Discussion Paper Series
No. 1707
November 2017

Efficient Estimation of Linear Panel Data Models with Sample Selection and Fixed Effects

Chirok Han Goeun Lee

The Institute of Economic Research - Korea University
Anam-dong, Sungbuk-ku, Seoul, 136-701, South Korea, Tel: (82-2) 3290-1632, Fax: (82-2) 928-4948

Efficient Estimation of Linear Panel Data Models with Sample Selection and Fixed Effects*

Chirok Han† Goeun Lee‡

October 2017

Abstract

For linear panel data models with endogenous selectivity, the popular pooled ordinary least squares with bias correction and its minimum distance variant can suffer from severe efficiency loss in the presence of large random effects. To resolve this problem, we algebraically derive an efficient estimator based on the moment restrictions used by the pooled ordinary least squares and make the estimator feasible under the conventional error-component assumption. The efficient estimation involves heavy computation, and we propose a convenient suboptimal estimator based on a novel common weighting transformation. We also consider partial and full aggregation of information in pairwise differences, where unobserved fixed effects are completely eliminated. Efficient estimation based on pairwise differences is discussed, and a computationally affordable method of estimating nuisance higher-order moments is proposed. Analytic standard errors are provided for all considered estimators. Simulations suggest that a convenient suboptimal estimator and the fully-aggregated pairwise-differencing estimator exhibit remarkable performances. The methods are applied to estimating earnings equation for married women using the Korean Labor and Income Panel Study data.

Keywords: Fixed effects, selection bias correction, efficiency, correlated random effects, pairwise differencing

JEL Classificiation: C23

*We thank Yoon-Jae Whang and Hyungsik Roger Moon for very helpful comments. This research was supported by the Korean Government (2014S1A2A2027803).

†Department of Economics, Korea University, 145 Anam-ro Seongbuk-gu, Seoul 02841, Republic of Korea. E-mail: chirokhan@korea.ac.kr.

‡Department of Economics, Korea University, 145 Anam-ro Seongbuk-gu, Seoul 02841, Republic of Korea. E-mail: gelee@korea.ac.kr.

1 Introduction

Bias due to sample selection has been an important research consideration in various fields of economic applications. To take some recent examples, Machado (2017) deals with selection bias in gender wage gap estimation by examining women who are always employed; Balleer, Gehrke, Lechthaler and Merkl (2016) estimate models with selectivity in analyzing the role of publicly subsidized working time reduction as a fiscal stabilizer; Vossmeyer (2016) studies the effectiveness of lender of last resort policies in a framework for estimating multivariate treatment effect models in the presence of sample selection; Chen and Flores (2015) address sample selection and noncompliance in assessing the wage effects of a job training program for disadvantaged youth in the United States; Rupert and Zanella (2015) consider sample selection when they investigate life cycle profiles of wage rates and hours of market work; Alva, Gray, Mihaylova and Clarke (2014) estimate sample selection models to account for possibly endogenous nonresponses in an analysis of the effect of diabetes complications on health-related quality of life; Jiménez, Ongena, Peydró and Saurina (2014) use a panel-data version of the sample-selection model to identify the effects of monetary policy on credit risk-taking; Revelli (2013) considers endogenous selection in a study on the local tax mix determination. Selectivity is also an area of ongoing theoretical research. Hoonhout and Ridder (2017) examine attrition in panels with refreshment samples by proposing the sequentially additive nonignorable attrition model; Semykina and Wooldridge (2017) consider estimation of binary-response panel data models with selection; Malikov, Kumbhakar and Sun (2016) extend Kyriazidou's (1997) framework to varying coefficient panel data models; Jochmans (2015) considers sample selection in models with multiplicative errors; Sasaki (2015) studies nonparametric identification for dynamic panel data models with selection.

The present paper revisits the selectivity issue in linear panel data models with an emphasis on efficiency. In particular, we address in depth the issues introduced by the presence of unobservable individual effects. The extant approaches can be classified into two categories with regard to handling the fixed-effects. One approach, developed by Wooldridge (1995), deals with the fixed effects by using the convenient correlated random effects (CRE) framework originating from Chamberlain (1980) and Mundlak (1978). The other, taken by Kyriazidou (1997) and Rochina-Barrachina (1999), accounts for the unobservable fixed effects by comparing two periods in pairs. Malikov *et al.* (2016) extend Kyriazidou's (1997) methodology to varying coefficient panel data models. The present paper contributes to the literature on both of the approaches.

In Wooldridge's CRE approach, fixed effects are decomposed into the component correlated with the exogenous regressors and the remainder term assumed to be independent of the regressors. Unlike the case

with balanced panel data, however, the unobservable individual effects are not fully controlled for by the Chamberlain-Mundlak device even for linear models if the sample is partially observed due to sample selection. As a result, the pooled ordinary least squares (POLS) and the minimum distance estimation (MDE) of the equation augmented with bias-correction terms may suffer from substantial efficiency loss, especially when the variance of the individual effects are large (see Examples 1 and 2 later). We address this issue by deriving an efficient estimator based on the moment restrictions used by Wooldridge's POLS. The efficient estimation is computational heavy, however, and we propose a convenient suboptimal estimator as an alternative.

The second approach to accounting for fixed effects, which is called the pairwise differencing (PD) method, is bias-correction applied to differences, simply put. Although PD completely eliminates the fixed effects, it involves estimating nuisance covariances of selection equation errors for implementation, as explained in detail in Section 3. There are two flavors available at present for handling those nuisance parameters. Kyriazidou (1997) and Malikov *et al.* (2016) consider the case where the covariance parameters are irrelevant under the assumption that the difference in the exogenous regressors is uncorrelated with the difference in the errors if the propensity to selection is the same at the two considered periods. Rochina-Barrachina (1999) does not make such an assumption and discusses methods of directly estimating the nuisance parameters.

Our contribution to the PD literature relates to Rochina-Barrachina's (1999) parametric estimation method. We propose an intuitive and approachable way of pooling information over all available periods as a simpler alternative to the MDE suggested by Rochina-Barrachina (1999). The resulting "full-aggregation" estimator is a weighted within-group estimator with bias-correction terms, where the right weighting is crucial for consistent estimation. Remarkably, both the full-aggregation estimator and the MDE are inefficient, and we derive an efficient estimator, although the advantage we get from the efficient estimation is limited once the fixed effects are already eliminated by pairwise differencing. We also provide a practically convenient fail-proof method of estimating serial correlation in the selection equation error, which plays a pivotal role in the implementation of bias correction.

All the estimators discussed in this paper are multi-step estimators, so that calculating valid standard errors is highly involved due to the generated-regressors problem (Pagan, 1984). We provide with general analytic formulae that can be used for constructing valid confidence intervals for all the estimators considered in the present paper. All the claims are verified by simulations.

The rest of the paper is organized as follows. In Section 2, we consider efficient estimation of the CRE

model and present a simple suboptimal estimator that practically solves the large individual-effects problem. Section 3 discusses the PD approach and proposes a convenient pooling ("weighted within-group") estimator. Section 4 presents the simulation results and an application to earnings equation for married women using the Korean Labor and Income Panel Study data. The last section concludes this paper. For all the estimation methods considered in this paper, we provide analytical formulae of standard errors in the appendix.

2 Estimation of the CRE Model

In this section we consider the linear panel data model

$$y_{it} = \mathbf{x}_{it}\beta + \mathbf{z}_i\gamma + u_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{x}_{it}$ is the $1 \times k$ vector of strictly exogenous time-varying regressors, and $\mathbf{z}_i$ is a row vector of time-invariant exogenous variables that contain $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ and possibly other time-invariant variables including a constant term. This model can be derived from the conventional panel data model with fixed effects $y_{it} = \mathbf{x}_{it}\beta + \alpha_i^0 + \varepsilon_{it}^0$ by applying the Chamberlain-Mundlak device $\alpha_i^0 = \mathbf{z}_i\gamma + a_i$, where a_i and $\mathbf{z}_i$ are assumed to be independent. We then obtain (1) by letting $u_{it} = a_i + \varepsilon_{it}^0$. The error term u_{it} has a zero mean, contains random effects, and is arbitrarily correlated over t. For the analysis in this section, $\mathbf{x}_{it}$ needs not be strictly exogenous and may in fact include predetermined variables as long as $\mathbf{z}_i$ is properly modified to contain only exogenous variables (see Dustmann and Rochina-Barrachina, 2007, and Semykina and Wooldridge, 2010). We assume that the regressors are strictly exogenous for simplicity and for direct comparison to the PD approach in Section 3.

The indicator of y_{it} being observed is denoted by s_{it}. The covariates $\mathbf{x}_{it}$ and $\mathbf{z}_i$ are observed regardless of the value of s_{it}. Wooldridge (1995) suggests that

$$s_{it} = 1[\mathbf{z}_i\pi_t + v_{it} > 0], \quad v_{it} \sim N(0,1) \tag{2}$$

for every t, where $\mathbf{z}_i$ contains $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$. Following Heckman (1976, 1979) and Wooldridge (1995), we let $u_{it} = \delta_t v_{it} + \varepsilon_{it}$, where ε_{it} is independent of v_{it} conditional on $\mathbf{z}_i$, so that

$$\mathrm{E}(y_{it}|\mathbf{z}_i, s_{it} = 1) = \mathbf{x}_{it}\beta + \mathbf{z}_i\gamma + \delta_t \mathrm{E}(v_{it}|\mathbf{z}_i, s_{it} = 1) = \mathbf{x}_{it}\beta + \mathbf{z}_i\gamma + \delta_t\lambda_{it}. \tag{3}$$

Above, λ_{it} is the inverse Mills ratio $\lambda_{it} = \lambda(\mathbf{z}_i\pi_t) = \phi(\mathbf{z}_i\pi_t)/\Phi(\mathbf{z}_i\pi_t)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the density function and distribution function of the standard normal distribution. Note that ε_{it} need not be normal if it is independent of v_{it} as Wooldridge (1995) points out.

The moment restrictions considered in this section are that $\mathrm{E}(e_{it}|\mathbf{z}_i, s_{it} = 1) = 0$ for every t, where $e_{it} = y_{it} - \mathbf{x}_{it}\beta - \mathbf{z}_i\gamma - \delta_t\lambda_{it}$. The model described by (2) and (3) specifies an average characteristic of y_{it} over the observed population. It is thus a population-averaged model.

2.1 Pooled OLS and Minimum Distance Estimation

For consistent estimation of the parameters in (3), Wooldridge (1995) proposes a two-step procedure. In the first step, λ_{it} is estimated by the probit regression of s_{it} on $\mathbf{z}_i$ for each t. In the second step, β, γ and $\delta_1, \ldots, \delta_T$ are estimated by the pooled OLS (POLS) regression of y_{it} on $\mathbf{x}_{it}$, $\mathbf{z}_i$ and $\hat{\lambda}_{it} = \lambda(\mathbf{z}_i\hat{\pi}_t)$ properly interacted with time dummies, using observations with $s_{it} = 1$, where $\hat{\pi}_t$ are the first-step probit estimators. For formality and future use, let $\hat{\mathbf{w}}_{it} = (\mathbf{x}_{it}, \mathbf{z}_i, 0, \ldots, 0, \hat{\lambda}_{it}, 0, \ldots, 0)$ and $\theta = (\beta', \gamma', \delta_1, \ldots, \delta_T)'$. Letting $y_i = (y_{i1}, \ldots, y_{iT})'$, $S_i = \mathrm{diag}(s_{i1}, \ldots, s_{iT})$ and $\hat{W}_i = (\hat{\mathbf{w}}_{i1}', \ldots, \hat{\mathbf{w}}_{iT}')'$, Wooldridge's POLS estimator is written as

$$\hat{\theta}_{pols} = \left(\sum_{i=1}^{n} \hat{W}_i' S_i \hat{W}_i \right)^{-1} \sum_{i=1}^{n} \hat{W}_i' S_i y_i.$$

Note that all the elements of $S_i y_i$ are observed even though some elements of y_i are not. Obtaining valid standard errors for $\hat{\theta}_{pols}$ is involved because of the "generated regressors" problem (Pagan, 1984). Wooldridge (1995) presents a way of calculating standard errors for POLS. We address this issue in Appendix A.3 in a more general manner applicable to all the estimators considered in the present paper.

Although $\mathbf{z}_i$ contains $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$, the fixed effects are not eliminated and the POLS estimator is different from the within-group estimator when the dependent variable is only partially observed. For example, if y_{it} is observed for only one t for some i, then the individual effect remains for that i. As a result, the performance of $\hat{\theta}_{pols}$ can deteriorate seriously if the random effects in the error term have a large variance. The following example is illustrative.

