(z_i^1)$, which gives

$$\triangle F_u^{-1}(\widehat{P}(\omega_{it}), \widehat{P}(\omega_{it-1})) \equiv \phi_0 \triangle d_{it-1} + \phi_1 \triangle y_{it-1} + \triangle z_{it}\gamma + \triangle \upsilon_{it} + \triangle \xi_{it}. \tag{36}$$

As discussed in section 2 (see Manski, 1987 or Ichimura, 1993 for details), we normalize the parameter of the continuous lagged outcome ϕ_1 to -1, [4] so that (36) can be expressed as

$$\triangle y_{it-1} \equiv \phi_0 \triangle d_{it-1} + \triangle z_{it}\gamma - \triangle F_u^{-1}(\widehat{P}(\omega_{it}), \widehat{P}(\omega_{it-1})) + \triangle \upsilon_{it} + \triangle \xi_{it}.$$

Finally, to estimate the selection parameters, one need to find a consistent nonparametric estimator of the conditional selection probability. Among many types available in the nonparametric literature, we will choose the kernel density estimator

$$\widehat{P}(\omega_{it}) = \sum_{j=1}^{n} \frac{d_{jt} K_h(\omega_{jt} - \omega_{it})}{\sum_{k=1}^{n} K_h(\omega_{jt} - \omega_{it})}, \tag{37}$$

where h is a positive smoothing parameter that goes to zero as the sample size increases, $K_h(u) = \frac{1}{h^{d_\omega}} K(\frac{u}{h})$ for a given kernel K (with compact support S_ω) and d_ω is the number of continuous variables in ω_{it}.

To ease the exposition, the following additional notations are required. Define $\omega_i = \{\omega_{is}\}_{s=1}^{T}, \omega_{i-1} = \{\omega_{is-1}\}_{s=1}^{T}, \omega_{i-2} = \{\omega_{is-2}\}_{s=2}^{T}, \triangle\omega_i = \{\omega_{is}\}_{s=2}^{T}, \triangle\omega_{i-1} = \{\omega_{is}\}_{s=1}^{T-1}$. We define the same way the variables y_i, y_{i-1}, y_{i-2}, d_i, d_{i-1}, d_{i-2}, z_i, z_{i-1}, $z_{i-2}, \overline{v}_i, \overline{v}_{i-1}$, $\overline{v}_{i-2}$, $\triangle y_i$, $\triangle y_{i-1}$, $\triangle y_{i-2}$, $\triangle d_i$, $\triangle d_{i-1}$, $\triangle d_{i-2}$, $\triangle z_i$, $\triangle z_{i-1}$, $\triangle z_{i-2}$, $\triangle \overline{v}_i, \triangle \overline{v}_{i-1}$, $\triangle \overline{v}_{i-2}$.

Let $P(\omega_i) \equiv \{P(\omega_{is})\}_{s=2}^{T}$ and $P(\omega_{i-1}) \equiv \{P(\omega_{is})\}_{s=1}^{T-1}$. Let $\theta_1 \equiv (\phi_0, \gamma, \triangle F^{-1}(.))' \in \Theta_1$ denote the vector of parameters to be estimated where the infinite-dimensional parameter space $\Theta_1 \equiv \Lambda_1 \times \mathcal{H}_1$ can be decomposed into a finite-dimensional space, Λ_1, and an infinite-dimensional space $\mathcal{H}_1$ with $(\phi_0, \gamma) \in \Lambda_1$ and $\triangle F^{-1}(.) \in \mathcal{H}_1$.

Let

$$\ell(\theta_1, \omega_i, \widehat{P}(\omega_i), \widehat{P}(\omega_{i-1})) = -\frac{1}{2}\left([\widehat{\triangle \upsilon_i + \triangle \xi_i}]'[\widehat{\triangle \upsilon_i + \triangle \xi_i}]\right)', \tag{38}$$

where $\widehat{\triangle \upsilon_i + \triangle \xi_i} = \triangle y_{i-1} - \phi_0 \triangle d_{i-1} - \triangle z_i \gamma + \triangle F^{-1}(\widehat{P}(\omega_i), \widehat{P}(\omega_{i-1}))$. We then define our Least Squared (LS) estimator of θ_1 as

$$\sup_{\theta_1 \in \Theta_1} Q_{1n}(\theta_1, \widehat{P}(\omega_i), \widehat{P}(\omega_{i-1})) = \sup_{\theta_1 \in \Theta_1} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta_1, \omega_i, \widehat{P}(\omega_i), \widehat{P}(\omega_{i-1})). \tag{39}$$

[4] Note that we could alternatively normalize any other continuous variable in z_{it}.

However, since $\mathcal{H}_1$ is infinite dimensional, maximizing over Θ_1 may not be well defined; even if the maximizer exists, it will generally be too difficult to compute. Instead, the maximization will be restricted to a sequence of approximating spaces, Θ_{1n}, such that $\bigcup_n \Theta_{1n}$ is dense in Θ_1. These types of estimators are called sieve LS in the econometric literature (see Chen, 2005, for a comprehensive survey of this literature). Following this literature, the selection equation estimator can be redefined as

$$\widehat{\theta}_1 = \underset{\theta_1 \in \Theta_{1n}}{\arg\max} Q_{1n}(\theta_1, \widehat{P}(\omega_i), \widehat{P}(\omega_{i-1})), \tag{40}$$

where $\Theta_{1n} \equiv \Lambda_1 \times \mathcal{H}_{1n}$ such that $\bigcup_n \mathcal{H}_{1n}$ is dense in $\mathcal{H}_1$.

We will restrict our analysis to linear sieve spaces that are compact, nondecreasing (i.e., $\mathcal{H}_{1n} \subseteq \mathcal{H}_{1n+1} \subseteq \ldots, \subseteq \mathcal{H}_1$). We provide below the example of two univariate linear sieves, which bivariate extensions will be used to estimate $\triangle F^{-1}(P(\omega_i), P(\omega_{i-1}))$. Since $P(\omega_i)$ lies in the interval $[0,1]$, let $Pol(J_n)$ denote the space of polynomials on $[0,1]$ of degree J_n or less; that is,

$$Pol(J_n) = \left\{ \sum_{k=0}^{J_n} a_k x^k,\ x \in [0,1] :\ a_k \in \mathcal{R} \right\}. \tag{41}$$

Let $TriPol(J_n)$ denote the space of trigonometric polynomials on $[0,1]$ of degree J_n or less; that is

$$TriPol(J_n) = \left\{ \frac{a_0}{2} + \sum_{k=1}^{J_n} a_k \cos(2k\pi x) + b_k \sin(2k\pi x),\ x \in [0,1] : a_k, b_k \in \mathcal{R} \right\}. \tag{42}$$

Since we will assume that $\triangle F_0^{-1}(P(\omega_i), P(\omega_{i-1}))$ belongs to a Hölder space, say $\Lambda_{c_1}^{p_1}\left([0,1]^2\right)$, it will be well approximated by the bivariate versions of both $Pol(J_n)$ and $TriPol(J_n)$. We will consider the tensor product linear sieve space $\mathcal{H}_{1n}$, which is constructed as a tensor product space of the univariate linear approximating spaces $\mathcal{H}_{1n1}, \mathcal{H}_{1n2}$. Let $dim\left(\mathcal{H}_{1n}\right) = k_{1n}$ and $[p]$ be the biggest integer satisfying $[p] < p_1$. Then, the approximation error rates for polynomials or orthogonal wavelets for $\triangle F_0^{-1}$ are of order $O(k_{1n}^{-\frac{p_1}{2}})$ (see Timan, 1963).

3.2 Outcome Equation Estimation

For notational ease, the following additional notations will be required. Let $E[g(\theta_2, Z_i) \mid \digamma_i] \equiv \{E[g_s(\theta_2, Z_{is}) \mid \digamma_{is}]\}_{s=4}^T$ where $\theta_2 \equiv (\rho, \beta, \triangle\overline{\lambda}(.))$,
$Z_{is} \equiv (d_{is}, d_{is-1}, \triangle y_{is}, \triangle y_{is-1}, \triangle x_{is}, \overline{v}_{is}, \overline{v}_{is-1}, \overline{v}_{is-2}, z_i^1)$, $Z_i \equiv \{Z_{is}\}_{s=4}^T$, $\digamma_i \equiv \{\digamma_{is}\}_{s=4}^T$ and

$$g_s(\theta_2, Z_{is}) \equiv d_{is} d_{is-1} (\triangle y_{is} - \rho \triangle y_{is-1} - \triangle x_{is}\beta - \triangle\overline{\lambda}(\overline{v}_{is}, \overline{v}_{is-1}, \overline{v}_{is-2}, z_i^1)). \tag{43}$$

From Proposition 2, we have a set of conditional moment restrictions, which can be expressed as

$$E[g(\theta_{20}, Z_i) \mid \digamma_i] = 0, \tag{44}$$

where θ_{20} is the true value of θ_2.

Let $\widehat{m}(\theta_2, \digamma_i) \equiv \{\widehat{m}_s(\theta_2, \digamma_{is})\}_{s=4}^{T}$, where $\widehat{m}_s(\theta_2, \digamma_{is})$ is a consistent nonparametric estimator of the sth element of the vector $E[g(\theta_2, Z_i) \mid \digamma_i]$. Following Ai and Chen (2003), our third-step estimator can be expressed as

$$\inf_{\theta_2 \in \Theta_{2n}} \widehat{Q}_{2n}(\theta_2) = \inf_{\theta_2 \in \Theta_{2n}} \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(\theta_2, \digamma_i)' [\widehat{\Sigma}_o(\digamma_i)]^{-1} \widehat{m}(\theta_2, \digamma_i), \tag{45}$$

where $\Theta_{2n} = \Lambda_2 \times \mathcal{H}_{2n}$, $\widehat{\Sigma}_o(\digamma_i) \longrightarrow \Sigma_o(\digamma_i)$ in probability and $\Sigma_o(\digamma_i)$ is a positive definite weighting matrix of the same dimension as $E[g(\theta_2, Z_i) \mid \digamma_i]$. We restrict our analysis to linear sieve spaces that are compact, nondecreasing (i.e., $\mathcal{H}_{2n} \subseteq \mathcal{H}_{2n+1} \subseteq \ldots, \subseteq \mathcal{H}_2$).

Let $b_{oj}(\digamma_{is}), j = 1, 2, \ldots, k_{s,n}\}$ be a sequence of known basis functions that approximate any real-valued L$_2$-functions of $\digamma_{is}$ well as $k_{s,n} \longrightarrow \infty$. Denote by $B^{k_{s,n}}(\digamma_{is}) = (b_{o1}(\digamma_{is}), \ldots, b_{ok_{s,n}}(\digamma_{is}))'$ and $B_s = (B^{k_{s,n}}(\digamma_{1s}), \ldots, B^{k_{s,n}}(\digamma_{ns}))'$. Following Ai and Chen (2003),[5] a series LS estimator of the conditional expectation $E[g_s(\theta_2, Z_{is}) \mid \digamma_{is}]$ is

$$\widehat{m}_s(\theta_2, \digamma_{is}) = \frac{1}{n} \sum_{j=1}^{n} g_s(\theta_2, Z_{js}) B^{k_{s,n}}(\digamma_{js})' (B_s' B_s)^{-1} B^{k_{s,n}}(\digamma_{is}) \tag{46}$$

As with the standard GMM type estimator, in order to implement the above estimator, one needs to be able to estimate $\Sigma_o(\digamma_i)$. One method is to use a nonparametric conditional-variance estimator of the moment restrictions calculated from a preliminary consistent estimator of θ_2 (see Robinson, 1987, for example). We use a LS estimator of the conditional variance, $\Sigma_o(\digamma_i)$ as $\widehat{\Sigma}_o(\digamma_i) \equiv \widehat{\Sigma}_o(\digamma_i, \widetilde{\theta}_2) \equiv \{\sigma_{0st}(\digamma_{is}, \widetilde{\theta}_2)\}_{s,t=2,\ldots,T}$, where

$$\sigma_{0st}(\digamma_{is}, \widetilde{\theta}_2) \equiv \begin{cases} \frac{1}{n} \sum_{j=1}^{n} g_s(\widetilde{\theta}_2, Z_{js}) g_t(\widetilde{\theta}_2, Z_{jt}) B^{k_{s,n}}(\digamma_{js})' (B_s' B_s)^{-1} B^{k_{s,n}}(\digamma_{is}) & \text{for s=t} \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

and $\widetilde{\theta}_2$ is a preliminary consistent estimator, normally obtained by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{m}(\theta_2, \digamma_i)' \widehat{m}(\theta_2, \digamma_i). \tag{48}$$

[5]Note that any consistent nonparametric estimator of $E[g(\theta_2, Z_i) \mid \digamma_i]$ could be used. However, we follow Ai and Chen (2003) because of the possible large dimension of $\digamma_i$.

The line in (47) follows from the fact the $\digamma_{is} \subset \digamma_{is+1}$ in conjunction with Proposition 1. At this stage, we will use the multivariate version of $Pol(J_n)$ or $TriPol(J_n)$ for estimating $\triangle\overline{\lambda}(.)$ and corresponding tensor product sieves for the estimation of $\widehat{m}(\theta_2, \digamma_i)$ and $\widehat{\Sigma}_o(\digamma_i)$ (see Chen, 2005).

4 Large-Sample Properties

Recall that $\omega_i \equiv \{\omega_{is}\}_{s=1}^{T}$. Let $d = \dim(\omega_i)$. For ease of exposition, we redefine the first-step kernel estimator as a component of

$$\widehat{p}(\omega_i) \equiv \left(\widehat{p}_1(\omega_i)', \widehat{p}_2(\omega_i)'\right)' \equiv \sum_{j=1}^{n} \widetilde{\mathrm{d}}_j K_h(\omega_j - \omega_i), \tag{49}$$

where $\widetilde{\mathrm{d}}_j = [1, \mathrm{d}_j]'$ so that $\widehat{p}(\omega_i)$ is the kernel estimate of $p_0(\omega_i) \equiv (p_{01}', p_{02}')' \equiv f_\omega(\omega_i) E[\widetilde{d}_i \mid \omega_i]$ and $f_\omega(.)$ denotes the marginal density of ω. In particular, $\widehat{P}(\omega_i) = \frac{\widehat{p}_2(\omega_i)}{\widehat{p}_1(\omega_i)}$. This notational change for the conditional expectation[6] eases the exposition in the results that follow. The conditions summarized in Assumption 4.1 below, ensure that $\widehat{p}$ is close to p_0 for n large enough (see Newey and McFadden, 1994).

Assumption 4.1 $(\forall i = 1, ..., n)$**:**

1. *There is a version of $p_0(\omega_i)$ that is continuously differentiable of order q $(> r)$ with bounded derivatives on an open set,*[7] *and $p_{01}(\omega_i) = f_\omega(\omega_i)$ is bounded away from 0 on S_ω, the compact support of ω_i.*

2. *$K(u)$ is differentiable of order q, $K(u)$ is zero outside a bounded set, $\int K(u)du = 1$, and there is a positive integer L, such that for all $j < L$, $\int K(u)(\bigotimes_{l=1}^{j} u)du = 0$.*

3. *There is $p \geq 4$ such that $E\left(\left\|\widetilde{d}_i\right\|^p\right) < \infty$ and $E\left(\left\|\widetilde{d}_i\right\|^p \mid \omega_i\right) f_\omega(\omega_i)$ is bounded.*

4. *The bandwidth h satisfies $h(n) \to 0$ and $n^{1-\frac{2}{p}} h(n)^d \ln n \to \infty$.*

Assumption 4.2 $(\forall i = 1, ..., n)$**:**

1. *$\Theta_{1n} \subseteq \Theta_{1n+1} \subseteq \cdots \subseteq \Theta_1$ compact for all $n \geq 1$ and for any $\theta_1 \in \Theta_1$ there exists $\pi_n\theta_1 \in \Theta_{1n}$ such that $\|\theta_1 - \pi_n\theta_1\| = o(1)$ as n gets large.*

[6] Now $E[\mathrm{d} \mid \omega] = p_{02}(\omega)/p_{01}(\omega) = f_\omega(\omega)E[\widetilde{\mathrm{d}} \mid \omega]/f_\omega(\omega)$.

[7] Note that r is the dimension of the continuous components of ω.

2. $\sigma_1^2(P_0, \omega_i) = E\{(\triangle \upsilon_i)^2 \mid P_0, \omega_i\} < \infty$.

3. $\triangle F^{-1}(.) \in H_1 = \Lambda_{c_1}^{p_1}[s_l, s_u]^2$, *with* $p_1 > d/2$ *and* $[s_l, s_u] \subset [0, 1]$.

Theorem 1: *Under Assumptions 2.1, 2.2, 4.1 and 4.2, if* $k_{1n} = O(n^{1/(2p_1+d)})$ *then* $\left\|\widehat{\theta}_1 - \theta_{10}\right\| = O_p(n^{-p_1/(2p_1+d)})$.

Assumption 4.1 is standard in the nonparametric literature (see Newey, 1994, and Newey and McFadden, 1994, for a proof and discussions of this result).

Assumption 4.2.1 is the regularity condition on the sieve space (see Chen and Shen, 1998). Assumption 4.2.2 is standard, bounding the second conditional moments. Assumption 4.2.3 imposes a smoothness condition on the class of functions. Theorem 1 is a consistency and rate of convergence result, which is an application of Theorem 1 in Chen and Shen (1998).[8]

To study the asymptotic distribution of $\widehat{\theta}_1$, we use a linear approximation of the criterion difference by the corresponding derivatives and the degree of smothness of $g(.)$, where $g(\theta_1)$ is a real functional of θ_1. We will show that the sieve estimator has a normal distribution and is efficient when the empirical criterion satisfies certain stochastic equicontinuity conditions. The degree of smoothness of $g(.)$ can compensate for the slowness of the convergence rate of the estimate.

Let $P_i \equiv (P(\omega_i), P(\omega_{i-1}))$ and $P_{i0} \equiv (P_0(\omega_i), P_0(\omega_{i-1}))$. Let $\omega \equiv (d_{-1}, y_{-1}, z', z^1)' \equiv \{\omega_i\}_{i=1}^n$, $P \equiv \{P_i\}_{i=1}^n$ and $P \equiv \{P_{i0}\}_{i=1}^n$ and all other relevant variables in the same manner.

Let

$$D_{u_1^*}(\omega)' \equiv \begin{bmatrix} \triangle d_{-1} \\ \triangle z \end{bmatrix} - u_1^*(P_0), \tag{50}$$

where $u_1^*(P)$ solves the following programming problem

$$\inf_{u_1 : E\left[\left|\begin{bmatrix} \triangle d_{-1} \\ \triangle z \end{bmatrix} - u_1(P_{i0})\right|_e^2\right] > 0} E\left[\left(\begin{bmatrix} \triangle d_{-1} \\ \triangle z \end{bmatrix} - u_1(P_{i0})\right)\left(\begin{bmatrix} \triangle d_{-1} \\ \triangle z \end{bmatrix} - u_1(P_0)\right)'\right]. \tag{51}$$

Assumption 4.3 $(\forall i = 1, ..., n)$**:**

1. $(\phi_{00}, \gamma_0) \in int(\Theta_1)$.

[8] These results are summarized in Chen (2005).

2. $E[D_{u_1^*}(\omega_i)' D_{u_1^*}(\omega_i)]$ *is positive definite;* $\Sigma_1(\omega_i) = E\left[(\triangle \upsilon_i + \triangle \xi_i)^2 \,|\omega_i\right]$ *is positive definite.*

3. *Each element of* $u_1^*(P_i)$ *belongs to the Hölder space* $\Lambda^{p_{11}}$ *with* $p_{11} > \frac{d}{2}$.

Theorem 2: Under Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3, if $k_{1n} = O(n^{1/(2p_1+d)})$, then
$\sqrt{n}[(\widehat{\phi}_0, \widehat{\gamma})'\text{-}(\phi_{00}, \gamma_0')'] \Longrightarrow N(0, \Omega_1)$, where

$$\Omega_1 = E[D_{u_1^*}(\omega_i)' D_{u_1^*}(\omega_i)]^{-1} E[D_{u_1^*}(\omega_i)' \Sigma_1(\omega) D_{u_1^*}(\omega_i)] E[D_{u_1^*}(\omega_i)' D_{u_1^*}(\omega_i)]^{-1}.$$

Assumption 4.3 is a standard regularity condition in sieve estimation (see Shen, 1997, Chen and Shen, 1998, and Chen, 2005, for discussions of these conditions). Note that as with the standard Berkson minimum chi-square estimation in standard binary-choice models, the asymptotic variance of the estimator depends only on the variance of $\triangle \upsilon_i$ and not on that of $\triangle \xi_i$ (see Amemiya, 1994, p. 277 for an example). Let k_{2n} denote the dimension of the approximating sieve space, $\mathcal{H}_{2n}$. Let

$$\Sigma_o(\digamma_i) \equiv \begin{pmatrix} \sigma_2^2(\digamma_{i2}) & ... & o \\ . & . & \\ 0 & & \sigma_T^2(\digamma_{iT}) \end{pmatrix}$$

where $\sigma_2^2(\digamma_{is}) = E[g_s(\theta_{20}, Z_{is})^2 \mid \digamma_{is}]$.

Assumption 4.4 $(\forall i = 1, ..., n)$**:**

1. α_i *and* z_i^1 *are i.i.d over individuals.*

2. *The support of* $\{x_{it}, z_{it}, y_{io}^*, \alpha_i, \varepsilon_{it}^*\}_{t=1}^T$ *is compact with nonempty support.*

3. $(\rho, \beta, \phi_0, \phi_1, \gamma) \in \Lambda_1 \times \Lambda_2$*, with* Λ_1, Λ_2 *are compact with nonempty interior and* $|\rho| < 1$, $|\phi_0| < 1$.

4. *The density of* $\{x_{it}, z_{it}, y_{io}^*, d_{io}, \alpha_i, \varepsilon_{it}^*, u_{it}\}_{t=1}^T$ *is bounded and bounded away from zero.*

5. *Either* $B^{k_{s,n}}(\digamma_{is})$ *is a tensor product of Fourier series with* $k_{s,n} k_{2n} \ln(n)/\sqrt{n} = o(1)$ *or a tensor product power series with* $k_{s,n}^2 k_{2n} \ln(n)/\sqrt{n} = o(1)$, where k_{2n} is defined in (70).

6. $\dim(g(.))\, k_{s,n} \geq 1 + \dim(x_{it}) + k_{2n}$*;* $k_{s,n}^{-\frac{p_3}{\dim(F)}} = o(n^{-\frac{1}{4}})$ *and* $k_{2n}^{-\frac{p_2}{3+\dim(z^1)}} = o(n^{-\frac{1}{4}})$.

7. *The smallest and largest eigenvalues of* $E\{B^{k_{s,n}}(\digamma_{is})B^{k_{s,n}}(\digamma_{is})'\}$ *are bounded and bounded away from zero for all* $k_{s,n}$.

8. *(i)* $\sigma_2^2(\digamma_i,\theta_2)>0$ *and* $\Sigma_o(\digamma_i,\theta_2)$ *is finite positive definite uniformly over* $\digamma_i\in\Im$, $\theta_2\in N_{on}$*; (ii) each element of* $g(\theta_2,Z_i)g(\theta_2,Z_i)'$ *satisfies an envelope condition and is Hölder continuous in* $\theta_2\in N_o$*; (iii) Each element of* $\Sigma_o(\digamma_i,\theta_2)$ *is in* $\Lambda_{c_\Sigma}^{p_\Sigma}$ *with* $p_\Sigma>\frac{d_{\digamma}}{2}$ *for all* $\theta_2\in N_{on}$.

9. $\triangle\overline{\lambda}_0(.)\in H_2=\Lambda_{c_2}^{p_2}(R^{3+\dim(z^1)})$ *with* $p_2>1$ *and* $E[(1+(\overline{v}_i,\overline{v}_{i-1},\overline{v}_{i-2},z_i^1)'(\overline{v}_i,\overline{v}_{i-1},\overline{v}_{i-2},z_i^1)^a\mid \digamma_i]$ *is bounded for some* $a>p_2$.

10. $E[dd_{-1}\triangle y_i\mid \digamma_i=\overline{f}]$, $E[dd_{-1}\triangle x_i\mid \digamma_i=\overline{f}]$, $E[dd_1q^{k_{2n}}(\overline{v}_i,\overline{v}_{i-1},\overline{v}_{i-2},z_i^1)\mid \digamma_i=\overline{f}]\in\Lambda_{c_3}^{p_3}(\Im)$, $p_3>(\dim(\digamma_i))/2$.

11. *(i)* $E(|\triangle y_i|^4)<\infty$*; (ii)* $E(\|\triangle y_i\|_E\mid \digamma_i]<\infty$, $E(\|\triangle y_{i-1}\|_E\mid \digamma_i]<\infty$, $E(\|\triangle x_i\|_E\mid \digamma_i]<\infty$ *and* $E(\sup_{\theta_1\in\Theta_1}\left\|\triangle\overline{\lambda}_0(\overline{v}_i(\theta_1),\overline{v}_{i-1}(\theta_1),\overline{v}_{i-2}(\theta_1),z_i^1)\right\|_E\mid \digamma_i]<\infty$.

Let v denote $(\overline{v}_{is},\overline{v}_{is-1},\overline{v}_{is-2},z_i^1)$ for $i=1,...,n,s=3,...,T$. Let $\digamma\equiv\{\digamma_i\}_{i=1}^n$ and denote by $\frac{dm(\digamma,\theta_{20})}{d\theta_2}[\theta_{21}-\theta_{22}]$ the first pathwise derivative of $m(.)$ at the direction $[\theta_{21}-\theta_{22}]$ evaluated at θ_{20}.

Theorem 3: For any $\theta_2,\theta_{21},\theta_{22}\in\Theta_2$, define $\|\theta_{21}-\theta_{22}\|=\sqrt{E\left\{\left\{\frac{dm(\digamma,\theta_{20})}{d\theta_2}[\theta_{21}-\theta_{22}]\right\}'\Sigma(\digamma,\theta_2)^{-1}\frac{dm(\digamma,\theta_{20})}{d\theta_2}[\theta_{21}-\theta_{22}]\right\}}$. *Under Assumptions 2.1-2.4 and 4.1–4.4,* $\left\|\widehat{\theta}_2-\theta_{20}\right\|=o_p(n^{-\frac{1}{4}})$.

Assumptions 4.4.1-4.4.10 are standard in the sieve minimum-distance literature (see Ai and Chen, 2003, for a complete discussion on the importance of these regularity conditions). Assumption 4.4.11 is an additional boundedness condition needed to ensure that our moment conditions are continuous in the preliminary estimates of the selection equation.

Let $u_2^*=(u_{21}^*,\ldots,u_{2,1+\dim(x)}^*)$ be the solution to

$$\min_{u_{21}^*\in\overline{\mathcal{U}}}E\{[\triangle y_{-1}-E\{u_{21}^*(\overline{v},\overline{v}_{-1},\overline{v}_{-2},z^1)\mid \digamma\}]^2\} \tag{52}$$

and

$$\min_{u_{2,1+j}^*\in\overline{\mathcal{U}}}E\{[\triangle x_j-E\{u_{21}^*(\overline{v},\overline{v}_{-1},\overline{v}_{-2},z^1)\mid \digamma\}]^2\}\qquad(j=1,\ldots,dim(x)) \tag{53}$$

Let $D_{u_2^*}(\digamma) = [\triangle y_{-1}, \triangle x] - E\{u_2^*(\overline{v}, \overline{v}_{-1}, \overline{v}_{-2}, z^1) \mid \digamma\}$,
$H_1 = E\left[D_{u_2^*}(\digamma)'\Sigma_0(\digamma, \theta_{20})^{-1} dd_{-1}\left\{\frac{\partial\overline{\lambda}_0(\overline{v},\overline{v}_{-1},z^1)}{\partial\overline{v}}(d_{-1}, z)' + \frac{\partial\overline{\lambda}_0(\overline{v},\overline{v}_{-1},z_i^1)}{\partial\overline{v}}(d_{-2}, z_{-1})'\right\}\right]$,
$\Omega_2 = E\{D_{u_2^*}(\digamma)'\Sigma_0(\digamma, \theta_{20})^{-1} D_{u_2^*}(\digamma)\}$ and
$H_2 = E\{D_{u_2^*}(\digamma)'\Sigma_0(\digamma, \theta_{20})^{-1} D_{u_2^*}(\digamma)\}$.

Assumption 4.5:

1. $(\rho_0, \beta_0) \in \Lambda_2$.
2. *(i) For j=1,. . . ,dim(*β*)* $+ 1$,
   $E[u_{2j}^*(\overline{v}, \overline{v}_{-1}, \overline{v}_{-2}, z^1) \mid \digamma = \overline{f}] \in \Lambda_{c_4}^{p_4}(\Im)$, *for* $p_4 > (\dim(\digamma))/2$*; (ii)* H_1 *is bounded.*

Theorem 4: *Under Assumptions 2.1-2.4, 4.1–4.5,* $\sqrt{n}[(\widehat{\rho}, \widehat{\beta})' - (\rho_0, \beta_0)'] \Longrightarrow N(0, V)$, *where* $V = H_2^{-1}[\Omega_2 + H_1\Omega_1 H_1']H_2^{-1}$.