Example 1. Consider the two-period panel data model $y_{it} = \alpha + \beta t + \mu_i + \varepsilon_{it}$ for $t = 0, 1$. We are interested in the β parameter. The dependent variable is observed for all units in the initial period ($t = 0$), but y_{i1} is observed only for selected observations. Let s_i be the dummy variable indicating the selection. Let $s_i = I(z_i\pi + v_i > 0)$ for some exogenous z_i and $v_i \sim N(0, 1)$, where v_i is assumed to be independent of z_i, μ_i and ε_{it} for simplicity. Let π be known for simplicity so that the inverse Mills ratio λ_i is directly observed. The POLS estimator of α and β is obtained by regressing $S_i y_i$ on $S_i W_i$, where $S_i = \mathrm{diag}(1, s_i)$,

$y_i = (y_{i0}, y_{i1})'$, and

$$W_i = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & \lambda_i \end{pmatrix}.$$

We have

$$\frac{1}{n}\sum_{i=1}^n W_i' S_i W_i = \begin{pmatrix} 1+\bar{s} & \bar{s} & \overline{s\lambda} \\ \bar{s} & \bar{s} & \overline{s\lambda} \\ \overline{s\lambda} & \overline{s\lambda} & \overline{s\lambda^2} \end{pmatrix}, \quad \frac{1}{n}\sum_{i=1}^n W_i' S_i y_i = \begin{pmatrix} \bar{y}_0 + \overline{sy_1} \\ \overline{sy_1} \\ \overline{s\lambda y_1} \end{pmatrix},$$

where

$$\bar{s} = \frac{1}{n}\sum_{i=1}^n s_i, \quad \overline{s\lambda} = \frac{1}{n}\sum_{i=1}^n s_i\lambda_i, \quad \overline{s\lambda^2} = \frac{1}{n}\sum_{i=1}^n s_i\lambda_i^2,$$
$$\bar{y}_0 = \frac{1}{n}\sum_{i=1}^n y_{i0}, \quad \overline{sy_1} = \frac{1}{n}\sum_{i=1}^n s_i y_{i1}, \quad \overline{s\lambda y_1} = \frac{1}{n}\sum_{i=1}^n s_i\lambda_i y_{i1}.$$

Let $Q_{WW} = n^{-1}\sum_{i=1}^n W_i' S_i W_i$ and $Q_{Wy} = n^{-1}\sum_{i=1}^n W_i S_i y_i$, both of which are observed. We have

$$(HQ_{WW}H')^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \bar{s} & \overline{s\lambda} \\ 0 & \overline{s\lambda} & \overline{s\lambda^2} \end{pmatrix}^{-1} = \frac{1}{d}\begin{pmatrix} d & 0 & 0 \\ 0 & \overline{s\lambda^2} & -\overline{s\lambda} \\ 0 & -\overline{s\lambda} & \bar{s} \end{pmatrix},$$

where $d = \bar{s}\cdot\overline{s\lambda^2} - \overline{s\lambda}^2$, and

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus, $\hat{\beta}$ is the second element of $H'(HQ_{WW}H')^{-1}HQ_{Wy}$, i.e.,

$$\hat{\beta} = d^{-1}(-1,1,0)\begin{pmatrix} d & 0 & 0 \\ 0 & \overline{s\lambda^2} & -\overline{s\lambda} \\ 0 & -\overline{s\lambda} & \bar{s} \end{pmatrix}\begin{pmatrix} \bar{y}_0 \\ \overline{sy_1} \\ \overline{s\lambda y_1} \end{pmatrix} = -\bar{y}_0 + \left(\frac{\overline{s\lambda^2}}{d}\right)\overline{sy_1} - \left(\frac{\overline{s\lambda}}{d}\right)\overline{s\lambda y_1}$$
$$= \frac{1}{n}\sum_{i=1}^n (q_i s_i y_{i1} - y_{i0}), \quad q_i = d^{-1}(\overline{s\lambda^2} - \overline{s\lambda}\lambda_i).$$

By plugging in $y_{it} = \alpha + \beta t + \mu_i + \varepsilon_{it}$, we have

$$\hat{\beta} = \beta + \frac{1}{n}\sum_{i=1}^n (q_i s_i - 1)\mu_i + \frac{1}{n}\sum_{i=1}^n (q_i s_i \varepsilon_{i1} - \varepsilon_{i0}).$$

Note that $n^{-1}\sum_{i=1}^n (q_i s_i - 1) = 0$ but $n^{-1}\sum_{i=1}^n (q_i s_i - 1)\mu_i \neq 0$ in general when μ_i is present. The variance of μ_i does matter even when μ_i and s_i are independent. If the variance of μ_i is large, the performance of $\hat{\beta}$ can be poor. For example, if the standard deviation of μ_i is of order $\sqrt{n}$, then $\hat{\beta}$ is not even consistent. ■

MDE is available as an alternative to pooled regression (Wooldridge, 1995). For example, after y_{it} is regressed on $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ and λ_{it} for each t using the observations with $s_{it} = 1$, the structural parameters β and γ can be recovered by a minimum distance procedure based on the relationship $\tau_{tr} = \beta I(t = r) + \gamma_r$, where τ_{tr} is the coefficient of $\mathbf{x}_{ir}$ for period t. Importantly, the issue in Example 1 is not resolved by this MDE. See the following example.

Example 2. Under the same settings as in Example 1, the MDE after fitting a model for each t does not provide a solution. Let the structural parameters be α, β and δ, where δ is the the coefficient of λ_i. The "reduced-form" equations are $y_{i0} = \alpha_0 + u_{i0}$ and $y_{i1} = \alpha_1 + \delta_1\lambda_i + e_{i1}$, where $u_{i0} = \mu_i + \varepsilon_{i0}$ and $e_{i1} = \mu_i + \varepsilon_{i1} - \delta_1\lambda_i$. Then,

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \delta_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \delta \end{pmatrix}, \text{ and the MDE is } \begin{pmatrix} \hat{\alpha}_{mde} \\ \hat{\beta}_{mde} \\ \hat{\delta}_{mde} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\delta}_1, \end{pmatrix},$$

where $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\delta}_1$ are the OLS reduced-form estimators. We have $\hat{\beta}_{mde} = \hat{\alpha}_1 - \hat{\alpha}_0$, which is identical to the pooled OLS estimator. ■

Example 2 illustrates that MDE does not solve the problem manifested in Example 1. This phenomenon is not limited to this particular example, but also occurs for more general models and more general T. In the presence of observation-wise heteroskedasticity and serial correlation, it is natural to expect that substantial efficiency gain can come only by observation-wise transformation, not by a linear combination of reduced-form estimators. In the next section, we show how such an efficient estimator is derived.

2.2 Efficient Estimation

The POLS and the MDE are not efficient, and a problem is illustrated in Examples 1 and 2 in the presence of random effects in u_{it}. In this section we derive an efficient estimator based on the moment restrictions used by the CRE approach. To proceed, we assume that π_t's are known and thus λ_{it}'s are observed. Let $\mathbf{w}_{it} = (\mathbf{x}_{it}, \mathbf{z}_i, 0, \ldots, 0, \lambda_{it}, 0, \ldots, 0)$ and $\theta = (\beta', \gamma', \delta_1, \ldots, \delta_T)'$ as before, so that the bias-corrected main equation is $y_{it} = \mathbf{w}_{it}\theta + e_{it}$, where $e_{it} = u_{it} - \delta_t\lambda_{it}$. The moment restrictions derived for the CRE model are $\mathrm{E}(s_{it}e_{it}|\mathbf{z}_i) = 0$ for $t = 1, \ldots, T$. Stacked vertically, those T conditional moment restrictions are written in matrix notation as

$$\mathrm{E}(S_i e_i | \mathbf{z}_i) = 0, \tag{4}$$

where $S_i = \text{diag}(s_{i1}, \ldots, s_{iT})$ and $e_i = (e_{i1}, \ldots, e_{iT})'$. Letting y_i be the T-vector of y_{it} and $W_i = (\mathbf{w}_{i1}', \ldots, \mathbf{w}_{iT}')'$, (4) is written as $\mathrm{E}[S_i(y_i - W_i\theta)|\mathbf{z}_i] = 0$, where $S_i y_i$ is observed although y_i is not completely. With these notations used, Wooldridge's POLS is the method of moments estimator based on the unconditional moment restrictions $\mathrm{E}[W_i' S_i(y_i - W_i\theta)] = 0$, where W_i is replaced with $\hat{W}_i$. These unconditional moment conditions are not optimal given the conditional moment restrictions (4), due to the presence of heteroskedasticity and correlation within $S_i e_i$ conditional on $\mathbf{z}_i$.

We derive an optimal set of unconditional moment restrictions implied by the conditional (4) using Chamberlain's (1992) arguments. For this, introduce the notation $g_i(\theta) = S_i(y_i - W_i\theta)$, and let $D_i = -\mathrm{E}[\frac{\partial}{\partial\theta'} g_i(\theta)|\mathbf{z}_i]$ and $\Omega_i = \mathrm{E}[g_i(\theta)g_i(\theta)'|\mathbf{z}_i]$, both of which are evaluated at the true parameter. Then $D_i'\Omega_i^{-1} g_i(\theta)$ is an optimal set of unrestricted moment functions.

Let us evaluate D_i and Ω_i. First, D_i is straightforward to evaluate. Because W_i is a function of $\mathbf{z}_i$, we have

$$D_i = \mathrm{E}(S_i W_i|\mathbf{z}_i) = P_i W_i,$$

where P_i is the $T \times T$ diagonal matrix of $p_{it} = \mathrm{E}(s_{it}|\mathbf{z}_i) = \Phi(\mathbf{z}_i\pi_t)$. Next, because of the identity $e_{it} = y_{it} - \mathbf{w}_{it}\theta = u_{it} - \delta_t\lambda_{it} = \varepsilon_{it} + \delta_t(v_{it} - \lambda_{it})$, the (t, r) element of $\Omega_i = \mathrm{E}(S_i e_i e_i' S_i|\mathbf{z}_i)$ equals

$$\omega_{i,tr} = \mathrm{E}(s_{it}s_{ir}e_{it}e_{ir}|\mathbf{z}_i) = p_{i,tr}\,\mathrm{E}(\varepsilon_{it}\varepsilon_{ir}) + \delta_t\delta_r m_{i,tr}, \tag{5}$$

where $p_{i,tr} = \mathrm{E}(s_{it}s_{ir}|\mathbf{z}_i)$ and $m_{i,tr} = \mathrm{E}[s_{it}s_{ir}(v_{it} - \lambda_{it})(v_{ir} - \lambda_{ir})|\mathbf{z}_i]$. Note that $p_{i,tt} = \mathrm{E}(s_{it}|\mathbf{z}_i) = p_{it}$, and $p_{i,tr} = \Phi_2(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr})$ for $t \neq r$, where $\rho_{tr} = \mathrm{E}(v_{it}v_{ir})$ and $\Phi_2(a, b; \rho)$ is the bivariate normal (mean 0, variance 1, and covariance ρ) cumulative distribution function evaluated at (a, b). Also, from the known facts about univariate and bivariate truncated normal distributions (see Greene, 2012, and Rosenbaum, 1961), we have

$$\begin{aligned} m_{i,tt} &= p_{i,tt}(1 - z_{it}\lambda_{it} - \lambda_{it}^2), \\ m_{i,tr} &= \rho_{tr}(p_{i,tr} - z_{it}h_{i,tr} - z_{ir}h_{i,rt}) + \sqrt{\frac{1-\rho_{tr}^2}{2\pi}}\,\phi\left(\sqrt{\frac{z_{it}^2 - 2\rho_{tr}z_{it}z_{ir} + z_{ir}^2}{1-\rho_{tr}^2}}\right) \\ &\quad - \lambda_{ir}(h_{i,tr} + \rho_{tr}h_{i,rt}) - \lambda_{it}(h_{i,rt} + \rho_{tr}h_{i,tr}) + p_{i,tr}\lambda_{it}\lambda_{ir}, \end{aligned} \tag{6}$$

where $z_{it} = \mathbf{z}_i\pi_t$, $h_{i,tr} = \phi(z_{it})\Phi(z_{i,rt}^*)$ and $z_{i,rt}^* = (z_{ir} - \rho_{tr}z_{it})(1 - \rho_{tr}^2)^{-1/2}$. Note that $m_{i,tt}$ is also derived from the $m_{i,tr}$ formula using $\rho_{tt} = 1$, $z_{i,tt}^* = 0$ and $h(z_{it}, 0) = \frac{1}{2}\phi(z_{it})$.

Therefore, the infeasible optimal estimator $\tilde{\theta}_{opt}$ solves $\sum_{i=1}^n D_i'\Omega_i^{-1}S_i(y_i - W_i\theta) = 0$, that is,

$$\tilde{\theta}_{opt} = \left(\sum_{i=1}^n W_i'P_i\Omega_i^{-1}S_iW_i\right)^{-1}\sum_{i=1}^n W_i'P_i\Omega_i^{-1}S_iy_i, \tag{7}$$

which is the instrumental variable (IV) estimator for the equation $S_i y_i = S_i W_i \theta + S_i e_i$ using $\Omega_i^{-1} P_i W_i$ as instruments. This estimator attains the asymptotic semiparametric efficiency bound (Chamberlain, 1992) for consistent estimators using the CRE moment restrictions (4).

The infeasible optimal estimator in (7) is derived under the assumption that π_t are known. When π_t are estimated from a first-step probit regression, the optimality of $\tilde{\theta}_{opt}$ is unclear. It is known in the literature that the infeasible and feasible estimators may have different asymptotic distributions (see, e.g., Hirano, Imbens and Ridder, 2003, and Han and Kim, 2011, for asymptotic comparison of feasible and infeasible estimators related with π_t), and there may be a hybrid estimator that is asymptotically more efficient than both the infeasible and the feasible estimators. Theoretically interesting as it may be, we do not pursue this issue further, because our primary goal is to improve efficiency by dealing with the fact that $s_{it} e_{it}$ is heteroskedastic and autocorrelated. An interested reader might benefit from following Han and Kim's (2011) approach as a simple means of analysis.

To make the procedure (7) feasible, we need to consistently estimate π_t, ρ_{tr}, δ_t and $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir})$ for all t and (t,r) pairs. Of them, π_t can be estimated by the initial probit for each t in a standard manner. The ρ_{tr} parameters can be estimated in various ways. One way is to use T-variate probit to estimate all π_t's and ρ_{tr}'s simultaneously, but the computational burden is heavy even for $T = 3$. Rochina-Barrachina (1999) suggests pairwise bivariate probit regressions to estimate π_t, π_r and ρ_{tr} for each pair (t,r). She also proposes estimating ρ_{tr} by a two-step procedure after estimating p_{it} nonparametrically for every t. Our suggestion is close to this latter method but is simpler. Specifically, we first estimate π_t for every t from the probit regressions, and then use the fact that $\Pr(s_{it}s_{ir} = 1|\mathbf{z}_i) = \Phi_2(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr})$ to estimate ρ_{tr} using a likelihood method. That is, $\hat{\rho}_{tr}$ is the maximizer of

$$\ln L(\rho) = \sum_{i=1}^{n} \Big\{ s_{it}s_{ir} \ln \Phi_2(\hat{z}_{it}, \hat{z}_{ir}; \rho) + (1 - s_{it}s_{ir}) \ln[1 - \Phi_2(\hat{z}_{it}, \hat{z}_{ir}; \rho)] \Big\}, \tag{8}$$

where $\hat{z}_{it} = \mathbf{z}_i\hat{\pi}_t$ for simplicity. (Singularity at $\rho = \pm 1$ can be avoided using Fisher's (1915, 1921) logarithmic transformation of reparametrizing ρ to $\operatorname{arctanh}(\rho) = \frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$.) While numerical procedures for bivariate probit often fail to converge, the maximization of (8) mostly works according to our experiments for Section 4.

Once π_t and ρ_{tr} are all estimated, $p_{i,tr}$ and $m_{i,tr}$ in (5) are naturally calculated because they are measurable functions of π_t and ρ_{tr}. The remaining parameters to estimate for the construction of Ω_i are δ_t and $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir})$. First, δ_t can be replaced with Wooldridge's POLS estimators. Next, for the estimation of $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir})$, we make the 'random effects (RE) assumption' that $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir}) = \sigma_a^2 + \sigma_b^2[t = r]$ for some $\sigma_a^2 \geq 0$ and $\sigma_b^2 > 0$.

Because $\omega_{i,tr}$ denotes $\mathrm{E}(s_{it}s_{ir}e_{it}e_{ir}|\mathbf{z}_i)$, we have from (5) that

$$\mathrm{E}[p_{i,tr}^{-1}(s_{it}s_{ir}e_{it}e_{ir} - \delta_t\delta_r m_{i,tr})|\mathbf{z}_i] = \mathrm{E}(\varepsilon_{it}\varepsilon_{ir}).$$

Letting $\tilde{p}_{i,tr}$, $\tilde{m}_{i,tr}$ and $\tilde{\delta}_t$ denoting the initial estimators of $p_{i,tr}$, $m_{i,tr}$ and δ_t, respectively, constructed using the initial estimates of π_t, π_r and ρ_{tr} as explained above, one way of estimating $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir})$ is to get the cross sectional average of $\tilde{p}_{i,tr}^{-1}(s_{it}\tilde{e}_{it}s_{ir}\tilde{e}_{ir} - \tilde{\delta}_t\tilde{\delta}_r\tilde{m}_{i,tr})$, where $s_{it}\tilde{e}_{it}$ is the POLS residual that is replaced with zero when $s_{it} = 0$. Under the standard RE (i.e., exchangeability) assumption that $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir}) = \sigma_a^2 + \sigma_b^2[t = r]$, we may further average those averages over all (t, t) and all (t, r) with $t \neq r$. There are also other methods such as regressing $s_{it}\tilde{e}_{it}s_{ir}\tilde{e}_{ir} - \tilde{\delta}_t\tilde{\delta}_r\tilde{m}_{i,tr}$ on $\tilde{p}_{i,tr}$ and its interaction with the dummy for $t = r$ to estimate σ_a^2 and σ_b^2. Massive experiments by the authors suggest that the following procedure works well.

1. Regress $\tilde{p}_{i,tr}^{-1}(s_{it}\tilde{e}_{it}s_{ir}\tilde{e}_{ir} - \tilde{\delta}_t\tilde{\delta}_r\tilde{m}_{i,tr})$ on a constant and the dummy variable $1[t = r]$ using POLS by pooling over all i, t and r. Let the intercept and the slope estimates be denoted $\hat{\varphi}_a$ and $\hat{\varphi}_b$.

2. Let $\hat{\sigma}_a^2 = \max(\hat{\varphi}_a, 0)$ and $\hat{\sigma}_b^2 = |\hat{\varphi}_b|$, where the max function is taken in order to ensure that $\hat{\sigma}_a^2$ is nonnegative and the absolute value to ensure positivity of $\hat{\sigma}_b^2$. Note that a similar consideration is made in the usual random-effects feasible generalized least squares estimation of panel data models under strict exogeneity (Baltagi, 2013, p. 24; Maddala and Mount, 1973).

3. $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir})$ is estimated by $\hat{\sigma}_a^2 + \hat{\sigma}_b^2[t = r]$.

Now that all the components of (7) are estimated, we have the following feasible optimal estimator:

$$\hat{\theta}_{opt} = \left(\sum_{i=1}^{n} \hat{W}_i'\tilde{P}_i\hat{\Omega}_i^{-1}S_i\hat{W}_i\right)^{-1} \sum_{i=1}^{n} \hat{W}_i'\tilde{P}_i\hat{\Omega}_i^{-1}S_i y_i, \tag{9}$$

where $\tilde{P}_i$ is the diagonal matrix of $\tilde{p}_{it} = \Phi(\mathbf{z}_i\hat{\pi}_t)$ with $\hat{\pi}_t$ being the initial probit coefficient estimates as before, and $\hat{\Omega}_i$ is the $T \times T$ matrix of $\hat{\omega}_{i,tr}$. Standard errors can be obtained using the method described in Appendix A.3.

2.3 A Convenient Suboptimal Estimator

The feasible optimal procedure in the previous section involves inverting n covariance matrices, which can be nuisance. Alternatively, we can use a common covariance $\Omega = n^{-1}\sum_{i=1}^{n}\Omega_i$ in place of Ω_i in (7). It is convenient that the common weight Ω can be consistently estimated by $\tilde{\Omega} = n^{-1}\sum_{i=1}^{n} S_i\tilde{e}_i\tilde{e}_i'S_i$ without estimating nuisance parameters (with "consistency" meaning $\tilde{\Omega} - \Omega \to_p 0$ as $n \to \infty$), where $S_i\tilde{e}_i$ are the

POLS residuals with the unobserved residuals replaced with zero. The resulting *common weighting* (CW) estimator is

$$\hat{\theta}_{cw} = \left(\sum_{i=1}^{n} \hat{W}_i' \tilde{P}_i \tilde{\Omega}^{-1} S_i \hat{W}_i \right)^{-1} \sum_{i=1}^{n} \hat{W}_i' \tilde{P}_i \tilde{\Omega}^{-1} S_i y_i. \tag{10}$$

Unlike the feasible optimal estimator in Section 2.2, the CW estimation requires estimating only π_t, and $\tilde{\Omega}$ can be estimated nonparametrically using the POLS residuals. The CW estimator can also be understood as the IV estimator obtained from the regression of $S_i y_i$ on $S_i \hat{W}_i$ using $\tilde{\Omega}^{-1} \tilde{P}_i \hat{W}_i$ as instruments. Note that the CW estimator in (10) is different from MDE.

Although the CW estimator accounts for error serial correlation somehow, the POLS and the CW estimators are, in fact, not unanimously rankable in terms of efficiency, and we can construct a better estimator by linearly combining them. For this, let $\hat{\theta}(C) = (I - C)\hat{\theta}_{pols} + C\hat{\theta}_{cw} = \hat{\theta}_{pols} - C(\hat{\theta}_{pols} - \hat{\theta}_{cw})$ for given C. When π_t are assumed known so W_i are observable, it turns out that an optimal choice of C is $A_1 A_2^{-1}$, where $A_1 = \mathrm{Cov}(\hat{\theta}_{pols}, \hat{\theta}_{pols} - \hat{\theta}_{cw})$ and $A_2 = \mathrm{Var}(\hat{\theta}_{pols} - \hat{\theta}_{cw})$. The resulting optimal linear combination $\hat{\theta}(A_1 A_2^{-1})$ is at least as efficient as both POLS and CW (with π_t known). We have the following result, where 'Avar' denotes the asymptotic variance.

Theorem 1. $\mathrm{Avar}(\hat{\theta}(A_1 A_2^{-1})) - \mathrm{Avar}(\hat{\theta}(C))$ *is negative semidefinite for all* C.

A proof is given in the appendix.

Remark 1.1. The asymptotic variance comparison in Theorem 1 is valid when π_t are known and thus λ_{it} are observed. When π_t are estimated by first-step probit regressions, the estimator $\hat{\theta}(A_1 A_2^{-1})$ may not be an optimal linear combination of the POLS and CW estimators. But the estimator is anyway inefficient, and the possible efficiency loss is in practice unimportant. ■

We next consider estimating A_1 and A_2. For this, let

$$F_1 = \left(\sum_{i=1}^{n} \hat{W}_i' S_i \hat{W}_i \right)^{-1}, \quad F_2 = \left(\sum_{i=1}^{n} \hat{W}_i' \tilde{P}_i \tilde{\Omega}^{-1} S_i \hat{W}_i \right)^{-1}$$

and

$$G_{11} = \sum_{i=1}^{n} \hat{W}_i' \tilde{\Omega}_i \hat{W}_i, \;\; G_{12} = \sum_{i=1}^{n} \hat{W}_i' \tilde{\Omega}_i \tilde{\Omega}^{-1} \tilde{P}_i \hat{W}_i, \;\; G_{22} = \sum_{i=1}^{n} \hat{W}_i' \tilde{P}_i \tilde{\Omega}^{-1} \tilde{\Omega}_i \tilde{\Omega}^{-1} \tilde{P}_i \hat{W}_i,$$

where $\tilde{\Omega}_i = S_i \tilde{e}_i \tilde{e}_i' S_i$ and $\tilde{\Omega} = n^{-1} \sum_{i=1}^{n} \tilde{\Omega}_i$. Then A_j are consistently estimated by $\hat{A}_j$, where

$$\hat{A}_0 = F_1 G_{11} F_1', \;\; \hat{A}_1 = \hat{A}_0 - F_1 G_{12} F_2', \;\; \hat{A}_2 = \hat{A}_1 - (F_1 G_{12} F_2')' + F_2 G_{22} F_2'.$$

Our proposed estimator (denoted the PO-CW estimator, hereafter) is, thus,

$$\hat{\theta}_* = (I - \hat{C}_*)\hat{\theta}_{pols} + \hat{C}_*\hat{\theta}_{cw}, \quad \hat{C}_* = \hat{A}_1\hat{A}_2^{-1}. \tag{11}$$

This estimator is also obtained by the IV regression of $S_i y_i$ on $S_i\hat{W}_i$ using $Z_i = \hat{W}_i F_1(I-\hat{C}_*)' + \tilde{\Omega}^{-1}\tilde{P}_i\hat{W}_i F_2\hat{C}_*'$ as instruments.

As repeatedly noted so far, the two-step procedures leads to the generated regressors problem, and standard errors should be obtained with this fact accounted for. A convenient general-purpose Delta-method procedure is explained in Appendix A.3. Alternatively, block bootstrapping may be used. The bootstrap variance estimators of the feasible estimators considered in this paper seem to perform well according to the simulations in Section 4.

3 Estimation Based on Pairwise Differencing

In the CRE approach in Section 2, fixed effects are dealt with by the Chamberlain-Mundlak device, where random effects remain and affect the statistical properties of the estimators. In this section we examine fixed-effects approaches, which completely eliminate the time-invariant individual effects, and we propose a convenient method that turns out to work remarkably well.

The model we consider is $y_{it} = \alpha_i + \mathbf{x}_{it}\beta + u_{it}$, where α_i are unobservable fixed effects. As briefly mentioned in the introduction, Kyriazidou (1997) considers the case where selection bias disappears by differencing if the selection propensity remains the same over time. Rochina-Barrachina (1999) considers a general case. We focus on Rochina-Barrachina's parametric approach and propose a convenient data pooling method for better performance.

3.1 First-Difference Estimation

We first consider the first-difference (FD) estimation using the observations with $s_{it} = 1$ as an introduction. The differenced equation is $\Delta y_{it} = \Delta\mathbf{x}_{it}\beta + \Delta u_{it}$, and we use data only for i and t such that Δy_{it} is observed, i.e., $s_{it}s_{it-1} = 1$. The pooled OLS estimator of this equation in differences is inconsistent because of selectivity. In order to derive the correction term, we note that

$$\mathrm{E}(\Delta y_{it}|\mathbf{z}_i, s_{it}s_{it-1} = 1) = \Delta\mathbf{x}_{it}\beta + \delta_t\,\mathrm{E}(v_{it}|\mathbf{z}_i, s_{it}s_{it-1} = 1) - \delta_{t-1}\,\mathrm{E}(v_{it-1}|\mathbf{z}_i, s_{it}s_{it-1} = 1) \tag{12}$$

under the assumption that $u_{it} = \delta_t v_{it} + \varepsilon_{it}$ and ε_{it} is independent of v_{it}. The $\mathrm{E}(v_{it}|\mathbf{z}_i, s_{it}s_{it-1} = 1)$ and $\mathrm{E}(v_{it-1}|\mathbf{z}_i, s_{it}s_{it-1} = 1)$ terms are evaluated using the following facts about the bivariate standard normal variables (x, y) with correlation ρ:

$$\mathrm{E}(x|x > -a, y > -b) = \frac{\phi(a)\Phi(b^*)}{\Phi_2(a, b; \rho)} + \rho \frac{\phi(b)\Phi(a^*)}{\Phi_2(a, b; \rho)} =: \psi(a, b; \rho), \tag{13}$$

where $a^* = (a - \rho b)(1 - \rho^2)^{-1/2}$ and $b^* = (b - \rho a)(1 - \rho^2)^{-1/2}$. (See Rosenbaum, 1961, and Maddala, 1983.) Note that $\psi(a, a; 1) = \lambda(a)$ and $\psi(a, b; 0) = \lambda(a)$ as special cases, where $\lambda(a) = \phi(a)/\Phi(a)$ as before. From (13) we derive that

$$\mathrm{E}(v_{it}|\mathbf{z}_i, s_{it}s_{ir} = 1) = \psi(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr}) =: \psi_{i,tr}, \tag{14}$$

where $\rho_{tr} = \mathrm{E}(v_{it}v_{ir})$ as before. By combining this with (12), we have

$$\mathrm{E}(\Delta y_{it}|\mathbf{z}_i, s_{it}s_{it-1} = 1) = \Delta\mathbf{x}_{it}\beta + \delta_t\psi_{i,t,t-1} - \delta_{t-1}\psi_{i,t-1,t}. \tag{15}$$

After $\psi_{i,tr}$ are estimated using the first-step estimators of the selection equation parameters (see Section 2 for the estimation of π_t and ρ_{tr}), one can regress

$$\begin{pmatrix} s_{i1}s_{i2}\Delta y_{i2} \\ s_{i2}s_{i3}\Delta y_{i3} \\ \vdots \\ s_{iT-1}s_{iT}\Delta y_{iT} \end{pmatrix} \text{ on } \begin{bmatrix} s_{i1}s_{i2}(\Delta\mathbf{x}_{i2} & -\hat{\psi}_{i,12} & \hat{\psi}_{i,21} & 0 & \cdots & 0) \\ s_{i2}s_{i3}(\Delta\mathbf{x}_{i3} & 0 & -\hat{\psi}_{i,23} & \hat{\psi}_{i,32} & \cdots & 0) \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ s_{iT-1}s_{iT}(\Delta\mathbf{x}_{iT} & 0 & 0 & 0 & \cdots & \hat{\psi}_{i,T,T-1}) \end{bmatrix}$$

using POLS. The parameters to estimate are $(\beta', \delta_1, \ldots, \delta_T)'$. In actual implementation, one can simply omit the observations with $s_{it-1}s_{it} = 0$. The correction terms are accounted for differently from Rochina-Barrachina (1999) in that the δ_t parameters are exactly identified. Note that the error terms in the differenced equation are heteroskedastic and serially correlated, and hence the pooled OLS estimation of the differenced equation is inefficient, and robust variance estimation is required for inferences.