Assumptions 4.5.1 and 4.5.2(i) are standard in the sieve minimum-distance literature (see Ai and Chen, 2003). Assumption 4.5.2 (ii) is an additional condition needed to correct the asymptotic variance for the preliminary estimates of the selection equation.

It is important to note that as the number of variables in $\digamma$ increases, the smoothness requirement on $E[\triangle y \mid \digamma = \overline{f}]$ and $E[q^{k_{2n}}(\overline{v}, \overline{v}_{-1}, \overline{v}_{-2}, z^1) \mid \digamma = \overline{f}]$ where $q^{k_{2n}}$ is defined as in (70), increases as well. This suggests that if we use less instruments in our estimation, it may have better small-sample properties. This is not new in this literature as it was pointed out by Ahn and Schmidt (1995) in linear panel-data models. To this end it may be practical to use less variables in the conditioning set when estimating $E[g(\theta_2, Z) \mid \digamma]$.

5 Monte Carlo Study

In this section, we present simulation results illustrating the performance of the estimation procedure described in the preceding sections. We consider two models: the first illustrates how our complete model (including both exogenous and predetermined variables in the selection equation) performs in a limited Monte Carlo study estimating both the selection and outcome equations. The second compares the performance of our estimator of the outcome equation to that of Kyriazidou (2001). However, because her estimator cannot handle predetermined variables in the selection equation, we do not include any such regressors in the second model.

5.1 Model 1 : Predetermined Variable in the Selection Equation

Data for the Monte Carlo experiment of this model are generated for $t = 0, 1, 2, 3$ according to the following specification.

For $i = 1, \ldots, n$; $t = 1, 2, 3$,

$$y_{it}^* = \rho y_{it-1}^* + x_{it}\beta + \alpha_i + \varepsilon_{it}^*,$$
$$y_{it} = d_{it} y_{it}^*,$$
$$d_{it} = \mathbb{I}\{\phi_0 d_{it-1} + \phi_1 y_{it-1} + \gamma_1 z_{1it} + \gamma_2 z_{2it} + \eta_i + u_{it} \geq 0\}.$$

In the selection equation, the predetermined variables are generated so that $z_{1it} = 6 + 0.5y_{it-1}^* + v_{it}$, with v_{it} independently distributed as $U(0,1)$, whereas the strictly exogenous variables z_{2it} are distributed as $N(0,1)$ and are the same strictly exogenous latent variables as those of the outcome equation, x_{it}. The individual-specific effect $\eta_i = \frac{1}{3}(z_{2i1} + \cdots + z_{2i3}) + 2\xi - 0.75$, where ξ is an independently distributed $U(0,1)$, whereas the time-varying error term, u_{it}, is distributed as $N(0, 0.5)$. In order to satisfy the scale normalization needed for identification of parameters in the selection equation, we normalized $\gamma_2 = 1$ in the estimation. Finally for simulation purposes, we assume that $\phi_0 = \phi_1 = 0.5, \gamma_1 = 1$. In the main equation, $\alpha_i = \frac{1}{3}(z_{2i1} + \cdots + z_{2i3})$ and $\varepsilon_{it}^* = 0.8\xi^* + 0.6\,(u_{it} - E\,(u_{it}))$, where ξ^* is an independent standard normal vector. The initial observation, y_{i0}, is generated as $y_{i0} = d_{i0}y_{i0}^*$, where y_{i0}^* is $N(-2, 1)$ and d_{i0} is generated according to a binomial distribution. We investigate the small-sample properties of the estimators of ρ and β, whereas their true values are respectively assumed to be $\rho = 0.5$, $\beta = 3$. Three sample sizes, n, are considered: 500, 1000, 2000.

We first estimate our selection equation. In all cases that follow, the approximation of unknown functions with power or Fourier series, respectively, use the results of the small-sample experiment of our selection equation estimator, displayed in Table 1 for 100 replications. For the design under investigation, we note that, in general, estimation using Fourier series leads to better mean and variance estimates than does the power series. We use the normal multivariate kernel with the bandwidth as the Silverman's rule of thumb. We trimmed the upper and low 2.5% of the data.

We nonparametrically estimate $\triangle F^{-1}(\widehat{P}(\omega_{it}), \widehat{P}(\omega_{it-1}))$ for $t = 2, 3$ and all i by using the the bivariate Fourier and power series basis functions. We use an order of ten for n=2000 (i.e. $k_{1,2000} = 10$), an order of seven for n=1000 (i.e. $k_{1,1000} = 7$), and an order of five for n=500 (i.e.$k_{1,500} = 5$). The results are reported in Table 1 below. As can been seen from the results, our estimator performs well in this limited Monte Carlo.

TABLE 1

FINITE-SAMPLE PERFORMANCE OF THE SECOND-STEP ESTIMATOR

Sample Size		500		1000		2000	
		Power Series	Fourier Series	Power Series	Fourier Series	Power Series	Fourier Series
ϕ_0	Mean	0.456	0.473	0.466	0.483	0.491	0.492
	Std Dev.	0.0122	0.107	0.092	0.079	0.031	0.025
	RMSE	0.012	0.010	0.007	0.006	2e-3	1e-3
ϕ_1	Mean	0.473	0.482	0.483	0.505	0.492	0.498
	Std Dev.	0.118	0.093	0.067	0.053	0.032	0.019
	RMSE	0.0113	0.0112	0.005	0.004	8e-3	1e-4
γ_1	Mean	0.943	0.963	0.973	0.978	1.003	1.001
	Std Dev.	0.145	0.134	0.076	0.091	0.008	0.006
	RMSE	0.011	9.4e-3	5.6e-3	0.003	8e-3	7e-4
γ_2	Mean	-	-	-	-	-	-
	Std Dev.	-	-	-	-	-	-
	RMSE	-	-	-	-	-	-

In order to estimate the structural parameters ρ and β of our model we use two versions of our estimator. The first version of our estimator is what we will call the sieve instrumental variable (SIV) estimator. This estimator is obtained by noting that the conditional moment restrictions can be written as an increasing sequence of unconditional moment restrictions. Using the identity matrix as the weighting matrix, our conditional moment restrictions imply the following unconditional moment restrictions

$$E[b_{0j}(\mathcal{F}_{it})g_t(\theta_{20}, Z_{it})] = 0$$

for $t = 2, ..., T$ and $j = 1, ..., k_{t,n}$. The model can now be estimated using any standard instrumental variable or two stage least square method with $\{b_{0j}(\mathcal{F}_{it})\}_{j=1}^{k_{t,n}}$ as the instruments (see Ai and Chen, 2003 and Chen, 2005 for details). The second version of our estimator is the standard sieve minimum distance (SMD) described in the text.

In this version of the Monte Carlo simulation, we have two moment restrictions for

$$g_3(\theta_2, Z_{i3}) = d_{i3}d_{i2}[\Delta y_{i3} - \rho\Delta y_{i2} - \Delta x_{i3}\beta - \Delta\overline{\lambda}(\overline{v}_{i3}, \overline{v}_{i2}, \overline{z}_{2i})],$$

and

$$g_2(\theta_2, Z_{i2}) = d_{i2}d_{i1}[\Delta y_{i2} - \rho\Delta y_{i1} - \Delta x_{i2}\beta - \Delta\overline{\lambda}(\overline{v}_{i2}, \overline{v}_{i1}, \overline{z}_{2i})],$$

where

$$\overline{v}_{i3} = \phi_0 d_{i2} + \phi_1 y_{i2} + \gamma_1 z_{1i3} + \gamma_2 z_{2i3},$$

$$\overline{v}_{i2} = \phi_0 d_{i1} + \phi_1 y_{i1} + \gamma_1 z_{1i2} + \gamma_2 z_{2i2},$$
$$\overline{v}_{i1} = \phi_0 d_{i0} + \phi_1 y_{i0} + \gamma_1 z_{1i1} + \gamma_2 z_{2i1},$$

and

$$\overline{z}_{2i} = \frac{1}{3}(z_{2i1} + z_{2i2} + z_{2i3}).$$

The conditioning sets becomes

$$\mathcal{F}_{i2} = \{y_{i0}, d_{i0}, x_{i1}, x_{i2}, z_{1i1}, \overline{z}_{2i}, d_{i2}d_{i1} = 1\}$$

and

$$\mathcal{F}_{i3} = \{y_{i0}, y_{i1}, d_{i0}, d_{i1}, x_{i1}, x_{i2}, x_{i3}, z_{1i1}, \overline{z}_{2i}, d_{i3}d_{i1} = 1\}.$$

The series LS estimator of the conditional expectation, $E[g(\theta_2, Z_i) \mid \mathcal{F}_i]$ used in the SMD estimator is

$$\widehat{m}(\theta_2, \mathcal{F}_i) = \{\widehat{m}_t(\theta_2, \mathcal{F}_{it})\}_{t=2,3}, \tag{54}$$

where

$$\widehat{m}_t(\theta_2, \mathcal{F}_{it}) = \sum_{j=1}^{n} g_t(\theta_2, Z_{jt}) B^{k_{t,n}}(\mathcal{F}_{jt})'(B_t'B_t)^{-1}B^{k_{t,n}}(\mathcal{F}_{it}) \tag{55}$$

and the LS estimator of the conditional variance $\Sigma_o(\mathcal{F}_i)$ as $\widehat{\Sigma}_o(\mathcal{F}_i) \equiv \{\sigma_{0ts}(\mathcal{F}_{it})\}_{t,s=2,3}$, where

(56)

$$\sigma_{0st}(\mathcal{F}_{is}) \equiv \begin{cases} \sum_{j=1}^{n} g_s(\widetilde{\theta}_{2SIV}, Z_{js}) g_t(\widetilde{\theta}_{2SIV}, Z_{jt}) B^{k_{t,n}}(\mathcal{F}_{js})'(B_t'B_t)^{-1}B^{k_{s,n}}(\mathcal{F}_{is}) & t = s \\ 0 & \text{otherwise} \end{cases}$$

where $\widetilde{\theta}_{2SIV}$ is the estimator from the SIV estimation above.

As mentioned above, we use both the multivariate Fourier and power series for our estimation. Furthermore, we increase the number of approximating basis functions with the sample size. Table 2 reports the results for the Mean, Standard Deviation, and Root Mean Squared Error of the estimates of ρ and β for 100 replications. For the design under investigation, we note that, both the power and Fourier series do a very good job of approximation. As is expected the SMD has much smaller standard deviation that the SIV. In all cases, however,the RMSE of the proposed estimators decreases as sample size increases at rate at least equal to $\sqrt{n}$.

TABLE 2

FINITE-SAMPLE PERFORMANCE OF THE THIRD-STEP ESTIMATOR

($\rho = 0.5$ *and* $\beta = 3$)

Sample Size			500		1000		2000	
			Power Series	Fourier Series	Power Series	Fourier Series	Power Series	Fourier Series
SIV	ρ	Mean	0.493	0.496	0.514	0.499	0.496	0.499
		Std Dev.	0.069	0.038	0.066	0.006	0.066	0.005
		RMSE	0.002	0.002	7.4e-4	1.1e-4	4.4e-4	1.6e-5
	β	Mean	3.026	2.999	3.095	3.000	3.002	3.000
		Std Dev.	0.710	0.019	0.459	0.002	0.020	0.001
		RMSE	0.013	5.8e-5	2.7e-3	5.7e-6	1.5e-5	2.4e-7
SMD	ρ	Mean	0.485	0.485	0.485	0.485	0.478	0.483
		Std Dev.	0.059	0.057	0.045	0.044	0.033	0.037
		RMSE	0.027	0.026	0.015	0.015	0.009	0.009
	β	Mean	2.940	2.956	2.974	2.991	2.968	2.976
		Std Dev.	0.173	0.198	0.136	0.127	0.094	0.101
		RMSE	0.082	0.090	0.043	0.040	0.022	0.023

5.2 Model 2: No Predetermined Variable in the Selection Equation

In order to see how our estimator performs in a small sample study, we compared our SMD estimator to the three estimators proposed in Kyriazidou (2001). Since Kyriazidou's (2001) estimator does not apply to the case in which there are variables in the selection that are predetermined with respect to the outcome we use the following framework for this study.

For $i = 1, \ldots, n$; $t = 1, 2, 3$

$$y^*_{it} = \rho y^*_{it-1} + \alpha_i + \varepsilon^*_{it},$$

$$\mathrm{y}_{it} = \mathrm{d}_{it}\mathrm{y}^*_{it},$$

$$\mathrm{d}_{it} = \mathbb{I}\{\gamma_1 z_{1it} + \eta_i + u_{it} > 0\},$$

where z_1 is strictly exogenous and normally distributed $N(0,1)$. Individual-specific effects as well as the time-varying error terms and initial observations follow the same structure and distributional assumptions as the one used in our model above. Namely $\eta_i = \frac{1}{3}(z_{1i1} + z_{1i2} + z_{1i3}) + 2\xi - 0.75$, where ξ is independently distributed $U(0,1)$, u_{it} is distributed as $N(0, 0.5)$. In the main equation, $\alpha_i = \frac{1}{3}(z_{2i1} + z_{1i2} + z_{2i3})$ and

$\varepsilon_{it}^* = 0.8\xi^* + 0.6\,(u_{it} - E\,(u_{it}))$, where ξ^* is an independent standard normal vector. The initial observation, y_{i0}, is generated as $y_{i0} = d_{i0}y_{i0}^*$, where y_{i0}^* is $N(-2,1)$ and d_{i0} is generated according to a binomial distribution. In both cases, we report the small-sample properties of the structural estimator, $\widehat{\rho}$, for the true selection parameter $\gamma_1 = 1$. Again we consider three sample sizes of 500, 1000, and 2000.