Comparing this FD estimation with the estimation of the CRE model (3), the CRE model describes the distribution of (y_{it}, s_{it}) conditional on $\mathbf{z}_i$ for *each* t, while the FD estimation requires information on the *joint* distribution of the pairs (y_{it}, s_{it}) and (y_{it-1}, s_{it-1}) for $t = 2, \ldots, T$. The moment conditions in (4) for the CRE estimation and those in (15) for the FD estimation do not overlap, so efficiency cannot be ranked between the CRE estimators and the FD estimator. Simulation results will be reported in Section 4.

In order to efficiently utilize the information in (15), we can again apply Chamberlain's (1992) argument by letting $g_{it}(\theta) = s_{it-1}s_{it}(\Delta y_{it} - \Delta\mathbf{x}_{it}\beta - \delta_t\psi_{i,t,t-1} + \delta_{t-1}\psi_{i,t-1,t})$, collecting them for $t = 2, \ldots, T$ to

form $g_i(\theta)$, and then getting "$D_i'\Omega_i^{-1}g_i(\theta)$" as optimal unconditional moment functions, where D_i is the expected first derivative and Ω_i is the covariance matrix conditional on the exogenous variables, and $\theta = (\beta', \delta_1, \ldots, \delta_T)'$. As before, the D_i term is not difficult to get, but the Ω_i term involves covariances conditional on $s_{it_1}, s_{it_2}, s_{it_3}$ and s_{it_4} for up to four different periods t_1, t_2, t_3 and t_4. (An expression is given in the appendix for a more general "full-aggregation" estimator explained in Section 3.2.) Theoretically interesting as it is, pursuing efficiency to this extent does not seem to be practically important because, first, making it feasible is much harder than the CRE approach case, and, second, fixed effects are already removed so there is not much room for improvement when the individual effects show large variability. Furthermore, we can gain efficiency by gathering information from other nonconsecutive pairwise differences similarly to the within-group estimation, as explained below.

3.2 Full Aggregation of Pairwise Differences

In the FD estimation in Section 3.1, only consecutive pairs are considered and we condition on the event that $s_{it}s_{it-1} = 1$. Extension to the full within-group (WG) estimation is more complex, because it involves averaging over all selected observations. One possible approach is to consider all combinations of $s_{i1}, \ldots, s_{iT}$ and exhaustively evaluate $\mathrm{E}(v_{it}|s_{i1}, \ldots, s_{iT})$. This might be possible for very small T, but working it out for general T would be impractical, if not impossible.

An alternative approach is to consider pairs of periods separately. This method is in fact closely related with the WG estimation of models with fixed effects. The connection between them is revealed by the identity $T\sum_{t=1}^{T}(a_t - \bar{a})(b_t - \bar{b}) = \sum_{t=2}^{T}\sum_{r=1}^{t-1}(a_t - a_r)(b_t - b_r)$, where $\bar{a}$ and $\bar{b}$ are the sample means of a_t and b_t, respectively. Due to this identity, the WG estimator for balanced panel data is also written as

$$\left[\sum_{i=1}^{n}\sum_{t=2}^{T}\sum_{r=1}^{t-1}(\mathbf{x}_{it} - \mathbf{x}_{ir})'(\mathbf{x}_{it} - \mathbf{x}_{ir})\right]^{-1}\sum_{i=1}^{n}\sum_{t=2}^{T}\sum_{r=1}^{t-1}(\mathbf{x}_{it} - \mathbf{x}_{ir})'(y_{it} - y_{ir}).$$

That is, the usual WG estimator is also the pooled OLS estimator using all possible pairwise differences (with each pair used once). For panel data with selectivity, too, a corresponding "WG" estimator can be obtained by pooling all pairwise differences. As shown above, for pairs of periods t and r such that $s_{it}s_{ir} = 1$, we have $y_{it} - y_{ir} = (\mathbf{x}_{it} - \mathbf{x}_{ir})\beta + u_{it} - u_{ir}$, where

$$\mathrm{E}(u_{it} - u_{ir}|\mathbf{z}_i, s_{it}s_{ir} = 1) = \delta_t\psi(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr}) - \delta_r\psi(\mathbf{z}_i\pi_r, \mathbf{z}_i\pi_t; \rho_{tr}) = \delta_t\psi_{i,tr} - \delta_r\psi_{i,rt},$$

and hence,

$$\mathrm{E}(y_{it} - y_{ir}|\mathbf{z}_i, s_{it}s_{ir} = 1) = (\mathbf{x}_{it} - \mathbf{x}_{ir})\beta + \delta_t\psi_{i,tr} - \delta_r\psi_{i,rt}.$$

Once $\psi_{i,tr}$ are estimated using (13) and (14), where π_t are estimated at the first-step regression and ρ_{tr} by maximizing (8) given the π_t and π_r estimates, we can pool all possible differences together without duplication and then run pooled OLS. For example, if $T = 4$ and $s_i = (1, 0, 1, 1)$ for an i, then the pooled left-hand side variables and right-hand side variables are respectively

$$\begin{pmatrix} y_{i3} - y_{i1} \\ y_{i4} - y_{i1} \\ y_{i4} - y_{i3} \end{pmatrix} \text{ and } \begin{pmatrix} \mathbf{x}_{i3} - \mathbf{x}_{i1} & -\hat{\psi}_{i,13} & 0 & \hat{\psi}_{i,31} & 0 \\ \mathbf{x}_{i4} - \mathbf{x}_{i1} & -\hat{\psi}_{i,14} & 0 & 0 & \hat{\psi}_{i,41} \\ \mathbf{x}_{i4} - \mathbf{x}_{i3} & 0 & 0 & -\hat{\psi}_{i,34} & \hat{\psi}_{i,43} \end{pmatrix}$$

for that i, and the associated parameter vector is $(\beta', \delta_1, \ldots, \delta_4)'$. The equation for $y_{ir} - y_{it}$ is identical to the equation for $y_{it} - y_{ir}$ except that (-1) is multiplied to both sides, and thus we use only the pairs with $t > r$. It is programmatically more convenient to stack the equations

$$s_{it}s_{ir}(y_{it} - y_{ir}) = s_{it}s_{ir}(\mathbf{x}_{it} - \mathbf{x}_{ir})\beta + \delta_t s_{it}s_{ir}\hat{\psi}_{i,tr} - \delta_r s_{it}s_{ir}\hat{\psi}_{i,rt} + \text{error}_{i,tr}$$

for all t and r such that $t > r$ including zeros that appear when $s_{it}s_{ir} = 0$. Note that this pooling estimation, which we call the *fully aggregated pairwise differencing* (FAPD) estimator, is considerably easier to implement than the MDE suggested by Rochina-Barrachina (1999).

It is worth noting that the FAPD estimator is not a bias-corrected WG estimator. While the WG estimator is

$$\hat{\beta}_{wg} = \left[\sum_{i=1}^{n}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\right]^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(y_{it} - \bar{y}_i),$$

the FAPD estimator corrects the bias of a 'weighted' WG estimator

$$\begin{aligned}\hat{\beta}_{wwg} &= \left[\sum_{i=1}^{n} T_i \sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\right]^{-1} \sum_{i=1}^{n} T_i \sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(y_{it} - \bar{y}_i) \\ &= \left[\sum_{i=1}^{n}\sum_{t=2}^{T}\sum_{r=1}^{t-1} s_{it}s_{ir}(\mathbf{x}_{it} - \mathbf{x}_{ir})'(\mathbf{x}_{it} - \mathbf{x}_{ir})\right]^{-1} \sum_{i=1}^{n}\sum_{t=2}^{T}\sum_{r=1}^{t-1} s_{it}s_{ir}(\mathbf{x}_{it} - \mathbf{x}_{ir})'(y_{it} - y_{ir}),\end{aligned}$$

where the weight for i is $T_i = \sum_{t=1}^{T} s_{it}$, $\bar{\mathbf{x}}_i = T_i^{-1}\sum_{t=1}^{T} s_{it}\mathbf{x}_{it}$, and $\bar{y}_i = T_i^{-1}\sum_{t=1}^{T} s_{it}y_{it}$. This exact weighting is crucial, due to which bias can be corrected by evaluating the expected errors conditional on only the pairs $s_{it} = 1$ and $s_{ir} = 1$ for each pair of t and r. Without it, the evaluation of the average of v_{it} conditional on the full $(s_{i1}, \ldots, s_{iT})$ is required because

$$\begin{aligned}\hat{\beta}_{wg} = \beta + \left[\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=2}^{T}\sum_{r=1}^{t-1} s_{it}s_{ir}(\mathbf{x}_{it} - \mathbf{x}_{ir})'(\mathbf{x}_{it} - \mathbf{x}_{ir})\right]^{-1} \cdot \\ \sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=2}^{T}\sum_{r=1}^{t-1} s_{it}s_{ir}(\mathbf{x}_{it} - \mathbf{x}_{ir})'(u_{it} - u_{ir}),\end{aligned}$$

the expectation of which involves conditioning on $T_i = \sum_{t=1}^{T} s_{it}$. The expected values conditional on T_i seem hard to evaluate except for very small T. This is the reason why bias correction is not plausible for the original unweighted WG estimator but is straightforward for the weighted WG estimator.

Letting $e_{i,tr} = (y_{it} - y_{ir}) - (\mathbf{x}_{it} - \mathbf{x}_{ir})\beta - \delta_t \psi_{i,tr} + \delta_r \psi_{i,rt}$, the FAPD estimator makes use of the moment restrictions that $\mathrm{E}(s_{it} s_{ir} e_{i,tr} | \mathbf{z}_i) = 0$ for all $t > r$. The way FAPD combines the moment restrictions is not efficient because the error term is heteroskedastic and serially correlated. Nor is the MDE efficient because it does not account for observation-wise heteroskedasticity and serial correlation. It is again possible to derive efficient unconditional moment functions using Chamberlain's (1992) method, and an expression for infeasible optimal estimation is given in Appendix A.2. But the efficient estimation involves parameters which are very hard to estimate. Furthermore, once fixed effects are eliminated, the extra benefit of this optimal procedure seems limited, as was briefly discussed at the end of Section 3.1. The FAPD estimator seems already to be a useful method, which practitioners can use without much difficulty. Simulations in Section 4 confirm this assertion.

Finally, standard errors for the FAPD estimator can be obtained by using the Delta method explained in Appendix A.3. Block-bootstrapping also works well according to simulations.

4 Simulations and an Application

4.1 Simulation Results

In this section we present results from simulations for the estimators considered in the previous sections. Data are generated by

$$x_{it} = \mu_i + \xi_i^0 + x_{it}^0, \quad \mu_i \sim N(0,1),\ \xi_i^0 \sim N(0,1),\ x_{it}^0 \sim iid\ N(0,1), \tag{16a}$$

$$v_{it} = (\sigma_\eta^2 + 1)^{-1/2}(\sigma_\eta \eta_i + v_{it}^0), \quad \sigma_\eta = 1,\ \eta_i \sim N(0,1),\ v_{it}^0 \sim iid\ N(0,1), \tag{16b}$$

$$s_{it} = 1[\pi_{0t} + \pi_{1t} x_{it} + v_{it} > 0], \quad \pi_{0t} = \pi_{1t} = 0.5, \tag{16c}$$

$$u_{it} = \delta_t v_{it} + \varepsilon_{it},\ \ \varepsilon_{it} = \sigma_\mu \mu_i + (1 - \rho_\varepsilon^2)^{1/2} \varepsilon_{it}^0,\ \ \varepsilon_{it}^0 = \rho_\varepsilon \varepsilon_{it-1}^0 + e_{it}^0, \tag{16d}$$

$$\delta_t = 0.75, \quad e_{it}^0 \sim iid\ N(0,1),$$

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}, \quad \beta_0 = -1,\ \beta_1 = 1, \tag{16e}$$

where the components μ_i, ξ_i^0, x_{it}^0, η_i, v_{it}^0 and e_{it}^0 are mutually independent. The explanatory variable x_{it} in (16a) consists of a time-invariant component ($\mu_i + \xi_i^0$) correlated with the fixed effects in the main equation

and idiosyncratic innovations (x_{it}^0). The selection error in (16b) has zero mean and unit variance, and is serially correlated because of a time-invariant component η_i. It is also correlated with the main regression error u_{it} in (16d). The ε_{it} component of u_{it} in (16d) consists of random effects $\sigma_\mu \mu_i$ and idiosyncratic errors $(1-\rho_\varepsilon^2)^{1/2}\varepsilon_{it}^0$, where ε_{it}^0 is serially correlated if $\rho_\varepsilon \neq 0$ as e_{it}^0 is generated as iid. If $\rho_\varepsilon = 0$, then the feasible optimal estimator is indeed optimal, but otherwise inefficient. All the random variables are iid across i. The correlation coefficient between selection errors in different times is $\rho_{tr} = \text{Corr}(v_{it}, v_{ir}) = \text{E}(v_{it}v_{ir}) = \sigma_\eta^2/(\sigma_\eta^2+1) = 0.5$ for all $t \neq r$. The dependent variable y_{it} is observed if $s_{it} = 1$. Selection is endogenous here because v_{it} is correlated with u_{it}.

We first estimate π_{0t} and π_{1t} for every t using separate probit regressions. Wooldridge's (1995) POLS estimator, the CW estimator, and an optimal linear combination of POLS and CW (denoted PO-CW) are then calculated following the explanation in Section 2. For the feasible optimal estimator $\hat{\theta}_{opt}$ (denoted OPT1), we estimate ρ_{tr} by maximizing (8) for every pair (t, r) with $t > r$. The infeasible optimal estimator $\tilde{\theta}_{opt}$ of the CRE model (denoted IOPT1) is computed only after π_{0t} and π_{1t} are estimated by the initial probit regressions, that is, using the true values of $\text{E}(\varepsilon_{it}\varepsilon_{ir})$ and ρ_{tr}. The FD estimator and the fully aggregated PD (denoted FA in short) estimator in Section 3 are computed given the estimates of π_{0t}, π_{1t} and ρ_{tr}.

Table 1 reports the simulated bias and variance for the alternative estimators, together with the infeasible optimal estimator (IOPT1), of β_1 obtained from 10,000 replications for various σ_μ values with $\rho_\varepsilon = 0$, $n = 500$ and $T = 5$. The POLS, FD, and WG estimators without bias correction are reported as POLS0, FD0, and WG0, respectively. The infeasible IOPT1 is obtained by (7) with π_t replaced by the first-step probit estimators, while the other parameters such as $\text{E}(\varepsilon_{it}\varepsilon_{ir})$, δ_t, δ_r and ρ_{tr} are regarded as known. The reported results suggest that the uncorrected POLS0, FD0, and WG0 estimators are severely biased as shown in panel (a) of Table 1, while the corrected estimators have little bias. In terms of efficiency, the performance of POLS deteriorates as the standard deviation σ_μ of the individual effects increases. See especially the case $\sigma_\mu = 10$ in the 'POLS' column in panel (b) of Table 1. The feasible optimal estimator (OPT1) works well in terms of both bias and variance. When σ_μ is large, the variance of OPT1 is noticeably larger than that of the infeasible IOPT1. This seems to originate from inefficient POLS which is used in the estimation of Ω_i. If we use the PO-CW estimator instead, the simulated variance decreases from 4.0722 to 3.1240 as noted in Table 1. It is clear that CW is better than POLS in terms of variance for large σ_μ (compare 'POLS' and 'CW' for the cases with $\sigma_\mu \geq 1$ in panel (b) of Table 1) but worse than POLS for $\sigma_\mu = 0$. Whereas, the linear combination PO-CW looks at least as efficient as both POLS and CW asymptotically, although the simulated variance of PO-CW sometimes slightly exceeds that of POLS for $\sigma_\mu = 0$ and that of CW for $\sigma_\mu = 3, 5$. This

seems a small sample issue. In an unreported experiment with $n = 1{,}000$ and $\sigma_\mu = 5$, PO-CW showed a smaller variance than that of CW, and than that of POLS of course. It is also noticed that CW and PO-CW are inferior to the infeasible optimal estimator, but even for $\sigma_\mu = 10$, the variance of PO-CW is less than twice that of IOPT1 and mounts to only 15% of the POLS estimator's variance. It seems that the inefficiency of PO-CW relative to OPT1 is small although PO-CW is not optimal, which seems to be explained by the fact that the initial consistent parameter estimators for OPT1 are obtained from the inefficient POLS. If PO-CW is used in the OPT1 procedure instead of POLS, the performance of OPT1 improves. Finally, FD and FA are completely free from the individual effects, which is why the estimates are identical for different σ_μ values. FA is more efficient than FD, and also than IOPT1 except for $\sigma_\mu = 0$. FA being more efficient than IOPT1 is not contradictory to the fact that IOPT1 is efficient because the moment restrictions used by FA are different from those used by IOPT1. They are not rankable in terms of efficiency. We have also examined MDE for both the CRE and the fixed effects models. No substantial differences have been noticed in the performances of the pooled regression and its MDE version for both POLS and FA (results not reported). In fact, the pooled regressions turned slightly better in terms of efficiency for all the data generating processes for Tables 1–3, while MDE outperforms pooled estimators if the idiosyncratic error shows large temporal heteroskedasticity.

Table 2 reports simulation results for $\rho_\varepsilon = 0.5$ (serial correlation in the idiosyncratic error). The infeasible optimal estimator IOPT1 is obtained using the true values of $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir})$, but the feasible estimator OPT1 is obtained under the (wrong) assumption that $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir}) = \sigma_a^2 + \sigma_b^2[t = r]$. The results are similar to the case with $\rho_\varepsilon = 0$. Though inefficient, OPT1 still performs better than POLS, CW, and PO-CW. It makes sense that FD and FA are similar in efficiency, because serial dependence in $\Delta\varepsilon_{it}$ is about the same as that in ε_{it}.

Table 3 examines the performance of various estimators as n and T change. We choose $n \in \{200, 400\}$ and $T \in \{5, 10\}$ for the case $\sigma_\mu = 1$ and $\rho_\varepsilon = 0$. All the estimators improve as the sample size increases. IOPT1 is supposed to be efficient relative to POLS, CW, and PO-CW, but not necessarily relative to the FD and FA estimators, because the latter estimators utilize moment restrictions different from those used by IOPT1, OPT1, POLS, CW, and PO-CW. The performance of FA is remarkable for the considered data generating processes. It even outperforms the IOPT1. FA seems a good practical method if the practitioner is willing to estimate the pairwise correlation coefficients ρ_{tr} in the selection errors v_{it}.

Tables 4 and 5 examine how the asymptotic variances perform by reporting the simulated sizes of the test when $\rho_\varepsilon = 0$. Statistical tests are conducted by comparing t-statistics with the standard normal critical values. Table 4 reports results using the Delta method in Appendix A.3, and Table 5 by using the block

bootstrapping (with 100 bootstrap replications) method for $n = 500$ and $T = 5$. The simulated sizes are reasonable.

4.2 Application: Earnings Equation for Married Women

We now apply the estimation methods to a wage equation for married women using a subset of the Korean Labor and Income Panel Study (KLIPS) data. The sample consists of 1,849 married women aged 30 to 60, observed during 2012–2015. About 40% of the observations are in the labor force. The main equation is the panel-data version of Wooldridge's (2010) Example 19.6:

$$\ln(wage_{it}) = \alpha_i + \beta_1 exper_{it} + \beta_2 exper_{it}^2 + \beta_3 educ_{it} + \theta_t + u_{it},$$

where $wage_{it}$ is average hourly wage, $exper_{it}$ is job-market experience, and $educ_{it}$ is years of education. Individual effects and common time effects are accounted for by α_i and θ_t. Determinants of selection (labor force participation) are annual household income except wife's (*nwifeinc*), age in 2012 (*age*), and the number of children aged 0–19 (*nkids*) together with $exper$, $exper^2$, $educ$. The model is identical to Wooldridge's (2010) except that *nkids* replaces number of children less than six years of age and number of children between 6 and 18 inclusive due to data availability. The regressors are assumed to be strictly exogenous. The selection equation is estimated for each t using explanatory variables in all periods. That is, when $\mathbf{z}_{1,it}$ and $\mathbf{z}_{2,i}$ are respectively time-varying and time-invariant instruments for the selection equation, the selection variable s_{it} is regressed on $\mathbf{z}_{1,i1}, \ldots, \mathbf{z}_{1,iT}$ and $\mathbf{z}_{2,i}$ for each t.

Table 6 shows summary statistics. Results of the main equation estimation are presented in Table 7, where the reported robust standard errors are obtained by using the Delta method described in Appendix A.3. The standard errors are smaller for PO-CW, OPT1, and FA than for POLS. Efficiency gain is noticeable.

5 Conclusion

In this paper we consider various estimators that correct selectivity bias in linear panel data models with fixed effects. For CRE models, Wooldridge's (1995) POLS and MDE are convenient but the performance can be compromised when the variance of the individual effects is large. An efficient procedure is derived as a means to overcome this problem, and a feasible version is provided under the conventional random-effects assumption. The simulated performance of the feasible optimal estimator is remarkable, even though the procedure involves cumbersome operations.

A convenient common-weighting (CW) estimator is also examined. It is useful in practice but can be worse than POLS when the variance of random effects is small. An asymptotically optimal linear combination of POLS and CW is considered, which practically solves the large random-effects variance problem and is asymptotically at least as efficient as both POLS and CW.

This paper also examines two estimators that eliminate fixed effects by pairwise differencing. We discuss a bias-corrected first-difference estimator and a new fully-aggregated pairwise differencing estimator, the latter of which corrects bias in a weighted within-group estimator. While Rochina-Barrachina (1999) proposes pairwise differencing and combining the pairs by the minimum distance procedure, we pool the differences in a simpler estimation procedure. Both procedures are inefficient, and an efficient version is provided in the appendix although its practical usefulness seems minor. Pairwise differencing methods require estimating pairwise serial correlations in the error term v_{it} of the selection equation. We propose a simple and workable two-step procedure while avoiding bivariate or multivariate probit regressions which sometimes turn out to be unstable in practical applications.

Simulations show that both the optimal CRE estimator and the convenient suboptimal estimator work well for samples with large n in comparison to POLS and MDE. The fully aggregated pairwise differencing estimator exhibits a remarkable performance for the data generating processes that are considered. Standard errors that account for the generated regressors problem are derived by using the Delta method explained in Appendix A.3.

References

Alva, M., Gray, A., Mihaylova, B. and Clarke, P. (2014). The effect of diabetes complications on health-related quality of life: The importance of longitudinal data to address patient heterogeneity, *Health Economics*, 23, 487–500.

Balleer, A., Gehrke, B., Lechthaler, W. and Merkl, C. (2016). Does short-time work save jobs? A business cycle analysis, *European Economic Review*, 84, 99–122.

Baltagi, B. H. (2013). *Econometric Analysis of Panel Data*, 5th edition, Wiley.

Chamberlain, G. (1980). Analysis of covariance with qualitative data, *Review of Economic Studies*, 47, 225–238.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression, *Econometrica*, 60, 567–596.

Chen, X. and Flores, C. A. (2015). Bounds on treatment effects in the presence of sample selection and noncompliance: The wage effects of job corps, *Journal of Business and Economic Statistics*, 33, 523–540.

Dustmann, C. and Rochina-Barrachina, M. E. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations, *Econometrics Journal*, 10, 263–293.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika*, 10, 507–521.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample, *Metron*, 1, 3–32.

Greene, W. H. (2012). *Econometric Analysis*, 7th edition, Pearson.

Han, C. and Kim, B. (2011). A GMM interpretation of the paradox in the inverse probability weighting estimation of the average treatment effect on the treated, *Economics Letters*, 110, 163–165.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica*, 47, 153–161.

Hirano, K., Imbens, G. W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, 71, 1161–1189.

Hoonhout, P. and Ridder, G. (2017). Nonignorable attrition in multi-period panels with refreshment samples, *Journal of Business and Economic Statistics*, forthcoming.

Jiménez, G., Ongena, S., Peydró, J. and Saurina, J. (2014). Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking?, *Econometrica*, 82, 463–505.

Jochmans, K. (2015). Multiplicative-error models with sample selection, *Journal of Econometrics*, 184, 315–327.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model, *Econometrica*, 65, 1335–1364.

Machado, C. (2017). Unobserved selection heterogeneity and the gender wage gap, *Journal of Applied Econometrics*, forthcoming.

Maddala, G. S. and Mount, T. D. (1973). A comparative study of alternative estimators for variance components models used in econometric applications, *Journal of the American Statistical Association*, 68, 324–328.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Malikov, E., Kumbhakar, S. C. and Sun, Y. (2016). Varying coefficient panel data model in the presence of endogenous selectivity and fixed effects, *Journal of Econometrics*, 190, 233–251.

Mundlak, Y. (1978). On the pooling of time series and cross section data, *Econometrica*, 46, 69–85.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors, *International Economic Review*, 25, 221–247.

Revelli, F. (2013). Tax mix corners and other kinks, *Journal of Law and Economics*, 56, 741–776.

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

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution, *Journal of the Royal Statistical Society: Series B*, 23, 405–408.

Rupert, P. and Zanella, G. (2015). Revisiting wage, earnings, and hours profiles, *Journal of Monetary Economics*, 72, 114–130.

Sasaki, Y. (2015). Heterogeneity and selection in dynamic panel data, *Journal of Econometrics*, 188, 236–249.

Semykina, A. and Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection, *Journal of Econometrics*, 157, 375–380.

Semykina, A. and Wooldridge, J. M. (2017). Binary response panel data models with sample selection and self-selection, *Journal of Applied Econometrics*, forthcoming.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution, *Journal of the Royal Statistical Society: Series B*, 23, 223–229.

Vossmeyer, A. (2016). Sample selection and treatment effect estimation of lender of last resort policies, *Journal of Business and Economic Statistics*, 34, 197–212.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions, *Journal of Econometrics*, 68, 115–132.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, MIT Press.

A Mathematical Appendix

In this appendix, we prove Theorem 1, provide an expression for the infeasible optimal pairwise-differencing estimators, and derive standard errors for various feasible estimators considered in this paper.

A.1 Proof of Theorem 1

We first prove Theorem 1.