In this version of the Monte Carlo study, we have two moment restrictions associated to

$$g_3\,(\theta_2, Z_{i3}) = d_{i3}d_{i2}[\Delta y_{i3} - \rho\Delta y_{i2} - \Delta\overline{\lambda}\,(\overline{v}_{i3}, \overline{v}_{i2}, \overline{z}_{2i})]$$

and

$$g_2\,(\theta_2, Z_{i2}) = d_{i2}d_{i1}[\Delta y_{i2} - \rho\Delta y_{i1} - \Delta\overline{\lambda}\,(\overline{v}_{i2}, \overline{v}_{i1}, \overline{z}_{2i})],$$

where

$$\begin{aligned}
\overline{v}_{i3} &= \phi_0 d_{i2} + \phi_1 y_{i2} + \gamma_1 z_{1i3} + \gamma_2 z_{2i3},\\
\overline{v}_{i2} &= \phi_0 d_{i1} + \phi_1 y_{i1} + \gamma_1 z_{1i2} + \gamma_2 z_{2i2},\\
\overline{v}_{i1} &= \phi_0 d_{i0} + \phi_1 y_{i0} + \gamma_1 z_{1i1} + \gamma_2 z_{2i1},
\end{aligned}$$

and

$$\overline{z}_{2i} = \frac{1}{3}(z_{2i1} + z_{2i2} + z_{2i3}).$$

The conditioning sets becomes

$$\mathcal{F}_{i2} = \{y_{i0}, d_{i0}, z_{1i1,}z_{1i2}, z_{1i3}, \overline{z}_{2i}, d_{i2}d_{i1} = 1\}$$

and

$$\mathcal{F}_{i3} = \{y_{i0}, y_{i1}, d_{i0}, d_{i1}, z_{1i1,}z_{1i2}, z_{1i3}, \overline{z}_{2i}, d_{i3}d_{i1} = 1\}.$$

Below we specify the kernel weighted moment restrictions used in Kyriazidou (2001):

$$\frac{1}{n}\sum_{i=1}^{n} d_{i0}d_{i1}d_{i2}y_{i0}\,(\Delta y_{i2} - \rho\Delta y_{i1})\,\overline{\omega}_{i2}, \tag{57}$$

$$\frac{1}{n}\sum_{i=1}^{n} d_{i0}d_{i1}d_{i2}d_{i3}y_{i0}\,(\Delta y_{i3} - \rho\Delta y_{i2})\,\overline{\omega}_{i3}, \tag{58}$$

$$\frac{1}{n}\sum_{i=1}^{n} d_{i1}d_{i2}d_{i3}y_{i1}\,(\Delta y_{i3} - \rho\Delta y_{i2})\,\overline{\omega}_{i3}, \tag{59}$$

$$\frac{1}{n}\sum_{i=1}^{n} d_{i0}d_{i1}d_{i2}d_{i3}\,(y_{i3} - \rho y_{i2})\,(\Delta y_{i2} - \rho\Delta y_{i1})\,\overline{\omega}_{i2}, \tag{60}$$

$$\frac{1}{n}\sum_{i=1}^{n} d_{i0}d_{i1}d_{i2}\left[(y_{i2}-\rho y_{i1})^2-(y_{i1}-\rho\Delta y_{i0})^2\right]\overline{\omega}_{i2}, \tag{61}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} d_{i1}d_{i2}d_{i3}\left[(y_{i3}-\rho y_{i2})^2-(y_{i2}-\rho\Delta y_{i1})^2\right]\overline{\omega}_{i3}, \tag{62}$$

where $\overline{\omega}_{ij}$ are the kernel weights constructed using a standard normal kernel with bandwidth parameter $h = n^{-\frac{1}{5}}$. We report here only estimates using the optimal weighting matrix. IV refers to the estimation using only moment restrictions (57), (58) and (59). GMM1 refers to the estimation where moment restriction (60) is added to those used in the IV estimation. Finally, GMM2 refers to the estimation using all the moment restrictions, (57)-(62).

TABLE 3

FINITE-SAMPLE PERFORMANCE OF SIEVE ESTIMATORS

Sample Size		500		1000		2000	
ρ_0= .5		Power Series	Fourier Series	Power Series	Fourier Series	Power Series	Fourier Series
SMD	Mean	0.493	0.467	0.487	0.468	0.486	0.460
	Std Dev.	0.070	0.099	0.057	0.087	0.036	0.050
	RMSE	0.031	0.047	0.018	0.029	0.009	0.014

TABLE 4

FINITE-SAMPLE PERFORMANCE OF WEIGHTED KERNEL ESTIMATORS

Sample Size		500	1000	2000
ρ_0= .5				
IV	Mean	0.5020	0.5017	0.5019
	Std Dev.	0.0844	0.0616	0.0447
	RMSE	0.0755	0.0390	0.0200
GMM1	Mean	0.5032	0.5020	0.5022
	Std Dev.	0.0858	0.0644	0.0457
	RMSE	0.0767	0.0407	0.0204
GMM2	Mean	0.5041	0.5003	0.5004
	Std Dev	0.0880	0.0653	0.0462
	RMSE	0.0787	0.0413	0.0206

As expected, our estimator performs as well if not better than Kyriazidou's (2001) estimators in terms of a smaller RMSE.

6 Conclusion

In this paper we consider the problem of identification and estimation in panel-data sample-selection models with a binary selection rule when the latent equations contain possible predetermined variables, lags of the dependent variables, and unobserved individual effects. The selection equation contains lags of the dependent variables from the latent equations and other possible predetermined variables relative to the latent equations. We derive a set of conditional moment restrictions that are then exploited to construct a three-step sieve extremum estimator for the parameters of the main equation including a nonparametric estimator of the sample-selection term. In the second step, the unknown parameters of the selection equation are consistently estimated using a transformation approach in the spirit of Berkson's minimum chi-square sieve method and a first-step kernel estimator for the selection probability. This second step estimator is of interest in its own right: it can be used to semiparametrically estimate a panel-data binary-response model with a nonparametric individual specific effect without making any other distributional assumptions. We show that both our second- and third-step estimators are $\sqrt{n}$-consistent and asymptotically normal. This has not been previously established for this class of dynamic sample-selection models. Our framework is also more general than the ones previously studied in that ours can estimate equations derived from an intertemporal utility maximization problem, whereas the alternative estimators in the literature cannot. The major limitation of our model is that it imposes an individual specific effect structure similar to Altug and Miller (1998) in the selection equation; however, it is still general enough since we do not make any parametric assumption about the functional form of either the mean component or the distribution function. Even this restriction can be relaxed further at the expense of making a exclusion-independence assumption. Our estimators perform well in a limited Monte Carlo study and does better that Kyriazidou's estimator in small sample.

7 Appendix

In this appendix, the letters c and C will denote diverse constants, not necessarily the same, and diverse occurences.

Proof of Proposition 1. First, note that for $i = 1, ..., n$, $t = 4, ...T$,

$$E[(\triangle y_{it} - \rho \triangle y_{it-1} - \triangle x_{it}\beta - \triangle \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, \overline{v}_{it-2}, z_i^1)) \mid \boldsymbol{F}_{it}] = E[\triangle \varepsilon_{it} \mid \boldsymbol{F}_{it}].$$

From (28),

$$E[\triangle \varepsilon_{it} \mid \boldsymbol{F}_{it}] = E[\varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \mid \boldsymbol{F}_{it}] - E[\varepsilon_{it-1}^* - \overline{\lambda}(\overline{v}_{it-1}, \overline{v}_{it-2}, z_i^1) \mid \boldsymbol{F}_{it}]. \quad (63)$$

Furthermore,

$$E[\varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \mid \digamma_{it}] = E[E[\varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \mid \zeta_{it}, \digamma_{it}] \mid \digamma_{it}] \tag{64}$$

$$= E[E[\varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \mid \zeta_{it}] \mid \digamma_{it}]. \tag{65}$$

The second equality follows from the independence of $\digamma_{it}/\zeta_{it}$ from ε_{it}^* (see Assumptions 2.1–2.2). Finally from (23),

$$E[\varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \mid \zeta_{it}] = 0, \tag{66}$$

which implies that

$$E[\varepsilon_{it}^* - \overline{\lambda}(\overline{v}_{it}, \overline{v}_{it-1}, z_i^1) \mid \digamma_{it}] = 0. \tag{67}$$

A similar reasoning on the second term of the right-hand side of (63) leads to

$$E[\varepsilon_{it-1}^* - \overline{\lambda}(\overline{v}_{it-1}, \overline{v}_{it-2}, z_i^1) \mid \digamma_{it}] = 0. \tag{68}$$

Finally, Equations (67) and (68) prove that (63) is equal to zero. ■

Proof of Theorem 1. First, by Assumption 2.3, $\{x_{it}, z_{it}, \alpha_i, \eta_i, \varepsilon_{it}^*, u_{it}\}_{t=1}^T$ is an i.i.d sample drawn from a distribution that satisfies (1)-(3). We observe a random sample $\{(y_{i0}, d_{i0})\}_{i=1}^n$ from a distribution that satisfies (2) where d_{i0} and y_{i0}^* take values on a subset of the real line. Assumptions 2.4 and 4.4.1 imply that y_{it}^* are i.i.d over individuals. From equation (3) and by Assumptions 2.1-2.2, 4.4.1 and 4.4.2 we obtain that $d_i \equiv \{d_{it}\}_{t=1}^T$ are i.i.d over individuals. Therefore, by definition of $\omega_{it} \equiv (d_{it-1}, y_{it-1}, z_{it}', z_i^1)'$, $\omega_i = \{\omega_{it}\}_{t=1}^T$ are i.i.d..

Let $\theta_1(\triangle d_{-1}, \triangle z, P) \equiv \phi_0 \triangle d_{-1} + \triangle z\gamma - \triangle F^{-1}(P)$, and $\widehat{\theta}_1(\triangle d_{-1}, \triangle z, P) \equiv \widehat{\phi}_0 \triangle d_{-1} + \triangle z\widehat{\gamma} - \widehat{\triangle F^{-1}}(P)$. Let $d^2(\theta_1, \theta_{10}) = \|\theta_1 - \theta_{10}\|^2 = E\{[\phi_0 - \phi_{00}]\triangle d_{-1} + \triangle z[\gamma - \gamma_0] - [\triangle F^{-1}(P_0) - \triangle F_0^{-1}(P_0)]\}^2$. Consistency of θ_1 is established using the following decomposition:

$$\begin{aligned} \left\|\widehat{\theta}_1\left(\widehat{P}\right) - \theta_{10}\right\| &\leq \left\|\widehat{\theta}_1(P_0) - \theta_{10}\right\| + \left\|\widehat{\theta}_1\left(\widehat{P}\right) - \widehat{\theta}_1(P_0)\right\| \\ &\leq A + B. \end{aligned}$$

Rate of Convergence of A.

The rate of convergence of A is obtained by application of Theorem 3-2 in Chen (2005). For clarity, these conditions are recalled here:

(C05)1.1: $Q_1(\theta_1, P_o) \equiv E[\ell(\theta_1, \omega, P_0)]$ *is uniquely maximized on* Θ_1 *at* $\theta_{10} \in \Theta_1$. *(C05)1.2 :* $\Theta_{1n} \subseteq \Theta_{1n+1} \subseteq \cdots \subseteq \Theta_1$ *for all* $n \geq 1$ *and for any* $\theta_1 \in \Theta_1$ *there exists* $\pi_n\theta_1 \in \Theta_{1n}$ *such that* $\|\theta_1 - \pi_n\theta_1\| = o(1)$ *as* n *gets large. (C05)1.3: The criterion* $Q_1(\theta_1, P_0)$ *is continuous in* $\theta_1 \in \Theta_1$ *with respect to d(.). (C05)1.4: The sieve spaces* Θ_{1n} *are compact under d(.). (C05)1.5:* $plim_{n\to\infty} sup_{\theta_1\in\Theta_{1n}} |Q_{1n}(\theta_1, P_0) - Q_1(\theta_1, P_0)| = 0$. *(C05)1.6 :* $\{\omega_i\}_{i=1}^n$ *are i.i.d. (C05)1.7: There is* $C_1 > 0$ *such that for all small* $\varepsilon > 0$, $\sup_{\theta_1\in\Theta_{1n}} V[\ell(\theta_1, \omega, P_0) - \ell(\theta_{10}, \omega, P_0)] \leq C_1\varepsilon^2$. *(C05)1.8: For any* $\delta > 0$, *there exists a constant* $s \in (0, 2)$ *such that*

$$\sup_{\theta_1\in\Theta_{1n}} |\ell(\theta_1, \omega, P_0) - \ell(\theta_{10}, \omega, P_0)| \leq \delta^s U(\omega),$$

where $E([U(\omega)]^\gamma) \leq C_2$ *for some* $\gamma \geq 2$. *(C05)1.9 : Let*

$$\mathcal{F}_n = \{\ell(\theta_1, \omega, P_0) - \ell(\theta_{10}, \omega, P_0) : d(\theta_1, \theta_{10}) \leq \delta, \theta_1 \in \Theta_{1n}\}$$

and for some constant b, *there exists* $\delta_n \in (0, 1)$ *such that*

$$\delta_n = \inf\left\{\delta \in (0,1) : \frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta}\sqrt{H_{[]}(s, \mathcal{F}_n, \|.\|_2)}ds \leq const.\right\}.$$

(C05)1.1 is a direct consequence of Proposition 1. (C05)1.2 is assumed in 4.2.1. (C05)1.3 follows from the properties of scalar products. Assumption 4.2.1 guaranties that the sieve spaces Θ_{1n} are compact, hence (C05)1.4 is satisfied. (C05)1.5 follows from the continuity (C05)1.3, from simple convergence of Q_{1n} and (C05)1.4. (C05)1.6 is implied by Assumption 2.3. The remaining of the proof consists in checking conditions (C05)1.7, (C05)1.8 and (C05)1.9 controlling for the rate of convergence. To simplify notations, we denote by $\triangle F^{-1} = \triangle F_u^{-1}(P_0)$ and $\triangle F_0^{-1} = \triangle F_{u0}^{-1}(P_0)$. Note that

$$\begin{aligned}
&\ell(\theta_1, \omega, P_0) - \ell(\theta_{10}, \omega, P_0)\\
&= -\tfrac{1}{2}\left\{\begin{array}{c}[\triangle y_{-1} - \phi_0\triangle d_{-1} - \triangle z\gamma + \triangle F^{-1}]'[\triangle y_{-1} - \phi_0\triangle d_{-1} - \triangle z\gamma + \triangle F^{-1}]\\ -\left[\triangle y_{-1} - \phi_{00}\triangle d_{-1} - \triangle z\gamma_0 + \triangle F_0^{-1}\right]'\left[\triangle y_{-1} - \phi_{00}\triangle d_{-1} - \triangle z\gamma_0 + \triangle F_0^{-1}\right]\end{array}\right\}\\
&= -\tfrac{1}{2}\left[(\phi_{00} - \phi_0)\triangle d_{-1} + \triangle z(\gamma_0 - \gamma) + \triangle F^{-1} - \triangle F_0^{-1}\right]'\\
&\times\left[2\triangle y_{-1} - (\phi_{00} + \phi_0)\triangle d_{-1} + \triangle z(\gamma_0 + \gamma) + \triangle F^{-1} + \triangle F_0^{-1}\right]\\
&= \left[(\phi_0 - \phi_{00})\triangle d_{-1} + \triangle z(\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1}\right]'\\
&\times\left[(\triangle\upsilon + \triangle\xi) + \tfrac{(\phi_0 - \phi_{00})}{2}\triangle d_{-1} + \triangle z\tfrac{(\gamma - \gamma_0)}{2} + \tfrac{\triangle F_0^{-1} - \triangle F^{-1}}{2}\right],
\end{aligned}$$

and by Assumption 4.2.2,

$$\begin{aligned}
&E[\ell(\theta_1, \omega, P_0) - \ell(\theta_{10}, \omega, P_0)]^2\\
&\leq 2C\|\theta_1 - \theta_{10}\|^2 + E\left[\begin{array}{c}\left(\begin{array}{c}\left((\phi_0 - \phi_{00})\triangle d_{-1} + \triangle z(\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1}\right)'\\ \times\left((\phi_0 - \phi_{00})\triangle d_{-1} + \triangle z(\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1}\right)\end{array}\right)'\\ \times\left(\begin{array}{c}\left((\phi_0 - \phi_{00})\triangle d_{-1} + \triangle z(\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1}\right)'\\ \times\left((\phi_0 - \phi_{00})\triangle d_{-1} + \triangle z(\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1}\right)\end{array}\right)'\end{array}\right].
\end{aligned}$$

Let D denote the second term on the right hand side of this inequality.

$$D \leq \sup_{\theta_1 \in \Theta_{1n}} \left[\begin{array}{c} \left((\phi_0 - \phi_{00}) \triangle d_{-1} + \triangle z (\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1} \right)' \\ \times \left((\phi_0 - \phi_{00}) \triangle d_{-1} + \triangle z (\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1} \right) \end{array} \right]$$
$$\times E \left[\begin{array}{c} \left((\phi_0 - \phi_{00}) \triangle d_{-1} + \triangle z (\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1} \right)' \\ \times \left((\phi_0 - \phi_{00}) \triangle d_{-1} + \triangle z (\gamma - \gamma_0) + \triangle F_0^{-1} - \triangle F^{-1} \right) \end{array} \right].$$

Using Theorem 1 of Gabushin (1967) and Lemma 2 in Chen and Shen (1998), by Assumption 4.2.3 we have for any $p_1 > 0$, $\|\theta_1 - \theta_{10}\|_\infty \leq C \|\theta_1 - \theta_{10}\|^{2p_1/(2p_1+d)}$, where $d = \dim(\omega)$. Therefore, $D \leq C_2 \|\theta_1 - \theta_{10}\|^{2(1+2p_1/(2p_1+d))} \leq C_2 \|\theta_1 - \theta_{10}\|^2$ and (C05)1.7 is satisfied.

Moreover,

$$\begin{aligned} &|\ell(\theta_1, \omega, P_0) - \ell(\theta_{10}, \omega, P_0)| \\ &\qquad \leq \|\theta_1 - \theta_{10}\|_\infty \left[|\triangle \upsilon + \triangle \xi| + \tfrac{1}{2} \left(\|\theta_1\|_\infty + \|\theta_{10}\|_\infty \right) \right] \text{ a.s.} \\ &\qquad \leq C \|\theta_1 - \theta_{10}\|^{2p_1/(2p_1+d)} \left[|\triangle \upsilon + \triangle \xi| + \tfrac{1}{2} \left(\|\theta_1\|_\infty + \|\theta_{10}\|_\infty \right) \right] \text{ a.s.} \end{aligned}$$

Using Lemma 2 in Chen and Shen (1998), we have that (C05)1.8 is also satisfied for $s = 2p_1/(2p_1 + d)$, $\gamma = 2$ and $U(\omega) = |\triangle \upsilon + \triangle \xi| + \frac{1}{2} \left(\|\theta_1\|_\infty + \|\theta_{10}\|_\infty \right)$.

Finally, in order for $\widehat{\theta}_1$ to converge to θ_{10} at a fast rate, not only does the approximation error $d(\theta_0, \pi_n \theta_0)$ have to approach zero suitably fast, the sieve space must not be too complex. More precisely, we denote by $H_{[]}(s, \mathcal{F}_n, \|.\|_2)$ the logarithm of the minimum number of closed intervals denoted by $N_{[]}(s, \mathcal{F}_n, \|.\|_2)$ and of the form $\{f : g \leq f \leq h\}$ that cover $\mathcal{F}_n$ for g and h given such that $\|h - g\|_r \leq s$. In $\mathbb{R}^2$ for example, these intervals are rectangles whose lower left and upper right summits are respectively g and h. Determination of the final convergence rate is obtained by setting $\delta_n \equiv d(\theta_{10}, \pi_n \theta_{10})$. To calculate δ_n from (C05)1.9, an upper bound of $H_{[.]}(s, \mathcal{F}_n, \|.\|_2)$ suffices.

Note that $\|\theta_{10} - \pi_n \theta_{10}\| \leq const. \left\| \Delta F_0^{-1} - \pi_n \Delta F_0^{-1} \right\|_\infty$. By Lemma 2.1 in Ossiander (1987), $H_{[.]}(s, \mathcal{F}_n, \|.\|_2) \leq \log N_{[.]}(s, \mathcal{H}_{1n}, \|.\|_2)$. Let $C = \sqrt{E[U(\omega)]^2}$, then $H_{[.]}(s, \mathcal{F}_n, \|.\|_2) \leq \log N_{[.]}\left(\frac{s}{C}, \mathcal{H}_{1n}, \|.\|_\infty\right)$. By Lorentz (1966), $\left\| \triangle F_0^{-1}(P_0) - \pi_n \triangle F_0^{-1}(P_0) \right\| = O(k_n^{-p_1})$ for $\mathcal{H}_1 \in \Lambda_{c_1}^{p_1}$. By Lemma 2.5 of Van de Geer (2000), $\log N_{[.]}\left(\frac{\omega}{C}, \mathcal{H}_{1n}, \|.\|_\infty\right) \leq const. k_n \log\left(1 + \frac{4c_1}{\omega}\right)$ and δ_n solves

$$\begin{aligned} \frac{1}{\sqrt{n}\delta_n^2} \int_{b\delta_n^2}^{\delta_n} \sqrt{H_{[]}(s, \mathcal{F}_n, \|.\|_2)} ds &\leq \frac{1}{\sqrt{n}\delta_n^2} \sqrt{k_n} \int_{b\delta_n^2}^{\delta_n} \log\left(1 + \frac{4c_1}{s}\right) ds \\ &\leq \frac{1}{\sqrt{n}\delta_n^2} \sqrt{k_n} \delta_n \leq const. \end{aligned}$$

The solution is $\delta_n \asymp \sqrt{\frac{k_n}{n}}$ where $\asymp$ means "bounded above and below". Finally, by Theorem 3.2 in Chen (2005), $\|\theta_1(P_0) - \theta_{10}\| = O_p\left(\max\left(\sqrt{\frac{k_n}{n}}, O(k_n^{-p_1})\right)\right)$ and for

$k_n = O\left(n^{\frac{1}{2p_1+d}}\right)$, $\|\theta_1(P_0) - \theta_{10}\| = O_p\left(n^{-\frac{p_1}{2p_1+d}}\right)$.

Rate of convergence of B.

By Lemma 8-10 in Newey and McFadden (1994) relying on Assumptions 4.1.1-4.1.3, we have $\sqrt{n}\left\|\widehat{p}(\omega) - p_0(\omega)\right\|^2 \to 0$. By Assumptions 4.1.1-4.1.4, Newey (1994) shows that

$$\|\widehat{p} - p_0\|_S = O_p\left((\ln n)^{\frac{1}{2}}\left(nh^{d+2q}\right)^{-\frac{1}{2}} + h^L\right).$$

Hence, we obtain that $\left\|\widehat{P} - P_0\right\|_S = O_p\left((\ln n)^{\frac{1}{2}}\left(nh^{d+2q}\right)^{-\frac{1}{2}} + h^L\right)$, where $\|.\|_S$ is the Sobolev norm defined as $\|P\|_S \equiv \max_{l\leq q}\sup_{\omega\in\mathcal{W}}\left\|\frac{\partial P(\omega)}{\partial\omega^l}\right\|$.

Finally, using the fact that π_n are uniformly Lipschitzian, we have $\left\|\widehat{\theta}_1\left(\widehat{P}\right) - \widehat{\theta}_1(P_0)\right\| = \left\|\pi_n \triangle F^{-1}(\widehat{P}) - \pi_n \triangle F^{-1}(P_0)\right\| \leq C\left\|\widehat{P} - P_0\right\|_S^{p_1}$.

Hence, if $\left\|P - \widehat{P}\right\|_S = O(k_n^{-1})$, then $\left\|\pi_n \triangle F^{-1}(\widehat{P}) - \pi_n \triangle F^{-1}(P)\right\| = O(k_n^{-p_1})$ and $\left\|\widehat{\theta}_1\left(\widehat{P}\right) - \widehat{\theta}_1(P_0)\right\| = O_p\left(n^{-\frac{p_1}{2p_1+d}}\right)$. Recall that $\left\|\widehat{P} - P_0\right\|_S = O_p\left((\ln n)^{\frac{1}{2}}\left(nh^{d+2q}\right)^{-\frac{1}{2}} + h^L\right)$ and note that $(\ln n)^{\frac{1}{2}}\left(nh^{d+2q}\right)^{-\frac{1}{2}} + h^L \leq (\ln n)^{\frac{1}{2}}\left[\left(nh^{d+2q}\right)^{-\frac{1}{2}} + h^L\right]$.

We choose the optimal bandwith h^* such that $\left(nh^{d+2q}\right)^{-\frac{1}{2}} = h^L$, that is $h^* = n^{-\frac{1}{2L+d+2q)}}$. Finally, $\left(n(h^*)^{d+2q}\right)^{-\frac{1}{2}} + (h^*)^L = n^{-\frac{L}{2L+d+2q}}$, so that $\left\|P - \widehat{P}\right\| = O_p\left((\ln n)^{\frac{1}{2}} n^{-\frac{L}{2L+d+2q)}}\right)$ and $\left\|\widehat{\theta}_1\left(\widehat{P}\right) - \widehat{\theta}_1(P_0)\right\| = O_p\left((\ln n)^{\frac{p_1}{2}} n^{-\frac{L}{2L+d+2q)}}\right)$.

Finally, we can choose q and L large enough such that $\frac{L}{2L+d+2q)} > \frac{p_1}{2p_1+d}$ and $\left\|\widehat{\theta}_1\left(\widehat{P}\right) - \theta_0\right\| = O_p\left(n^{-\frac{p_1}{2p_1+d}}\right)$. ■

Proof of Theorem 2. All computations being computed at the true value of the selection probability P_0, we will define $l(\theta_1, \omega, P_0) \equiv l(\theta_1, \omega)$ for the remaining of the proof. For all $\theta_1 \in \Theta_1$ and all ω, there exists $l'_{\theta_0}[\theta - \theta_0, \omega]$ such that the remainder in the linear approximation is

$$r(\theta_1 - \theta_{10}, \omega) = l(\theta_1, \omega) - l(\theta_{10}, \omega) - l'_{\theta_0}[\theta - \theta_0, \omega],$$

where $l'_{\theta_0}[\theta - \theta_0, \omega] = \lim_{t\to 0}[l(\theta_1(\theta_{10}, t), \omega) - l(\theta_{10}, \omega)]/t$ is the pathwise derivative of l at θ_0 and $\theta_1(\theta_{10}, t) \in \Theta_1$ is a path in t connecting θ_{10} to θ_1 such that $\theta_1(\theta_{10}, 0) = \theta_{10}$ and $\theta_1(\theta_{10}, 1) = \theta_1$. Suppose that L_2 norm defined above induces an inner product $\langle ., . \rangle$ on the completion of the space spanned by $\Theta_1 - \theta_{10}$ denoted V. Let ε_n denote any sequence satisfying $\varepsilon_n = o\left(n^{-\frac{1}{2}}\right)$ and let $\nu_n(g(\omega)) = \frac{1}{n}\sum_{i=1}^{n}(g(\omega_i) - E_0(g(\omega_i)))$ denote the empirical

process indexed by the function $g(.)$. Let $K(\theta_{10},\theta_1) \equiv n^{-1}\sum_{i=1}^{n} E_0\left[l(\theta_{10},\omega_i) - l(\theta_1,\omega_i)\right]$.

The proof consists in checking the following conditions (see Chen, 2005).

(C05)2.1: Suppose the functional of interest f has the following smoothness properties, (i) there is a $\alpha > 0$ such that $\left|f(\theta_1) - f(\theta_{10}) - f'_{\theta_{10}}[\theta_1 - \theta_{10}]\right| = O\left(\|\theta_1 - \theta_{10}\|^{\alpha}\right)$ uniformly in $\theta_1 \in \Theta_{1n}$ with $\|\theta_1 - \theta_{10}\| = o(1)$; (ii) $\sup_{\{\theta_1\in\Theta_{1n}:\|\theta_1-\theta_{10}\|>0\}} \frac{\left|f'_{\theta_{10}}[\theta_1-\theta_{10}]\right|}{\|\theta_1-\theta_{10}\|} < \infty$; (iii) there is $\pi_n v^* \in \Theta_{1n}$ such that $\|\pi_n v^* - v^*\| \left\|\widehat{\theta}_1 - \theta_{10}\right\| = o_p\left(n^{-\frac{1}{2}}\right)$. (C05)2.2:

$$\sup_{\{\theta\in\Theta_n:\|\theta_1-\theta_{10}\|\leq\delta_n\}} \nu_n\left(l(\theta_1,\omega) - l\left(\left(\theta_1 \pm \varepsilon_n\pi_n v^*\right),\omega\right) - l'_{\theta_{10}}([\pm\varepsilon_n\pi_n v^*,\omega]\right) = O_p\left(\varepsilon_n^2\right).$$

(C05)2.3: $K\left(\theta_{10},\widehat{\theta}_1\right) - K\left(\theta_{10},\widehat{\theta}_1 \pm \varepsilon_n\pi_n v^*\right) = \pm\varepsilon_n\left\langle\widehat{\theta}_1 - \theta_{10},\pi_n v^*\right\rangle + o\left(n^{-1}\right)$. *(C05)2.4: (i)* $\nu_n\left(l'_{\theta_{10}}[\pi_n v^* - v^*,\omega]\right) = o_p\left(n^{-\frac{1}{2}}\right)$ *; (ii)* $E[l'_{\theta_{10}}[\pi_n v^*,\omega]] = o\left(n^{-\frac{1}{2}}\right)$. *(C05)2.5:* $n^{\frac{1}{2}}\nu_n\left(l'_{\theta_{10}}[v^*]\right) \to N\left(0,\sigma^2_{v^*}\right)$ *where* $\sigma^2_{v^*} \equiv Var_0\left(l'_{\theta_{10}}[v^*]\right) > 0$ *for i.i.d data.*

By the Riesz representation theorem, there exists $v^* \in V$ such that for any $\theta_1 \in \Theta_1$

$$f'_{\theta_{10}}[\theta_1 - \theta_{10}] = \langle\theta_1 - \theta_{10}, v^*\rangle$$

if and only if $\left\|g'_{\theta_{10}}\right\| < \infty$.