Proof of Theorem 1. Let $\Delta\hat{\theta} = \hat{\theta}_{pols} - \hat{\theta}_{cw}$ so that $\hat{\theta}(C) = \hat{\theta}_{pols} - C\Delta\hat{\theta}$. As we deal with asymptotic covariances, we assume that Ω is known. Let 'Avar' and 'Acov' denote the asymptotic variance and asymptotic coveriance, respectively. We have $\text{Avar}(\hat{\theta}(C)) = A_0 - A_1C' - CA_1' + CA_2C'$, where $A_0 = \text{Avar}(\hat{\theta}_{pols})$, $A_1 = \text{Acov}(\hat{\theta}_{pols}, \Delta\hat{\theta})$ and $A_2 = \text{Avar}(\Delta\hat{\theta})$. Using $C = C_* = A_1A_2^{-1}$, we have $\text{Avar}(\hat{\theta}(C_*)) = A_0 - A_1A_2^{-1}A_1'$, and thus,

$$\begin{aligned}\text{Avar}(\hat{\theta}(C)) - \text{Avar}(\hat{\theta}(C_*)) &= A_1A_2^{-1}A_1' - A_1C' - CA_1' + CA_2C' \\ &= (A_1 - CA_2)A_2^{-1}(A_1 - CA_2)',\end{aligned}$$

which is positive semidefinite. ■

A.2 Derivation of an Optimal Pairwise-Differencing Estimator

We next derive an efficient estimator based on pairwise differences under the assumption that π_t and ρ_{tr} are known. The moment restrictions are $\text{E}(s_{it}s_{ir}e_{i,tr}|\mathbf{z}_i) = 0$ for all $t > r$, where $e_{i,tr} = (y_{it} - y_{ir}) - (\mathbf{x}_{it} - \mathbf{x}_{ir})\beta - (\delta_t\psi_{i,tr} - \delta_r\psi_{i,rt})$. For notational brevity, let $\theta = (\beta', \delta_1, \ldots, \delta_T)'$ as before and $\mathbf{w}_{i,tr} =$

$(\mathbf{x}_{it} - \mathbf{x}_{ir}, \psi_{i,tr}\mathbf{d}_t - \psi_{i,rt}\mathbf{d}_r)$, where $\mathbf{d}_t$ is the tth row of I_T. Let $y_{i,tr} = y_{it} - y_{ir}$. Then the differenced equation is $y_{i,tr} = \mathbf{w}_{i,tr}\theta + e_{i,tr}$. The moment conditions are $\mathrm{E}(s_{it}s_{ir}e_{i,tr}|\mathbf{z}_i) = 0$ for every t and r such that $t > r$, where $e_{i,tr} = y_{i,tr} - \mathbf{w}_{i,tr}\theta$. We again apply Chamberlain's (1992) method. The expected first derivative times -1 is the matrix of the $T(T-1)/2$ rows of $\mathrm{E}(s_{it}s_{ir}\mathbf{w}_{i,tr}|\mathbf{z}_i) = p_{i,tr}\mathbf{w}_{i,tr}$, where $p_{i,tr} = \mathrm{E}(s_{it}s_{ir}|\mathbf{z}_i) = \Phi_2(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr})$ as before. It is more complicated to evaluate the conditional covariance matrix. For the covariance of $s_{it_1}s_{ir_1}e_{i,t_1r_1}$ and $s_{it_2}s_{ir_2}e_{i,t_2r_2}$ conditional on $\mathbf{z}_i$, note that $e_{i,tr} = (\varepsilon_{it} - \varepsilon_{ir}) + \delta_t(v_{it} - \psi_{i,tr}) - \delta_r(v_{ir} - \psi_{i,rt})$. For $t_1 > r_1$ and $t_2 > r_2$, we have

$$\mathrm{E}(s_{it_1}s_{ir_1}e_{i,t_1r_1}s_{it_2}s_{ir_2}e_{i,t_2r_2}|\mathbf{z}_i) = \Pr(s_{it_1}s_{ir_1}s_{it_2}s_{ir_2} = 1|\mathbf{z}_i)\,\mathrm{E}(e_{i,t_1r_1}e_{i,t_2r_2}|\mathbf{z}_i, s_{it_1}s_{ir_1}s_{it_2}s_{ir_2} = 1).$$

The first probability on the right-hand side is obtained by integrating the density of correlated standard normal random variables, and the second term $\mathrm{E}(e_{i,t_1r_1}e_{i,t_2r_2}|\mathbf{z}_i, s_{it_1}s_{ir_1}s_{it_2}s_{ir_2} = 1)$ is

$$\begin{aligned}
&\mathrm{E}[(\varepsilon_{it_1} - \varepsilon_{ir_1})(\varepsilon_{it_2} - \varepsilon_{ir_2})] + \delta_{t_1}\delta_{t_2}\,\mathrm{E}[(v_{it_1} - \psi_{i,t_1r_1})(v_{it_2} - \psi_{i,t_2r_2})|\mathbf{z}_i, a_i] \\
&- \delta_{t_1}\delta_{r_2}\,\mathrm{E}[(v_{it_1} - \psi_{i,t_1r_1})(v_{ir_2} - \psi_{i,r_2t_2})|\mathbf{z}_i, a_i] - \delta_{r_1}\delta_{t_2}\,\mathrm{E}[(v_{ir_1} - \psi_{i,r_1t_1})(v_{it_2} - \psi_{i,t_2r_2})|\mathbf{z}_i, a_i] \\
&+ \delta_{r_1}\delta_{r_2}\,\mathrm{E}[(v_{ir_1} - \psi_{i,r_1t_1})(v_{ir_2} - \psi_{i,r_2t_2})|\mathbf{z}_i, a_i],
\end{aligned}$$

where a_i denotes the event that $s_{it_1}s_{ir_1}s_{it_2}s_{ir_2} = 1$ for notational brevity. The first term is straightforward, because ε_{it} are assumed to be independent of v_{it}. For the second term, we have

$$\begin{aligned}
\mathrm{E}[(v_{it_1} - \psi_{i,t_1r_1})(v_{it_2} - \psi_{i,t_2r_2})|\mathbf{z}_i, a_i] = {}& \mathrm{E}(v_{it_1}v_{it_2}|\mathbf{z}_i, a_i) - \mathrm{E}(v_{it_1}|\mathbf{z}_i, a_i)\psi_{i,t_2r_2} \\
& - \psi_{i,t_1r_1}\,\mathrm{E}(v_{it_2}|\mathbf{z}_i, a_i) + \psi_{i,t_1r_1}\psi_{i,t_2r_2},
\end{aligned}$$

and other terms are expanded similarly. Typical terms to evaluate are $\mathrm{E}(v_{it_j}|\mathbf{z}_i, s_{it_1}s_{it_2}s_{it_3}s_{it_4} = 1)$ and $\mathrm{E}(v_{it_j}v_{it_k}|\mathbf{z}_i, s_{it_1}s_{it_2}s_{it_3}s_{it_4} = 1)$ for $j, k \in \{1, 2, 3, 4\}$, which depend on the joint distribution of $(v_{it_1}, v_{it_2}, v_{it_3}, v_{it_4})$. The case $t_1 = t_3$ and $t_2 = t_4$ is readily obtained using Rosenbaum's (1961) results and has already been done in Section 2. For other cases, the conditional moments can be obtained by Tallis's (1961) moment generating function of the multivariate truncated normal distribution. The first conditional moment is

$$\begin{aligned}
&\mathrm{E}(v_{it_j}|\mathbf{z}_i, s_{it_1}s_{it_2}s_{it_3}s_{it_4} = 1) \\
&\quad = p_{i,\cdot}^{-1}\left\{\rho_{t_jt_1}\phi(z_{it_1})\Phi_3(z^*_{i,t_2t_1}, z^*_{i,t_3t_1}, z^*_{i,t_4t_1}; \rho_{t_2t_3.t_1}, \rho_{t_2t_4.t_1}, \rho_{t_3t_4.t_1})\right. \\
&\qquad + \rho_{t_jt_2}\phi(z_{it_2})\Phi_3(z^*_{i,t_1t_2}, z^*_{i,t_3t_2}, z^*_{i,t_4t_2}; \rho_{t_1t_3.t_2}, \rho_{t_1t_4.t_2}, \rho_{t_3t_4.t_2}) \\
&\qquad + \rho_{t_jt_3}\phi(z_{it_3})\Phi_3(z^*_{i,t_1t_3}, z^*_{i,t_2t_3}, z^*_{i,t_4t_3}; \rho_{t_1t_2.t_3}, \rho_{t_1t_4.t_3}, \rho_{t_2t_4.t_3}) \\
&\qquad \left. + \rho_{t_jt_4}\phi(z_{it_4})\Phi_3(z^*_{i,t_1t_4}, z^*_{i,t_2t_4}, z^*_{i,t_3t_4}; \rho_{t_1t_2.t_4}, \rho_{t_1t_3.t_4}, \rho_{t_2t_3.t_4})\right\},
\end{aligned}$$

where $p_{i,\cdot} = \Pr(s_{it_1}s_{it_2}s_{it_3}s_{it_4} = 1|\mathbf{z}_i)$. For $j = 1$, $\rho_{t_1t_1} = \mathrm{E}(v_{it_1}v_{it_1}) = 1$, $z_{it_1} = \mathbf{z}_i\pi_{t_1}$, $z^*_{i,t_2t_1} = (1 - \rho^2_{t_1t_2})^{-1/2}(z_{it_2} - \rho_{t_1t_2}z_{it_1})$ and $\rho_{t_2t_3.t_1} = \{(1-\rho^2_{t_1t_2})(1-\rho^2_{t_1t_3})\}^{-1/2}(\rho_{t_2t_3} - \rho_{t_1t_2}\rho_{t_1t_3})$ is the first-order partial correlation coefficient between v_{it_2} and v_{it_3} holding v_{it_1} fixed. The expressions for other j are similar. Also,

$$
\begin{aligned}
&\mathrm{E}(v_{it_j}v_{it_k}|\mathbf{z}_i, s_{it_1}s_{it_2}s_{it_3}s_{it_4} = 1) \\
&= \rho_{t_jt_k} + p^{-1}_{i,\cdot}\Big[\rho_{t_jt_1}\rho_{t_kt_1}z_{it_1}\phi(z_{it_1})\Phi_3(z^*_{i,t_2t_1}, z^*_{i,t_3t_1}, z^*_{i,t_4t_1}; \rho_{t_2t_3.t_1}, \rho_{t_2t_4.t_1}, \rho_{t_3t_4.t_1}) \\
&\quad + \rho_{t_jt_2}\rho_{t_kt_2}z_{it_2}\phi(z_{it_2})\Phi_3(z^*_{i,t_1t_2}, z^*_{i,t_3t_2}, z^*_{i,t_4t_2}; \rho_{t_1t_3.t_2}, \rho_{t_1t_4.t_2}, \rho_{t_3t_4.t_2}) \\
&\quad + \rho_{t_jt_3}\rho_{t_kt_3}z_{it_3}\phi(z_{it_3})\Phi_3(z^*_{i,t_1t_3}, z^*_{i,t_2t_3}, z^*_{i,t_4t_3}; \rho_{t_1t_2.t_3}, \rho_{t_1t_4.t_3}, \rho_{t_2t_4.t_3}) \\
&\quad + \rho_{t_jt_4}\rho_{t_kt_4}z_{it_4}\phi(z_{it_4})\Phi_3(z^*_{i,t_1t_4}, z^*_{i,t_2t_4}, z^*_{i,t_3t_4}; \rho_{t_1t_2.t_4}, \rho_{t_1t_3.t_4}, \rho_{t_2t_3.t_4}) \\
&\quad + \rho_{t_jt_1}\Big\{\phi_2(z_{it_1}, z_{it_2}; \rho_{t_1t_2})\Phi_2(z^*_{i,t_3t_2|t_1}, z^*_{i,t_4t_2|t_1}; \rho_{t_3t_4.t_1t_2})(\rho_{t_kt_2} - \rho_{t_1t_2}\rho_{t_kt_1}) \\
&\quad + \phi_2(z_{it_1}, z_{it_3}; \rho_{t_1t_3})\Phi_2(z^*_{i,t_2t_3|t_1}, z^*_{i,t_4t_3|t_1}; \rho_{t_2t_4.t_1t_3})(\rho_{t_kt_3} - \rho_{t_1t_3}\rho_{t_kt_1}) \\
&\quad + \phi_2(z_{it_1}, z_{it_4}; \rho_{t_1t_4})\Phi_2(z^*_{i,t_2t_4|t_1}, z^*_{i,t_3t_4|t_1}; \rho_{t_2t_3.t_1t_4})(\rho_{t_kt_4} - \rho_{t_1t_4}\rho_{t_kt_1})\Big\} \\
&\quad + \rho_{t_jt_2}\Big\{\phi_2(z_{it_1}, z_{it_2}; \rho_{t_1t_2})\Phi_2(z^*_{i,t_3t_1|t_2}, z^*_{i,t_4t_1|t_2}; \rho_{t_3t_4.t_1t_2})(\rho_{t_kt_1} - \rho_{t_1t_2}\rho_{t_kt_2}) \\
&\quad + \phi_2(z_{it_2}, z_{it_3}; \rho_{t_2t_3})\Phi_2(z^*_{i,t_1t_3|t_2}, z^*_{i,t_4t_3|t_2}; \rho_{t_1t_4.t_2t_3})(\rho_{t_kt_3} - \rho_{t_2t_3}\rho_{t_kt_2}) \\
&\quad + \phi_2(z_{it_2}, z_{it_4}; \rho_{t_2t_4})\Phi_2(z^*_{i,t_1t_4|t_2}, z^*_{i,t_3t_4|t_2}; \rho_{t_1t_3.t_2t_4})(\rho_{t_kt_4} - \rho_{t_2t_4}\rho_{t_kt_2})\Big\} \\
&\quad + \rho_{t_jt_3}\Big\{\phi_2(z_{it_1}, z_{it_3}; \rho_{t_1t_3})\Phi_2(z^*_{i,t_2t_1|t_3}, z^*_{i,t_4t_1|t_3}; \rho_{t_2t_4.t_1t_3})(\rho_{t_kt_1} - \rho_{t_1t_3}\rho_{t_kt_3}) \\
&\quad + \phi_2(z_{it_2}, z_{it_3}; \rho_{t_2t_3})\Phi_2(z^*_{i,t_1t_2|t_3}, z^*_{i,t_4t_2|t_3}; \rho_{t_1t_4.t_2t_3})(\rho_{t_kt_2} - \rho_{t_2t_3}\rho_{t_kt_3}) \\
&\quad + \phi_2(z_{it_3}, z_{it_4}; \rho_{t_3t_4})\Phi_2(z^*_{i,t_1t_4|t_3}, z^*_{i,t_2t_4|t_3}; \rho_{t_1t_2.t_3t_4})(\rho_{t_kt_4} - \rho_{t_3t_4}\rho_{t_kt_3})\Big\} \\
&\quad + \rho_{t_jt_4}\Big\{\phi_2(z_{it_1}, z_{it_4}; \rho_{t_1t_4})\Phi_2(z^*_{i,t_2t_1|t_4}, z^*_{i,t_3t_1|t_4}; \rho_{t_2t_3.t_1t_4})(\rho_{t_kt_1} - \rho_{t_1t_4}\rho_{t_kt_4}) \\
&\quad + \phi_2(z_{it_2}, z_{it_4}; \rho_{t_2t_4})\Phi_2(z^*_{i,t_1t_2|t_4}, z^*_{i,t_3t_2|t_4}; \rho_{t_1t_3.t_2t_4})(\rho_{t_kt_2} - \rho_{t_2t_4}\rho_{t_kt_4}) \\
&\quad + \phi_2(z_{it_3}, z_{it_4}; \rho_{t_3t_4})\Phi_2(z^*_{i,t_1t_3|t_4}, z^*_{i,t_2t_3|t_4}; \rho_{t_1t_2.t_3t_4})(\rho_{t_kt_3} - \rho_{t_3t_4}\rho_{t_kt_4})\Big\}\Big].
\end{aligned}
$$