Let $f(\theta_1) = \lambda'\begin{bmatrix}\phi_0\\ \gamma\end{bmatrix}$ where λ is a unit vector in $\mathbb{R}^{d_{\phi_0+\gamma}}$. Then, (C05)2.1 is satisfied for any arbitrary large α since

$$\left|f(\theta_1) - f(\theta_{10}) - f'_{\theta_0}[\theta_1 - \theta_{10}]\right| = \lambda'\begin{bmatrix}\phi_0\\ \gamma\end{bmatrix} - \lambda'\begin{bmatrix}\phi_{00}\\ \gamma_0\end{bmatrix} - \begin{bmatrix}\phi_0 - \phi_{00}\\ \gamma - \gamma_0\end{bmatrix}'\lambda = 0.$$

Moreover,

$$\begin{aligned}
\left\|f'_{\theta_{10}}\right\|^2 &= \sup_{\{\theta_1\in\Theta_{1n}:\|\theta_1-\theta_{10}\|>0\}} \frac{\left\{\begin{pmatrix}\phi_0 - \phi_{00}\\ \gamma - \gamma_0\end{pmatrix}'\lambda\right\}^2}{\|\theta_1 - \theta_{10}\|^2} \\
&= \sup_{\left\{(b,u_1):\left\|b'\begin{bmatrix}\triangle d_{-1}\\ \triangle z\end{bmatrix} - u_1(P_i)\right\|>0\right\}} \frac{\{b'\lambda\}^2}{b'E\left[\left(\begin{bmatrix}\triangle d_{-1}\\ \triangle z\end{bmatrix} - u_1(P)\right)\left(\begin{bmatrix}\triangle d_{-1}\\ \triangle z\end{bmatrix} - u_1(P)\right)'\right]b} \\
&= \lambda' E\left[D_{u_1^*}(\omega)' D_{u_1^*}(\omega)\right]^{-1}\lambda = \lambda'\Sigma_*^{-1}\lambda,
\end{aligned}$$

where $D_{u_1^*}(\omega) \equiv \left[\begin{array}{c} \triangle d_{-1} \\ \triangle z \end{array} \right] - u_1(P)$ solves

$$\inf_{\left(u_1 : \left\| \left[\begin{array}{c} \triangle d_{-1} \\ \triangle z \end{array} \right] - u_1(P) \right\| > 0\right)} E\left[\left(\left[\begin{array}{c} \triangle d_{-1} \\ \triangle z \end{array} \right] - u_1(P) \right) \left(\left[\begin{array}{c} \triangle d_{-1} \\ \triangle z \end{array} \right] - u_1(P) \right)' \right].$$

Using the definition of the norm above, we have

$$\left(\nu^*_{(\phi_0,\gamma)}, \nu^*_F \right) = \left(\Sigma_*^{-1}\lambda, -u_1^*\Sigma_*^{-1}\lambda \right).$$

Assumption 4.3 ensures that the denominator is different from zero, hence that $\left\| f'_{\theta_{10}} \right\|$ is bounded and (C05)2.1 (ii) is satisfied. Since $\left\| \widehat{\theta}_1 - \theta_{10} \right\| = O\left(n^{-\frac{p_1}{2p_1+d}} \right)$, therefore, by Assumption 4.3.3, (C05)2.1(iii) is satisfied. In order to check (C05)2.2, note that

$$l\left(\theta_1, \omega\right) - l(\theta_1 \pm \varepsilon_n \pi_n v^*, \omega) - l'_{\theta_{10}}\left[\pm \varepsilon_n \pi_n v^*, \omega\right] = r\left(\theta_1 - \theta_{10}, \omega\right) - r\left(\theta_1 \pm \varepsilon_n \pi_n v^* - \theta_{10}, \omega\right),$$

with

$$r\left(\theta_1 - \theta_{10}, \omega\right) = \left(\left(\phi_0 - \phi_{00}\right) \triangle d_{-1} + \triangle z \left(\gamma - \gamma_0\right) - \triangle F^{-1}(P_0) + \triangle F_0^{-1}(P_0) \right)^2$$

and

$$r\left(\theta_1 \pm \varepsilon_n \pi_n v^* - \theta_{10}, \omega\right) = \left(\begin{array}{c} \left(\phi_0 - \phi_{00}\right) \triangle d_{-1} + \triangle z \left(\gamma - \gamma_0\right) \\ -\left(\triangle F^{-1} - \triangle F_0^{-1} \right) \pm \varepsilon_n \pi_n v^* \end{array} \right)^2.$$

Therefore,

$$l\left(\theta_1, \omega\right) - l(\theta_1 \pm \varepsilon_n \pi_n v^*, \omega) - l'_{\theta_{10}}\left[\pm \varepsilon_n \pi_n v^*, \omega\right] = -\frac{1}{2}\left[\mp 2 \left(\theta_1 - \theta_{10}\right) \varepsilon_n \pi_n v^* - \left(\varepsilon_n \pi_n v^*\right)^2 \right].$$

Finally, let $S_n = \left\{ \left(\theta_1 - \theta_{10}\right) v^* : \left\| \theta_1 - \theta_{10} \right\| \leq \delta_n, \theta_1 \in \Theta_1 \right\}$. It follows from Kolmogorov and Tihomirov (1961) and by Lemma 4 of Shen and Wong (1994) that the convergence rate of the empirical process

$$\sup_{\left\{\theta_1 \in \Theta_1 : \left\| \theta_1 - \theta_{10} \right\| \leq \delta_n \right\}} n^{-\frac{1}{2}} \upsilon_n \left(\left(\theta_1 - \theta_{10}\right) v^* \right)$$

is of order $O_p\left(n^{-\frac{2p_1}{2p_1+d}} \right)$, and (C05)2.2 holds for $p_1 > \frac{d}{2}$.

Note that $\ell(\theta_1, \omega) - \ell(\theta_{10}, \omega) = \left[\left(\phi_0 - \phi_{00}\right) \triangle d_{-1} + \triangle z \left(\gamma - \gamma_0\right) + \triangle F_0^{-1} - \triangle F^{-1} \right]'$
$\times \left[\left(\triangle \upsilon + \triangle \xi \right) + \frac{\left(\phi_0 - \phi_{00}\right)}{2} \triangle d_{-1} + \triangle z \frac{\left(\gamma - \gamma_0\right)}{2} + \frac{\triangle F_0^{-1} - \triangle F^{-1}}{2} \right]$.

Therefore, $E\left[\ell(\theta_1,\omega)-\ell(\theta_{10},\omega)\right]=\frac{1}{2}\left\|\theta_1-\theta_{10}\right\|^2=K\left(\theta_{10},\theta_1\right)$ and

$$\begin{aligned}
& K\left(\theta_{10},\widehat{\theta}_1\right)-K\left(\theta_{10},\widehat{\theta}_1\pm\varepsilon_n\pi_n v^*\right) \\
= & \frac{1}{2}\left\langle\theta_{10}-\widehat{\theta}_1,\theta_{10}-\widehat{\theta}_1\right\rangle-\frac{1}{2}\left\langle\theta_{10}-\widehat{\theta}_1\pm\varepsilon_n\pi_n v^*,\theta_{10}-\widehat{\theta}_1\pm\varepsilon_n\pi_n v^*\right\rangle \\
= & \pm\frac{1}{2}\left(2\varepsilon_n\left\langle\theta_{10}-\widehat{\theta}_1,\pi_n v^*\right\rangle\right)-\varepsilon_n^2\left\langle\pi_n v^*,\pi_n v^*\right\rangle \\
= & \pm\varepsilon_n\left\langle\theta_{10}-\widehat{\theta}_1,\pi_n v^*\right\rangle+o_p\left(n^{-1}\right),
\end{aligned}$$

and (C05)2.3 is satisfied. Furthermore,

$$l'_{\theta_{10}}\left[\theta_1-\theta_{10},\omega\right]=\left(\triangle\upsilon+\triangle\xi\right)\left(\left(\phi_0-\phi_{00}\right)\triangle d_{-1}+\triangle z\left(\gamma-\gamma_0\right)+\triangle F_0^{-1}-\triangle F^{-1}\right).$$

Hence the expectation of this term is equal to zero and (C05)2.4 (ii) is satisfied.

Moreover,

$$\frac{1}{n}\sum_{i=1}^{n}l'_{\theta_{10}}\left[\pi_n v^*-v^*,\omega_i\right]=\frac{1}{n}\sum_{i=1}^{n}\left(\triangle\upsilon_i+\triangle\xi_i\right)\left\{\pi_n v_F^*-v_F^*\right\}.$$

Using Chebyshev's inequality, for any real number α,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}l'_{\theta_{10}}\left[\pi_n v^*-v^*,\omega_i\right]\right|>n^{\alpha}\right)\leq\frac{\sum_{i=1}^{n}\left\{\pi_n v_F^*-v_F^*\right\}^2V\left(\triangle\upsilon_i+\triangle\xi_i|\omega\right)}{n^{2\alpha+2}}.$$

Therefore, by Assumption 4.2.2, (C05)2.4(i) is satisfied for $\alpha=-\frac{1}{4}$.

We know that

$$l'_{\theta_{10}}\left[\pi_n v^*,\omega\right]=\left(\triangle\upsilon+\triangle\xi\right)\left\{\begin{pmatrix}\triangle d_{-1}\\ \triangle z\end{pmatrix}\nu^*_{(\phi_0,\gamma)}+\pi_n v_F^*\left(P\right)\right\}.$$

Hence, (C05)4.2(ii) is trivially satisfied.Using the expression above for $\nu^*_{(\phi_0,\gamma)}$ and $v_F^*\left(P\right)$, we obtain that

$$\left(\nu^*_{(\phi_0,\gamma)},\nu_F^*\right)=\left(\Sigma_*^{-1}\lambda,-u_1^*\Sigma_*^{-1}\lambda\right)$$

and

$$\begin{aligned}
l'_{\theta_{10}}\left[v^*,\omega\right] &= \left(\triangle\upsilon+\triangle\xi\right)\left\{\begin{pmatrix}\triangle d_{-1}\\ \triangle z\end{pmatrix}\Sigma_*^{-1}\lambda-u_1^*\Sigma_*^{-1}\lambda\right\} \\
&= \left(\triangle\upsilon+\triangle\xi\right)D_{u_1^*}\left(\omega\right)\Sigma_*^{-1}\lambda.
\end{aligned}$$

(C05)2.5 is satisfied under Assumption 4.3.1. Finally,

$$\begin{aligned} V\left(l'_{\theta_{10}}[v^*,\omega]\right) &= E\left(\lambda'\Sigma_*^{-1}D_{u_1^*}(\omega)'V\left(\triangle v+\triangle\xi|\omega\right)D_{u_1^*}(\omega)\Sigma_*^{-1}\lambda\right), \\ &= \lambda' E\left[D_{u_1^*}(\omega)'D_{u_1^*}(\omega)\right]^{-1}E\left(D_{u_1^*}(\omega)'V\left(\triangle v+\triangle\xi|\omega\right)D_{u_1^*}(\omega)\right) \\ &\quad \times E\left[D_{u_1^*}(\omega)'D_{u_1^*}(\omega)\right]^{-1}\lambda > 0. \end{aligned}$$

The result follows from Theorem 1 and a direct application of Theorem 4.2 in Chen (2005). ■

Proof of Theorem 3. As for the selection equation, we decompose the proof in successive stages:

$$\begin{aligned} \left\|\widehat{\theta}_2\left(\widehat{\theta}_1,\widehat{\Sigma}_o(\widetilde{\theta}_2)\right)-\theta_{20}\right\| &\leq \left\|\widehat{\theta}_2\left(\theta_{10},\Sigma_0(\theta_{20})\right)-\theta_{20}\right\| \\ &\quad +\left\|\widehat{\theta}_2\left(\widehat{\theta}_1,\Sigma_o(\theta_{20})\right)-\widehat{\theta}_2\left(\theta_{10},\Sigma_0\left(\theta_{20}\right)\right)\right\| \\ &\quad +\left\|\widehat{\theta}_2\left(\widehat{\theta}_1,\widehat{\Sigma}_o(\widetilde{\theta}_2)\right)-\widehat{\theta}_2\left(\widehat{\theta}_1,\Sigma_0\left(\widetilde{\theta}_2\right)\right)\right\| \\ &\quad +\left\|\widehat{\theta}_2\left(\widehat{\theta}_1,\Sigma_0\left(\widetilde{\theta}_2\right)\right)-\widehat{\theta}_2\left(\widehat{\theta}_1,\Sigma_0\left(\theta_{20}\right)\right)\right\|, \\ &\leq A'+B'+C'+D'. \end{aligned}$$

We first find the rate of convergence of A' and $\widetilde{\theta}_2$. This is done by verifying Assumptions 3.1-3.9 necessary to apply Theorem 3.1 of Ai and Chen (2003) for the respective weighting matrices $\Sigma_0(\theta_{20})$ and the identity I. For clarity, we recall these assumptions in the context of our model. We then show that the set of primitive Assumptions 4.1-4.11 are sufficient for these conditions to hold. Let us introduce the following norm:

$$\|\theta_2\|_s = \|(\rho,\beta)'\|_E + \sup_{v\in\mathcal{R}^3}\left|\overline{\triangle\lambda}(v)\times\left(1+\|v\|_E^2\right)^{-\frac{a}{2}}\right| \text{ or some } a > p_2. \tag{69}$$

(AC03)3.1: (i)The data $\{Z_i,F_i\}_{i=1}^n$ are i.i.d.; (ii) the support of F, $\Im$ is compact with a nonempty interior; (iii) the density of F is bounded and bounded away from zero; (AC03)3.2: (i) The smallest and largest eigenvalues of $E\{B^{k_{s,n}}(F)B^{k_{s,n}}(F)'\}$ are bounded and bounded away from zero for all $k_{s,n}$; (ii) for any $f(.)$ with $E[f(F)^2]<\infty$, there exists a $B^{k_{s,n}}(F)'\pi$ such that $E\left\{\left[f(F)-B^{k_{s,n}}(F)'\pi\right]^2\right\}=o(1)$; (iii) for any $f(.)\in\Lambda_c^\gamma(\Im)$ with $\gamma>d_F/2$, there exists $B^{k_{s,n}}(F)'\pi\in\Lambda_c^\gamma(\Im)$ such that

$$\underset{F\in\Im}{Sup}\left|f(F)-B^{k_{s,n}}(F)'\pi\right| = O(k_{s,n}^{-\gamma/d_F})$$

and $k_{s,n}^{-\gamma/d_F}=o(n^{-\frac{1}{4}})$. (AC03)3.3: $\theta_{20}\in\Theta_2$ is the only $\theta_2\in\Theta_2$ satisfying $E[g(\theta_2,Z)|F]=0$. (AC03)3.4: (i) $\Sigma_o(F)$ is finite positive definite uniformly over $F\in\Im$; (ii) $\widehat{\Sigma}_o(F)=$