For example, for $j = 1$ and $k = 3$, we have

$$
z^*_{i,t_3t_2|t_1} = \left\{(1 - \rho^2_{t_1t_3})(1 - \rho^2_{t_2t_3.t_1})\right\}^{-1/2}(z_{it_3} - \beta_{t_3t_1.t_2}z_{it_1} - \beta_{t_3t_2.t_1}z_{it_2}),
$$

where $\beta_{t_3t_1.t_2}$ is the partial regression coefficient of v_{it_3} on v_{it_1} controlling v_{it_2}, and $\rho_{t_3t_4.t_1t_2}$ is the second-order partial correlation coefficient between v_{it_3} and v_{it_4} controlling v_{it_1} and v_{it_2}. Other conditional moments can be obtained similarly. Also, Tallis (1961) presents the evaluated first and second moments conditional on three different periods. With the expected first derivative and the covariance (conditional on $\mathbf{z}_i$) obtained

in this way, we can use Chamberlain's (1992) arguments to construct optimal unconditional moment restructions and a resulting optimal estimator.

A.3 Standard Errors for Multi-Step Estimators

In this appendix, we derive the asymptotic variance matrices of various two-step and multi-step estimators considered in this paper, where asymptotics are derived with fixed T as $n \to \infty$. In all cases, we express the estimator vector $\hat{\xi}$ (containing all the multi-step parameter estimators) as the Z-estimator satisfying $n^{-1}\sum_{i=1}^{n} g_i(\hat{\xi}) = 0$. Specifically, a Taylor series expansion about the true parameter ξ gives

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\hat{\xi}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\xi) + \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_i(\xi)}{\partial \xi'}\right]\sqrt{n}(\hat{\xi}-\xi) + o_p(1),$$

and thus, the limit variance of $\sqrt{n}(\hat{\xi}-\xi)$ equals $D^{-1}CD^{-1\prime}$ under usual regularity, where C is the limit covariance matrix of $n^{-1/2}\sum_{i=1}^{n} g_i(\xi)$ and D is the probability limit of $n^{-1}\sum_{i=1}^{n}\frac{\partial g_i(\xi)}{\partial \xi'}$. This limit variance is estimated by $\hat{D}^{-1}\hat{C}\hat{D}^{-1\prime}$, where $\hat{D} = n^{-1}\sum_{i=1}^{n}\frac{\partial g_i(\hat{\xi})}{\partial \xi'}$ and $\hat{C} = n^{-1}\sum_{i=1}^{n} g_i(\hat{\xi})g_i(\hat{\xi})'$. For the estimators considered in this paper, the $g_i(\cdot)$ functions can be shown to satisfy the regularity conditions necessary for the approximations and convergences. Theoretically interesting as it is, we do not pursue mathematical details here, and we rather presume that the regularity conditions are satisfied as a high level assumption.

Let $\pi = (\pi_1', \ldots, \pi_T')'$ and $\lambda(z) = \phi(z)/\Phi(z)$. All the estimation procedures begin with the probit regression of the selection equation for each t. The associated moment functions for the first order conditions are

$$g_{1i}(\pi) = (g_{1i1}', \ldots, g_{1iT}')', \quad g_{1it} = \mathbf{z}_i' h(\mathbf{z}_i\pi_t)[s_{it} - \Phi(\mathbf{z}_i\pi_t)], \tag{17}$$

suppressing the arguments, where $h(z) = \phi(z)/\{\Phi(z)[1-\Phi(z)]\}$. Given the probit estimator $\tilde{\pi}$ defined by the first order condition derived from (17), the POLS estimator is obtained based on $g_i = [g_{1i}', g_{2i}']'$, where

$$g_{2i}(\pi, \theta_{pols}) = W_i(\pi)'[S_i y_i - S_i W_i(\pi)\theta_{pols}], \tag{18}$$

where $W_i(\pi) = [\mathbf{w}_{i1}(\pi_1)', \ldots, \mathbf{w}_{iT}(\pi_T)']'$ with $\mathbf{w}_{it}(\pi_t) = [\mathbf{x}_{it}, \mathbf{z}_i, 0, \ldots, 0, \lambda(\mathbf{z}_i\pi_t), 0, \ldots, 0]$, the parameter vector is $\xi = (\pi', \theta_{pols}')'$. The asymptotic variance of $(\tilde{\pi}', \hat{\theta}_{pols}')'$ is derived by the method explained at the beginning of this section after centering and rescaling. Wooldridge (1995) also derives the asymptotic variance of POLS.

Next, for the CW estimation, let the parameter vector be $(\pi', \theta_{pols}', \theta_{cw}')'$. Then the CW estimator is based on (17), (18), and

$$g_{3i}(\pi, \theta_{pols}, \theta_{cw}) = W_i(\pi)'P_i(\pi)\Omega(\pi, \theta_{pols})^{-1}[S_i y_i - S_i W_i(\pi)\theta_{cw}], \tag{19}$$

where $\Omega(\pi,\theta) = n^{-1}\sum_{i=1}^{n}[S_i y_i - S_i W_i(\pi)\theta][S_i y_i - S_i W_i(\pi)\theta]'$. Note that the same parameter θ is used separately for the POLS estimation and the CW estimation in the next step, so we use the θ_{pols} and θ_{cw} notations to distinguish them. The asymptotic variance of the centered and rescaled $(\tilde{\pi}', \hat{\theta}'_{pols}, \hat{\theta}'_{cw})'$ is now obtained as before.

As the joint asymptotic distribution of $(\tilde{\pi}', \hat{\theta}'_{pols}, \hat{\theta}'_{cw})'$ has been obtained above, the limit distribution of $\hat{\theta}_* = (I - \hat{C}_*)\hat{\theta}_{pols} + \hat{C}_*\hat{\theta}_{cw}$ in (11) is easily obtained under the regularity that $\hat{C}_*$ converges in probability to a nonrandom matrix. This is obvious, and its estimation is straightforward.

For the optimal weighted IV estimator OPT1 in (9), we should also consider the estimator of ρ_{tr}, the maximizer of $\ln L(\rho)$ in (8), for all (t,r) pairs. The first order condition for maximizing (8) is $n^{-1}\sum_{i=1}^{n} g_{4i,tr}(\tilde{\pi}, \hat{\rho}_{\text{tr}}) = 0$, where

$$g_{4i,tr}(\pi,\rho_{tr}) = \frac{\phi_2(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r;\rho_{tr})}{\Phi_{2i,tr}(\rho_{tr})[1-\Phi_{2i,tr}(\rho_{tr})]}[s_{it}s_{ir} - \Phi_{2i,tr}(\rho_{tr})], \tag{20}$$

and $\Phi_{2i,tr}(\rho_{tr}) = \Phi_2(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r;\rho_{tr})$. (See Greene, 2012, for the derivative of $\Phi_2(\cdot,\cdot;\rho)$.) Next, the first order conditions for the pooled least squares estimation of the σ_a^2 and σ_b^2 parameters in $\mathrm{E}(\varepsilon_{it}\varepsilon_{ir}) = \sigma_a^2 + \sigma_b^2[t=r]$ are

$$g_{5i}(\pi,\boldsymbol{\rho},\theta_{pols},\boldsymbol{\varphi}) = \sum_{t=1}^{T}\sum_{r=1}^{T}\begin{pmatrix}1\\ 1[t=r]\end{pmatrix}\times \left\{\left[\frac{s_{it}e_{it}^{pols}s_{ir}e_{ir}^{pols} - \delta_t^{pols}\delta_r^{pols}m_{i,tr}(\pi,\rho_{tr})}{p_{i,tr}(\pi,\rho_{tr})}\right] - \varphi_a - \varphi_b[t=r]\right\}, \tag{21}$$

where $s_{it}e_{it}^{pols} = s_{it}e_{it}(\pi,\theta_{pols})$, $\boldsymbol{\rho}$ is the $T(T-1)/2\times 1$ vector of ρ_{tr} for all $t > r$, $\boldsymbol{\varphi} = (\varphi_a,\varphi_b)'$, $p_{i,tr}(\pi,\rho_{tr}) = \Phi_2(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r;\rho_{tr})$, $m_{i,tr}(\pi,\rho_{tr})$ as given in (6), and $s_{it}e_{it}(\pi,\theta) = s_{it}y_{it} - s_{it}\mathbf{w}_{it}(\pi_t)\theta$. Now the optimal estimator is based on the moment function

$$g_{6i}(\xi) = W_i(\pi)'P_i(\pi)\Omega_i(\pi,\rho,\theta_{pols},\varphi)^{-1}[S_i y_i - S_i W_i(\pi)\theta_{opt}], \tag{22}$$

where $\xi = (\pi',\rho',\theta'_{pols},\varphi',\theta'_{opt})'$ and $\Omega_i(\pi,\rho,\theta_{pols},\varphi)$ is the $T\times T$ matrix of

$$\omega_{i,tr}(\pi,\rho,\theta_{pols},\varphi) = p_{i,tr}(\pi,\rho_{tr})(\max(0,\varphi_a) + |\varphi_b|[t=r]) + \delta_t^{pols}\delta_r^{pols}m_{i,tr}(\pi,\rho_{tr}),$$

with $\theta_{pols} = (\beta'_{pols},\gamma'_{pols},\delta_1^{pols},\ldots,\delta_T^{pols})'$. The collected moment function vector is $g_i = [g'_{1i}, g'_{2i}, g'_{4i}, g'_{5i}, g'_{6i}]'$, where g_{4i} is the collection of $g_{4i,tr}$ in (20) for all $t > r$.

We next consider the FAPD estimator. The procedure uses the probit estimator of π and the ML-like estimator of the serial correlation parameters ρ_{tr}. Thus, the g_i function contains (17) and (20) for $t > r$. The

generated regressors are $\psi(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr})$, where the ψ function is given in (13). The remaining moment functions are

$$g_{7i}(\xi) = \sum_{t=2}^{T}\sum_{r=1}^{t-1}\begin{bmatrix}(\mathbf{x}_{it} - \mathbf{x}_{ir})' \\ \psi(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr})J_t - \psi(\mathbf{z}_i\pi_r, \mathbf{z}_i\pi_t; \rho_{tr})J_r\end{bmatrix} s_{it}s_{ir}$$
$$\times \left[(y_{it} - y_{ir}) - (\mathbf{x}_{it} - \mathbf{x}_{ir})\beta - \delta_t\psi(\mathbf{z}_i\pi_t, \mathbf{z}_i\pi_r; \rho_{tr}) + \delta_r\psi(\mathbf{z}_i\pi_r, \mathbf{z}_i\pi_t; \rho_{tr})\right],$$

where J_t is the tth column of I_T, and the parameter vector is $\xi = (\pi', \rho', \beta', \delta_1, \ldots, \delta_T)'$. The collected moment function vector is $g_i = [g_{1i}', g_{4i}', g_{7i}']'$, where g_{4i} is the collection of $g_{4i,tr}$ in (20) for all $t > r$ as before. The FD estimation is handled similarly.