$\Sigma_o(F) + o_p(n^{-\frac{1}{4}})$ *uniformly over* $F \in \Im$. *(AC03)3.5: (i) There is a metric* $\|.\|_s$ *such that* $\Theta_2 \equiv \Lambda_2 \times H_2$ *is compact under* $\|.\|_s$*; (ii) for any* $\theta_2 \in \Theta_2$*, there exists* $\pi_n\theta_2 \in \Theta_{2n} \equiv \Lambda_2 \times H_{2n}$ *with* $\|\pi_n\theta_2 - \theta_2\|_s = o(1)$*; (iii) there is a constant* $\mu_1 > 0$ *such that for any* $\theta_2 \in \Theta_2$*, there exists* $\pi_n\theta_2 \in \Theta_{2n} \equiv \Lambda_2 \times H_{2n}$ *such that* $\|\pi_n\theta_2 - \theta_2\|_s = O(k_{2n}^{-\mu_1})$*, and* $k_{2n}^{-\mu_1} = o(n^{-\frac{1}{4}})$*. (AC03)3.6: (i)* $E[|g(\theta_{20}, Z)|^2 \,||F]$ *is bounded; (ii)* $g(\theta_2, Z)$ *is Hölder continuous in* $\alpha \in A$ *; (iii) each element of* $g(\theta_2, Z)$ *satisfies an envelope condition over* $\theta_2 \in \Theta_{2n}$ *that is, there exists a measurable function* $c(Z)$ *with* $E[c(Z)^4] < \infty$ *such that* $|g(\theta_2, Z)| \leq c(Z)$ *for all* Z *and* $\theta_2 \in \Theta_{2n}$*; (iv) each element of* $m(., \theta_2) \equiv E(g(., \theta_2)|F) \in \Lambda_c^\gamma(\Im)$ *with* $\gamma > d_F/2$ *for all* $\theta_2 \in \Theta_{2n}$*. (AC03)3.7: (i)* $d_g k_{s,n} \geq d_{1+\dim(x)} + k_{2n}$*,* $k_{2n} \to \infty$ *and* $k_{s,n}/n \to 0$*. Let* $\xi_{0n} \equiv \sup_{F \in \Im} \left\|B^{k_{s,n}}(F)\right\|_E$*, which is nondecreasing in* $k_{s,n}$*. Denote* $N(r, \Theta_2, \|.\|_s)$ *as the minimal number of radius* r *covering balls of* H_{2n} *under* $\|.\|_s$*. (ii)* $k_{2n} \times \ln n \times \xi_{0n}^2 \times n^{-\frac{1}{2}} = o(1)$*. (AC03)3.8:* $\ln[N(\varepsilon^{1/\kappa}, \Theta_{2n}, \|.\|_s)] \leq const. \times k_{2n} \times \ln(k_{2n}/\varepsilon)$*. (AC03)3.9: (i)* Θ_2 *is convex in* θ_{20}*, and* $g(\theta_2, Z)$ *is pathwise differentable at* θ_{20}*; (ii) for some* C_1*,* $C_2 > 0$

$$C_1 E\{g(F, \theta_2)'\Sigma(F)^{-1}g(F, \theta_2)\} \leq \|\theta_2 - \theta_{20}\|^2 \leq C_2 E\{g(F, \theta_2)'\Sigma(F)^{-1}g(F, \theta_2)\}$$

for all $\theta_2 \in \Theta_{2n}$ with $\|\theta_2 - \theta_{20}\|_s = o(1)$.

As is standard we need to show that the criterion function is continuous in the first step estimator. This is done by adding a boundedness condition in 4.4.7 that ensures that our moment conditions are continuous in the preliminary estimates of the selection equation. Given the definition of Z_{is} and F_{is}, the data $\{Z_i, F_i\}_{i=1}^n$ are i.i.d (see proof of Theorem 1) and (AC03)3.1(i) is satisfied. Assumptions 4.4.2, 4.4.3, 4.4.4 directly imply that the support of $\{F_i\}_{i=1}^n$ is compact with a nonempty interior and (AC03)3.1 (ii) is satisfied. (AC03)3.1(iii) is directly assumed in Assumption 4.4.4. (AC03)3.2(i) is directly assumed by Assumption 4.4.7. Note that (AC03)3.2(i) is satisfied with the linear sieves satisfying Assumption 4.4.5 (see Newey (1997) for details). Note that 3.2 (ii) is implied by Assumptions 4.4.5, while (AC03)3.2(iii) is implied by Assumptions 4.4.5 and 4.4.6 (ii). Propositions 2 and 3 ensure that θ_2 is identified, hence (AC03)3.3 is satisfied. Note that (AC03)3.4(ii) and (iii) are trivally satisfied with an identity weighting matrix or for the true value Σ_0. Otherwise Assumptions 4.4.8 (ii) and (iii) are sufficient to satisfy this condition. By Assumptions 4.4.3 and 4.4.9, Θ_2 is compact under the norm $\|\theta_2\|_s = \|(\rho, \beta)\|_E + \sup_{v \in \mathcal{R}^{3+\dim(z^1)}} \left|\triangle\overline{\lambda}(v) \times (1 + \|v\|_E^2)^{-\frac{a}{2}}\right|$ for some $a > p_2$. From Chen, Hansen and ScheinKman (1997) for any $\theta_2 \in \Theta_2$, $\|\pi_n\theta_2 - \theta_2\|_s = \left\|\pi_n\triangle\overline{\lambda} - \triangle\overline{\lambda}\right\|_{\infty,\varpi} = \sup_v \left|\left[\pi_n\triangle\overline{\lambda}(v) - \triangle\overline{\lambda}(v)\right]\varpi(v)\right| \leq C(k_{2n})^{-\frac{p_2}{3+\dim(z^1)}}$, where $\varpi(v) = (1 + \|v\|_E^2)^{-\frac{a}{2}}$ for some $a > p_2$. Hence (AC03)3.5(i) is satisfied by Assumption 4.4.6. By the same argument (AC03)3.5(iii) is satisfied with $\mu_1 = p_2/(3 + \dim(z^1))$. (AC03)3.6(i) is satisfied by As-

sumptions 4.4.2, 4.4.3 and 4.4.8 while for any θ_{21},$\theta_{22} \in \Theta_2$,

$$\begin{aligned} |g(\theta_{21}, Z) - g(\theta_{22}, Z)| &\leq |dd_{-1}\triangle y_{-1}(\rho_1 - \rho_2)| + |dd_{-1}\triangle x(\beta_1 - \beta_2)| \\ &\quad + \left|dd_{-1}(\triangle\overline{\lambda}_1(v) - \triangle\overline{\lambda}_2(v))\right| \\ &\leq \|dd_{-1}\triangle y_{-1}\|_E \|\rho_1 - \rho_2\|_E + \|dd_{-1}\triangle x\|_E \|\beta_1 - \beta_2\|_E \\ &\quad + [\varpi(v)]^{-1} \|dd_{-1}\|_\infty \left\|\triangle\overline{\lambda}_1 - \triangle\overline{\lambda}_2\right\|_{\infty,\varpi} \\ &\leq C(Z) \|\theta_{21} - \theta_{22}\|_s , \end{aligned}$$

where $C(Z) = \|dd_{-1}\triangle y_{-1}\|_E + \|dd_{-1}\triangle x\|_E + [\varpi(v)]^{-1} \|dd_{-1}\|_\infty$ and $E[C^2(Z)|\boldsymbol{F}] < \infty$. Therefore, (AC03)3.6(ii) for $\kappa = 1$ by Assumptions 4.4.2, 4.4.9 and 4.4.11(ii). For any $\theta_2 \in \Theta_{2n} \equiv \Theta \times \mathcal{H}_{2n}$, with $\mathcal{H}_{2n}$,being the sieve space for the tensor product of Fourier or power series of the form

$$\mathcal{H}_{2n} = \left\{ \begin{array}{c} \triangle\overline{\lambda}(\overline{v}, \overline{v}_{-1}, \overline{v}_{-2}, z^1) = q^{k_{2n}}(\overline{v}, \overline{v}_{-1}, \overline{v}_{-2}, z^1)'\delta \\ \text{for all } \delta \text{ satisfying } \left\|\triangle\overline{\lambda}\right\|_{\Lambda^{p_2}} \leq c \end{array} \right\}, \tag{70}$$

we have

$$\begin{aligned} \sup_{\theta_2 \in \Theta_{2n}} |g(Z, \theta_2)| &\leq |dd_{-1}\triangle y| + |dd_{-1}\triangle y_{-1}| + \sup_{\beta \in \Lambda_2} |dd_{-1}\triangle x\beta| \\ &\quad + |dd_{-1}| \sup_{\triangle\overline{\lambda} \in \mathcal{H}_n} \left|\triangle\overline{\lambda}(v)\right| \leq C(Z) \end{aligned}$$

with $C(Z) = |\triangle y| + |\triangle y_{-1}| + \sup_\beta |\triangle x\beta| + \sup_{\triangle\overline{\lambda} \in \mathcal{H}_n} \left|\triangle\overline{\lambda}(v)\right|$. Hence (AC03)3.6(iii) is satisfied by Assumptions 2.2, 4.4.11(i) and 4.4.2-4.4.4. Note that

$$\begin{aligned} m(\boldsymbol{F}, \theta_2) &= E[dd_{-1}\triangle y \mid \boldsymbol{F}] - \rho E[dd_{-1}\triangle y_{-1} \mid \boldsymbol{F}] \\ &\quad - E[dd_{-1}\triangle x|\boldsymbol{F}]\beta - E[dd_{-1}\triangle\overline{\lambda}(v)|\boldsymbol{F}]. \end{aligned}$$

Therefore (AC03)3.6(iv) is satisfied by Assumption 4.4.10. The first part of (AC03)3.7(i) directly assumed by Assumption 4.4.6(i), (AC03)3.7(ii) is trivially implied by Assumption 4.4.5 and 4.4.6(ii). Consider the case of power series. From Newey (1997), if a) $var(g(\theta_{20}, Z)|\boldsymbol{F}]$ is bounded; b) The support of $\boldsymbol{F}$ is compact with nonempty interior with probability density bounded away from zero and c) $m(\boldsymbol{F}, \theta) \in \Lambda_c^\gamma(\boldsymbol{F})$, then $\xi_{0n} \leq k_{s,n}$. Note that a) is satisfied by Assumption 4.4.8, b) was shown above when proving (AC03)3.1(ii) while c) was just verified by (AC03)3.6(ii). We now have

$$k_{2n} \times \ln n \times \xi_{0n}^2 \times n^{-\frac{1}{2}} \leq k_{2n} \times \ln n \times k_{s,n}^2 \times n^{-\frac{1}{2}} = o(1)$$

where the final equality comes from Assumption 4.4.5. In the case of Fourier Series (see Ai and Chen (2003) pp. 1807), $\xi_{0n} = k_{s,n}^{\frac{1}{2}}$, therefore

$$k_{2n} \times \ln n \times \xi_{0n}^2 \times n^{-\frac{1}{2}} = k_{2n} \times \ln n \times k_{s,n} \times n^{-\frac{1}{2}} = o(1)$$

where the last equality comes from Assumption 4.4.5. From Shen and Chen (1998) and Shen and Wong (1994), for $\kappa = 1$ and the linear sieves space in Assumption 4.4.5,

$$\ln[N(\varepsilon, \Theta_{2n}, \|.\|_s)] \leq const. \times \ln(1/\varepsilon) \leq const. \times k_{2n} \times \ln(k_{2n}/\varepsilon)$$

Note that $g(\theta_2, Z)$ is linear and continuous in θ_{20}, hence differenciable. As noted in Ai and Chen (2003), when $g(Z, \theta_2)$ is linear in θ_2 then $\|\theta_2 - \theta_{20}\|^2 = E\{m(\digamma, \theta_2)'\Sigma(X)^{-1}m(\digamma, \theta_2)\}$. Assumption 3.9(ii) is satisfied and by Theorem 3.1 in Ai and Chen (2003), we have $\|\theta_2(\Sigma_0(\theta_{20})) - \theta_{20}\| = o_p\left(n^{-\frac{1}{4}}\right)$, $\left\|\widetilde{\theta}_2(I) - \theta_{20}\right\| = o_p\left(n^{-\frac{1}{4}}\right)$. By Assumption 4.4.1-4.4.8 and application of Lemma A1(A and B) in Ai and Chen (2003), we obtain

$$\widehat{\Sigma}_o(\digamma, \theta_2) = \Sigma_o(\digamma, \theta_2) + o_p(n^{-\frac{1}{4}}) \text{ uniformly over } \digamma \in \Im, \theta_2 \in \mathcal{N}_{on}.$$

Finally for $\Theta_{2n} \subseteq \Theta_{2n+1} \subseteq \ldots \subseteq \Theta_2$ with $\Theta_2 \in \Lambda_{c_2}^{p_2}$, we obtain

$$\begin{aligned}\left\|\widehat{\theta}_2\left(\widehat{\theta}_1, \widehat{\Sigma}_o(\widetilde{\theta}_2)\right) - \theta_{20}\right\| &= o_p\left(n^{-\frac{1}{4}}\right) + O_p\left(n^{-\frac{p_2 p_1}{2p_1+d}}\right) + o_p\left(n^{-\frac{p_2}{4}}\right) + o_p\left(n^{-\frac{p_2 p_\Sigma}{4}}\right) \\ &= o_p\left(n^{-\frac{1}{4}}\right),\end{aligned}$$

for $p_2 > 1$, $p_\Sigma \geq 1$ and $p_1 > \frac{d}{2}$, where the first equality follows from the fact that $\left\|\widehat{\theta}_1 - \theta_{10}\right\| \leq \left\|\widehat{\theta}_1 - \theta_{10}\right\|_2$ for all $\theta_1 \in \Theta_1$. ■

Proof of Theorem 4. This theorem is proved by first verifying Assumptions 4.1–4.6 of Theorem 4.1 in Ai and Chen (2003) given that all the conditions in Theorem 3 are satisfied: *(AC03)4.1: (i)* $E\{D_{u_2^*}(\digamma)'\Sigma_0(\digamma)^{-1}D_{u_2^*}(\digamma)\}$ *is positive definite; (ii)* $(\rho_0, \beta_0) \in int(\Lambda_2)$*; (iii)* $\Sigma_0(\digamma) \equiv var[g(\theta_{20}, Z)|\digamma]$ *is positive definite for all* $\digamma \in \mathcal{F}$*. (AC03)4.2: There is* $v_n^* = \left(v_{\theta_2}^*, -\pi_n u_2^* v_{\theta_2}^*\right) \in \Theta_{2n} - \theta_{20}$ *such that* $\|v_n^* - v^*\| = O\left(n^{-\frac{1}{4}}\right)$*. Let* $\mathcal{N}_{on} \equiv \left\{\theta_2 \in \Theta_{2n} : \|\theta_2 - \theta_{20}\|_s = o(1), \|\theta_2 - \theta_{20}\| = o(n^{-\frac{1}{4}})\right\}$ *and* $\mathcal{N}_o$ *is defined the same way with* Θ_{2n} *replaced by* Θ_2*. Denote* $\frac{dg(\theta_2, Z)}{d\theta_2} \equiv \frac{dg(Z, \theta_2 + \tau v)}{d\tau}$ *a.s.* Z *and* $\frac{dm(\digamma, \theta_2)}{d\theta_2}[v] \equiv E\left\{\frac{dg(\theta_2, Z)}{d\theta_2}[v]\,|\digamma\right\}$ *a.s.* $\digamma$*. (AC03)4.3: (i) For all* $\theta_2 \in N_o$*, the pathwise derivative* $\frac{dg(Z, \theta_2(t))}{d\theta_2}[v]$ *exists a.s.* $Z \in Z$*. Moreover, each element of* $\frac{dg(Z, \theta_2)}{d\theta_2}[v_n^*]$ *satisfies an envelope condition and is Hölder continuous in* $\theta_2 \in \mathcal{N}_{on}$*; (ii) each element of* $\frac{dm(\digamma, \theta_2)}{d\theta_2}[v_n^*]$ *is in* Λ_c^γ *with* $\gamma > \frac{d_\digamma}{2}$ *for all* $\theta_2 \in N_o$*; (AC03)4.4: Uniformly over* $\theta_2 \in \mathcal{N}_{on}$*,*

$$E\left[\left\|\frac{dm(\digamma, \theta_2)}{d\theta_2}[v_n^*] - \frac{dm(\digamma, \theta_{20})}{d\theta_2}[v_n^*]\right\|_E^2\right] = o\left(n^{-\frac{1}{2}}\right).$$

(AC03)4.5: Uniformly over $\theta_2 \in \mathcal{N}_o, \overline{\theta}_2 \in \mathcal{N}_{on}$,

$$E\left[\begin{array}{c}\left\{\frac{dm(F,\theta_{20})}{d\theta_2}\left[v^*\right]\right\}'\Sigma_0(F)^{-1} \\ \times\left\{\frac{dm(F,\theta_2)}{d\theta_2}\left[\overline{\theta}_2-\theta_{20}\right]-\frac{dm(F,\theta_{20})}{d\theta_2}\left[\overline{\theta}_2-\theta_{20}\right]\right\}\end{array}\right]=o\left(n^{-\frac{1}{2}}\right).$$

(AC03)4.6: For all $\theta_2 \in \mathcal{N}_{on}$*, the pathwise second derivative* $\left.\frac{d^2g(Z,\theta_2+\tau v_n^*)}{d\tau^2}\right|_{\tau=0}$ *exists a.s.* $Z \in Z$ *and is bounded by a measurable function* $c(Z)$ *with* $E\left[c^2(Z)\right]<\infty$.

Assumption (AC03)4.1(i) is implied by Proposition 3 and Assumption 2.4.2. Assumptions (AC03)4.1(ii) and (iii) are assumed directly in Assumptions 4.5.1 and 4.4.8. Assumption (AC03)4.2 is satisfied with Assumption 4.4.10 and Assumption 4.5.2(i). Assumption (AC03)4.3(i) is implied by Assumptions 4.4.1-4.4.4 and 4.5.2(i). Assumption (AC03)4.3(ii) is satisfied with Assumptions 4.5.2(i) and 2.4.2. Conditions (AC03)4.4 and (AC03)4.5 are trivially satisfied and, since the second derivative of $g(\theta_2, Z_i)$ is always zero, (AC03)4.6 is automatically satisfied. This implies that we can write

$$\begin{aligned}\sqrt{n}[(\widehat{\rho},\widehat{\beta})'-(\rho_0,\beta_0)'] &= \{-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_{u_2^*}(F_i)'\Sigma_0(F_i)^{-1}g(\theta_{20}(\widehat{\phi}_0,\widehat{\gamma}),Z_i)\} \\ &\times E\{D_{u_2^*}(F_i)'\Sigma_0(F_i)^{-1}D_{u_2^*}(F_i)\}^{-1}+o_p(1). \qquad (71)\end{aligned}$$

Taking a first-order Taylor series expansion of $g(\theta_{20}(\widehat{\phi}_0,\widehat{\gamma}))$ around (ϕ_{00},γ_0) gives us

$$\begin{aligned}\sqrt{n}[(\widehat{\rho},\widehat{\beta})'-(\rho_0,\beta_0)'] &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_{u_2^*}(F_i)'\Sigma_0(F_i,\theta_{20})^{-1}\{g(\theta_{20}(\phi_{00},\gamma_0),Z_i) \\ &\quad+\frac{\partial g(\theta_{20}(\phi_{00},\gamma_0),Z_i)}{\partial\theta_{20}'}\frac{\partial\theta_{20}(\phi_{00},\gamma_0)}{\partial(\phi_0,\gamma)'}[(\widehat{\phi}_0,\widehat{\gamma})'-(\phi_{00},\gamma_0)']\} \\ &\times E\{D_{u_2^*}(F_i)'\Sigma_0(F_i,\theta_{20})^{-1}D_{u_2^*}(F_i)\}^{-1} \\ &\quad-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_{u_2^*}(F_i)'\Sigma_0(F_i,\theta_{20})^{-1}o(\left\|(\widehat{\phi}_0,\widehat{\gamma})'-(\phi_{00},\gamma_0)'\right\|) \\ &\times E\{D_{u_2^*}(F_i)'\Sigma_0(F_i,\theta_{20})^{-1}D_{u_2^*}(F_i)\}^{-1}+o_p(1). \qquad (72)\end{aligned}$$

From Theorem 2,

$$\sqrt{n}\left[(\widehat{\phi}_0,\widehat{\gamma})'-(\phi_{00},\gamma_0)'\right]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\triangle\upsilon_i)D_{u_1^*}(\omega_i)E\{D_{u_1^*}(\omega_i)'D_{u_1^*}(\omega_i)\}^{-1}+o_p(1). \qquad (73)$$

Hence, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} D_{u_2^*}(\digamma_i)'\Sigma_0(\digamma_i)^{-1}\frac{\partial g(\theta_{20}(\phi_{00},\gamma_0),Z_i)}{\partial\theta_{20}'}\frac{\partial\theta_{20}(\phi_{00},\gamma_0)}{\partial(\phi_0,\gamma)'} \\ \longrightarrow\ & E\left[D_{u_2^*}(\digamma_i)'\Sigma_0(\digamma_i)^{-1}\frac{\partial g(\theta_{20}(\phi_{00},\gamma_0),Z_i)}{\partial\theta_{20}'}\frac{\partial\theta_{20}(\phi_{00},\gamma_0)'}{\partial(\phi_0,\gamma)'}\right]. \end{aligned} \tag{74}$$

In our framework,

$$\begin{aligned} \frac{\partial g(\theta_{20}(\phi_{00},\gamma_0),Z_i)}{\partial\theta_{20}'}\frac{\partial\theta_{20}(\phi_{00},\gamma_0)}{\partial(\phi_0,\gamma)'} &= d_i d_{i-1}\{\frac{\partial\overline{\lambda}_0(\overline{v}_i,\overline{v}_{i-1},z_i^1)}{\partial\overline{v}_i}(d_{i-1},z_i)' \\ &+ \frac{\partial\overline{\lambda}_0(\overline{v}_i,\overline{v}_{i-1},z_i^1)}{\partial\overline{v}_i}(d_{i-2},z_{i-1})'\}. \end{aligned} \tag{75}$$

Therefore by substituting Equations (73), (74) and (75) into Equation (72) and applying a standard CLT for i.i.d. data and the Generalized Slutsky's Theorem, we obtain Theorem 4. ■

Acknowledgement *We are grateful to the Editor and to three anonymous referees, whose comments significantly improved the content of this paper. We also thank Arup Bose, Mehmet Caner, David DeJong, Vilmos Komornik, Robert Miller, Whitney Newey, and Jean-Francois Richard for useful comments on the early draft of this paper. We would also like to thank seminar participants at Carnegie-Mellon, Cincinnati, Pittsburgh, SUNY-Stony Brook, Wisconsin-Madison, the 2003 North American Summer Meetings of the Econometric Society, the 2003 European Summer Meetings of the Econometric Society, and Semiparametrics in Rio 2004. Gayle's research is partially supported by the Andrew Mellon Reseach Fellowship. Viauroux's research is partially supported by Charles Phelps Taft research grants.*

References

[1] Ahn, H., and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

[2] Ahn, H., and P. Schmidt (1995): "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68, 5–27.

[3] Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 6, 1795–1843.

[4] Altug, S. and R. A. Miller (1998), "The Effect of Work Experience on Female Wages and Labour Supply," *Review of Economic Studies*, 45–85.

[5] Amemiya, T. (1994): *Advanced Econometrics*. Cambridge, Harvard University Press.

[6] Anderson, T. W., and C. Hsiao (1982): "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics*, 18, 47–82.

[7] Arellano, M., and S. R. Bond (1991): "Some Tests of Specifications for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277–297.

[8] Arellano, M., and O. Bover (1995): "Another Look at the Instrumental Variable Estimation of Error-Component Models," *Journal of Econometrics*, 68, 29–51.

[9] Arellano, M., and R. Carrasco (2003): "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics,* 115, 125-157.

[10] Blundell, R., Chen, X., Kristensen, D. (2004):"Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," CEMMAP Working Paper no CWP15/03, Institute for Fiscal Studies, University College of London.

[11] Bond, S., and C. Meghir (1994): "Dynamic Investment Models and the Firm's Financial Policy," *Review of Economic Studies*, 61, 197–222.

[12] Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

[13] Chen, S. (1998): "Root N Consistent Estimation of a Panel Data Sample Selection Model," Unpublished Manuscript, The Hong Kong University of Science and Technology.

[14] Chen, X. (2005): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Vol. 6, eds. J. Heckman and E. Leamer. Elsevier Science. Forthcoming.

[15] Chen, X., Hansen L. P., Sheinkman, J. (1997): "Shape-Preserving Estimation of Diffusions," *Unpublished Manuscript,* University of Chicago.

[16] Chen, X., Linton, O., Van Keilegom, I. (2003): "Estimation of Semiparametric Models when the Criterion Function is not Smooth," *Econometrica*, 71, 1591–1608.

[17] Chen, X., and X., Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

[18] Cosslett, S. R. (1991): "Distribution-Free Estimator of a Regression Model with Sample Selectivity," in *Nonparametric and Semiparametric Methods in Econometrics*, eds. W. A. Barnett, J. L. Powell and G. Tauchen. Cambridge, Cambridge University Press, 175–197.

[19] Darolles, S., Florens, J. P., Renault, E. (2003): "Nonparametric Instrumental Regression, IDEI Working Paper 228, University of Toulouse.

[20] Das, M., W. K. Newey and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

[21] Gabushin, V. N. (1967): "Inequalities for norms of functions and their derivatives in L_p Metric," *Matematicheskie Zametki,* 1, 291-298.

[22] Gayle, G. L., and R. A. Miller (2004): "Life-cycle Fertility Behavior and Human Capital Accumulation," Working Paper no. 2004-E, Carnegie Mellon University.

[23] Hahn (1997): "Bias Corrected Instrumental Variables Estimation for Dynamic Panel Models with Fixed Effects," Working Paper 2005-024, Boston University.

[24] Heckman, J. (1974): "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42, 679–694.

[25] Heckman, J. (1976): "Common Structures of Statistical Models of Truncations Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 15, 475–492.

[26] Heckman, J. (1976): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

[27] Honore, B. (1993): "Orthogonality conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables," *Journal of Econometrics*, 29, 35-61.

[28] Honore, B. and E. Kyriazidou (2000): " Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica,* 68, 839-874.

[29] Honore, B. E., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," *Econometrica*, 70, 2053–2063.

[30] Hotz, V. J., F. Kydland, and G. Sedlacek (1988): "Intertemporal Preferences and Labor Supply," *Econometrica*, 56, 335–360.

[31] Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

[32] Kydland, F. E. and E. C. Prescott (1982): "Time to Build and Aggregate Fluctuations," *Econometrica*, 50, 1345–1370.

[33] Kyriazidou, E. (1997): "Estimation of Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

[34] Kyriazidou, E. (2001): "Estimation of Dynamic Panel Data Sample Selection Model," *Review of Economic Studies*, 68, 543–572.

[35] Kolmogorov, A. N., Tihomirov, V. M. (1961): "*ε-entropy and ε-capacity sets in function spaces*," in American Mathematical Society Translations, 2, 227-304.

[36] Lorentz, G (1966): *Approximation of functions.* New York, Holt.

[37] Manski, C.F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.

[38] MaCurdy, T. E. (1981): " An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy,* 89, 1059-1085.

[39] Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 16, 1–32.

[40] Newey, W. K. (1994): "Kernel Estimation of Partial Means and A General Variance Estimator," *Econometric Theory,* 10, 233-253.

[41] Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics,* 79, 147-168.

[42] Newey, W. K. (1999): "Two-Step Series Estimation of Sample Selection Models," Working Paper, Department of Economics, MIT.

[43] Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, eds. R. F. Engle and D. L. McFadden. Amsterdam: North-Holland. 2111–2245.

[44] Newey, W. K. and J. L. Powell (2003): Instrumental Variable Estimation of Nonparametric Models, *Econometrica,* 71, 5, 1565-1578.

[45] 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.

[46] Ossiander, M. (1987): "A central limit theorem under metric entropy with L_2 bracketing," *The Annals of Probability*, 15, 897-919.

[47] Robinson, P. (1988): "Root N Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

[48] Shen, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

[49] Shen, X. and W. H. Wong (1994): "Convergence Rate of Sieve Estimates," *Annals of Statistics*, 22, 580–615.

[50] Timan, A.F. (1963): *Theory of Approximation of Functions of a Real Variable.* McMillan, New York.

[51] Van de Geer, S. (2000): *Empirical Processes in M-estimation.* Cambridge, Cambridge University Press.

[52] Zabel, J. (1992): "Estimating Fixed and Random Effects Model with Selectivity," *Economics Letters*, 40, 269–272.