Table 1: Comparison for various σ_μ values ($\rho_\varepsilon = 0$, 10,000 replications)

$$s_{it} = 1[0.5 + 0.5x_{it} + v_{it} > 0],\ v_{it} = (\eta_i + v_{it}^0)/\sqrt{2},\ v_{it}^0 \sim iid\ N(0,1),\ \eta_i \sim N(0,1),$$
$$y_{it} = -1 + x_{it} + u_{it},\ u_{it} = 0.75v_{it} + \varepsilon_{it},\ \varepsilon_{it} = \sigma_\mu \mu_i + (1-\rho_\varepsilon^2)^{1/2}\varepsilon_{it}^0,\ \varepsilon_{it}^0 = \rho_\varepsilon \varepsilon_{it-1}^0 + e_{it}^0,$$
$$e_{it}^0 \sim iid\ N(0,1),\ \mu_i \sim N(0,1),\ n = 500,\ T = 5.$$

(a) Simulated bias

	Without correction			With correction						
σ_μ	POLS0	FD0	WG0	POLS	CW	PO-CW	OPT1	IOPT1	FD	FA
0	-0.1518	-0.0840	-0.0933	-0.0053	-0.0051	-0.0079	-0.0075	-0.0063	-0.0081	-0.0093
1	-0.1518	-0.0840	-0.0933	-0.0060	-0.0065	-0.0074	-0.0080	-0.0080	-0.0081	-0.0093
2	-0.1518	-0.0840	-0.0933	-0.0067	-0.0074	-0.0068	-0.0075	-0.0083	-0.0081	-0.0093
3	-0.1518	-0.0840	-0.0933	-0.0075	-0.0078	-0.0067	-0.0059	-0.0083	-0.0081	-0.0093
5	-0.1518	-0.0840	-0.0933	-0.0089	-0.0083	-0.0066	-0.0021	-0.0081	-0.0081	-0.0093
10	-0.1517	-0.0840	-0.0933	-0.0125	-0.0088	-0.0064	0.0016	-0.0076	-0.0081	-0.0093

(b) Simulated variance $\times 100$

	Without correction			With correction						
σ_μ	POLS0	FD0	WG0	POLS	CW	PO-CW	OPT1	IOPT1	FD	FA
0	0.1079	0.1780	0.1190	0.5029	0.5806	0.5263	0.4738	0.4758	0.8516	0.5776
1	0.1237	0.1780	0.1190	0.7971	0.7183	0.6748	0.6082	0.6086	0.8516	0.5776
2	0.1710	0.1780	0.1190	1.6640	0.9004	0.8848	0.8063	0.7953	0.8516	0.5776
3	0.2498	0.1780	0.1190	3.1038	1.1087	1.1111	1.0390	0.9873	0.8516	0.5776
5	0.5022	0.1780	0.1190	7.7015	1.6962	1.7016	1.6503	1.4204	0.8516	0.5776
10	1.6852	0.1780	0.1190	29.2195	4.3476	4.2513	4.0722[a]	2.8149	0.8516	0.5776

Note: Estimates without correction are computed using observations with $s_{it} = 1$. POLS is Wooldridge's (1995) POLS estimator, CW is a common-weighting estimator that uses individual expected first derivatives and a common covariance, PO-CW is a feasible optimal linear combination of POLS and CW, OPT1 is the feasible optimal estimator, and IOPT1 is the infeasible optimal estimator which uses true parameters for all unknown parameters with the exception of π_t estimated by probit. FD is the bias-corrected pooled first-difference estimator, and FA is a feasible fully-aggregated pairwise-differencing estimator with proper correction terms.

[a] The simulated variance for $\sigma_\mu = 10$ of OPT1 is 3.1240 if the PO-CW estimators of δ_t are used instead of the POLS estimators.

Table 2: Comparison for various σ_μ values ($\rho_\varepsilon = 0.5$, 10,000 replications)

$$s_{it} = 1[0.5 + 0.5x_{it} + v_{it} > 0],\ v_{it} = (\eta_i + v_{it}^0)/\sqrt{2},\ v_{it}^0 \sim iid\ N(0,1),\ \eta_i \sim N(0,1),$$
$$y_{it} = -1 + x_{it} + u_{it},\ u_{it} = 0.75v_{it} + \varepsilon_{it},\ \varepsilon_{it} = \sigma_\mu \mu_i + (1-\rho_\varepsilon^2)^{1/2}\varepsilon_{it}^0,\ \varepsilon_{it}^0 = \rho_\varepsilon \varepsilon_{it-1}^0 + e_{it}^0,$$
$$e_{it}^0 \sim iid\ N(0,1),\ \mu_i \sim N(0,1),\ n = 500,\ T = 5.$$

(a) Simulated bias

	Without correction			With correction						
σ_μ	POLS0	FD0	WG0	POLS	CW	PO-CW	OPT1	IOPT1	FD	FA
0	-0.1520	-0.0843	-0.0935	-0.0053	-0.0058	-0.0070	-0.0079	-0.0073	-0.0081	-0.0092
1	-0.1520	-0.0843	-0.0935	-0.0060	-0.0067	-0.0068	-0.0080	-0.0080	-0.0081	-0.0092
2	-0.1519	-0.0843	-0.0935	-0.0067	-0.0074	-0.0066	-0.0067	-0.0080	-0.0081	-0.0092
3	-0.1519	-0.0843	-0.0935	-0.0074	-0.0078	-0.0066	-0.0045	-0.0079	-0.0081	-0.0092
5	-0.1519	-0.0843	-0.0935	-0.0089	-0.0082	-0.0065	-0.0009	-0.0077	-0.0081	-0.0092
10	-0.1519	-0.0843	-0.0935	-0.0125	-0.0088	-0.0064	0.0018	-0.0074	-0.0081	-0.0092

(b) Simulated variance $\times 100$

	Without correction			With correction						
σ_μ	POLS0	FD0	WG0	POLS	CW	PO-CW	OPT1	IOPT1	FD	FA
0	0.0865	0.0957	0.0862	0.5355	0.4761	0.4445	0.4225	0.3986	0.4634	0.4148
1	0.1018	0.0957	0.0862	0.8344	0.5535	0.5348	0.5074	0.4768	0.4634	0.4148
2	0.1487	0.0957	0.0862	1.7062	0.6976	0.6956	0.6727	0.6159	0.4634	0.4148
3	0.2271	0.0957	0.0862	3.1507	0.8955	0.9016	0.8870	0.7779	0.4634	0.4148
5	0.4786	0.0957	0.0862	7.7580	1.4812	1.4799	1.4654	1.1620	0.4634	0.4148
10	1.6595	0.0957	0.0862	29.2999	4.1383	4.0252	3.8482[a]	2.4427	0.4634	0.4148

Note: Estimates without correction are computed using observations with $s_{it} = 1$. See the notes in Table 1 for POLS, CW, PO-CW, OPT1, IOPT1, FD, and FA. IOPT1 is obtained correctly for $\rho_\varepsilon = 0.5$, while OPT1 is based on the false assumption that ε_{it}^0 is iid.
[a] The simulated variance for $\sigma_\mu = 10$ of OPT1 is 2.8341 if the PO-CW estimators of δ_t are used instead of the POLS estimators.

Table 3: Comparison for various (n, T) combinations ($\sigma_\mu = 1$, 10,000 replications)

$$s_{it} = 1[0.5 + 0.5x_{it} + v_{it} > 0],\ v_{it} = (\eta_i + v_{it}^0)/\sqrt{2},\ v_{it}^0 \sim iid\ N(0,1),\ \eta_i \sim N(0,1),$$
$$y_{it} = -1 + x_{it} + u_{it},\ u_{it} = 0.75v_{it} + \varepsilon_{it},\ \varepsilon_{it} = \sigma_\mu \mu_i + \varepsilon_{it}^0,\ \varepsilon_{it}^0 \sim iid\ N(0,1),\ \mu_i \sim N(0,1).$$

(a) Simulated bias

		Without correction			With correction						
n	T	POLS0	FD0	WG0	POLS	CW	PO-CW	OPT1	IOPT1	FD	FA
200	5	-0.1510	-0.0839	-0.0932	-0.0133	-0.0139	-0.0159	-0.0177	-0.0174	-0.0190	-0.0204
400	5	-0.1519	-0.0849	-0.0937	-0.0078	-0.0077	-0.0098	-0.0105	-0.0103	-0.0107	-0.0119
200	10	-0.1500	-0.0849	-0.0978	-0.0120	-0.0132	-0.0148	-0.0185	-0.0182	-0.0193	-0.0206
400	10	-0.1512	-0.0850	-0.0979	-0.0057	-0.0077	-0.0088	-0.0104	-0.0106	-0.0107	-0.0119

(b) Simulated variance $\times 100$

		Without correction			With correction						
n	T	POLS0	FD0	WG0	POLS	CW	PO-CW	OPT1	IOPT1	FD	FA
200	5	0.2985	0.4437	0.2918	1.8819	1.7022	1.7158	1.4747	1.4673	1.9185	1.2892
400	5	0.1506	0.2274	0.1478	0.9805	0.8837	0.8432	0.7535	0.7596	1.0563	0.7158
200	10	0.1468	0.2151	0.1309	1.1174	0.8577	0.8891	0.7110	0.7137	0.8956	0.5433
400	10	0.0721	0.1044	0.0643	0.6045	0.4371	0.4273	0.3729	0.3774	0.4765	0.3061

Note: Estimates without correction are computed using observations with $s_{it} = 1$. See the notes in Table 1 for POLS, CW, PO-CW, OPT1, IOPT1, FD, and FA.

Table 4: Simulated size of test using the Delta method ($n = 500$, $T = 5$, 1,000 replications)

$$s_{it} = 1[0.5 + 0.5x_{it} + v_{it} > 0],\ v_{it} = (\eta_i + v_{it}^0)/\sqrt{2},\ v_{it}^0 \sim iid\ N(0,1),\ \eta_i \sim N(0,1),$$
$$y_{it} = -1 + x_{it} + u_{it},\ u_{it} = 0.75v_{it} + \varepsilon_{it},\ \varepsilon_{it} = \sigma_\mu \mu_i + \varepsilon_{it}^0,\ \varepsilon_{it}^0 \sim iid\ N(0,1),\ \mu_i \sim N(0,1),$$
$$H_0 : \beta_1 = 1,\ H_1 : \beta_1 \neq 1.$$

(a) 10% significance level

σ_μ	POLS	CW	PO-CW	OPT1	FD	FA
0	0.109	0.113	0.124	0.097	0.093	0.103
1	0.123	0.117	0.115	0.107	0.093	0.103
2	0.115	0.111	0.117	0.100	0.093	0.103
3	0.115	0.103	0.116	0.094	0.093	0.103
5	0.113	0.102	0.114	0.081	0.093	0.103
10	0.114	0.090	0.098	0.063	0.093	0.103

(b) 5% significance level

σ_μ	POLS	CW	PO-CW	OPT1	FD	FA
0	0.054	0.060	0.070	0.047	0.052	0.046
1	0.061	0.055	0.070	0.054	0.052	0.046
2	0.064	0.055	0.058	0.053	0.052	0.046
3	0.063	0.056	0.060	0.044	0.052	0.046
5	0.064	0.052	0.060	0.039	0.052	0.046
10	0.063	0.041	0.051	0.026	0.052	0.046

(c) 1% significance level

σ_μ	POLS	CW	PO-CW	OPT1	FD	FA
0	0.016	0.015	0.023	0.013	0.005	0.007
1	0.011	0.011	0.019	0.013	0.005	0.007
2	0.012	0.009	0.014	0.009	0.005	0.007
3	0.015	0.006	0.013	0.009	0.005	0.007
5	0.014	0.007	0.009	0.005	0.005	0.007
10	0.017	0.004	0.004	0.003	0.005	0.007

Note: Variances are estimated by using the Delta methods explained in Appendix A.3. The t-statistics are compared with the critical values for the standard normal distribution.

Table 5: Simulated size of test using block bootstrap ($n = 500$, $T = 5$, 1,000 replications)

$$s_{it} = 1[0.5 + 0.5x_{it} + v_{it} > 0],\ v_{it} = (\eta_i + v_{it}^0)/\sqrt{2},\ v_{it}^0 \sim iid\ N(0,1),\ \eta_i \sim N(0,1),$$
$$y_{it} = -1 + x_{it} + u_{it},\ u_{it} = 0.75v_{it} + \varepsilon_{it},\ \varepsilon_{it} = \sigma_\mu \mu_i + \varepsilon_{it}^0,\ \varepsilon_{it}^0 \sim iid\ N(0,1),\ \mu_i \sim N(0,1),$$
$$H_0 : \beta_1 = 1,\ H_1 : \beta_1 \neq 1.$$

(a) 10% significance level

σ_μ	POLS	CW	PO-CW	OPT1	FD	FA
0	0.120	0.120	0.116	0.101	0.108	0.121
1	0.133	0.127	0.101	0.105	0.108	0.121
2	0.134	0.119	0.098	0.106	0.108	0.121
3	0.129	0.123	0.104	0.096	0.108	0.121
5	0.124	0.114	0.104	0.084	0.108	0.121
10	0.122	0.110	0.093	0.062	0.108	0.121

(b) 5% significance level

σ_μ	POLS	CW	PO-CW	OPT1	FD	FA
0	0.061	0.064	0.062	0.058	0.054	0.069
1	0.073	0.063	0.060	0.063	0.054	0.069
2	0.070	0.065	0.055	0.062	0.054	0.069
3	0.068	0.062	0.056	0.054	0.054	0.069
5	0.072	0.064	0.051	0.046	0.054	0.069
10	0.069	0.054	0.043	0.028	0.054	0.069

(c) 1% significance level

σ_μ	POLS	CW	PO-CW	OPT1	FD	FA
0	0.019	0.018	0.015	0.016	0.012	0.015
1	0.012	0.015	0.013	0.015	0.012	0.015
2	0.017	0.014	0.014	0.012	0.012	0.015
3	0.021	0.013	0.013	0.012	0.012	0.015
5	0.024	0.011	0.008	0.008	0.012	0.015
10	0.023	0.009	0.007	0.005	0.012	0.015

Note: The variance for each sample is estimated from 100 bootstrap replications. The *t*-statistics are compared with the critical values for the standard normal distribution.

Table 6: Summary statistics

Variable	Description	Total sample	Participants	Nonparticipants
inlf	Labor force participation	0.396	1.000	0.000
		(0.489)	(0.000)	(0.000)
wage	Average hourly earnings	–	10.549	–
	(thousand KRW)		(6.383)	
exper	Job-market experience (year)	9.582	13.763	6.836
		(7.707)	(7.217)	(6.721)
educ	Education (year)	12.522	12.737	12.380
		(2.970)	(2.943)	(2.980)
nwifeinc	Other household annual income	42.394	35.486	46.931
	(million KRW)	(27.411)	(23.076)	(29.040)
nkids	Number of children aged 0–19	1.071	1.062	1.076
		(0.983)	(0.946)	(1.006)
age	Age at year 2012	43.274	42.902	43.518
		(8.542)	(7.786)	(8.997)
Number of observations		7,396	2,932	4,464

Note: Sample means and sample standard deviations (in parentheses) are reported. Statistics for participants and nonparticipants are simple averages and standard deviations over the observations with $inlf = 1$ and $inlf = 0$, respectively.

Table 7: Estimation results of the earnings equation for married women

$\ln(wage)$	*exper*	*expersq*	*educ*	Wald test
POLS0	0.0823*** (0.0246)	-0.0008*** (0.0003)	0.0494* (0.0383)	–
FD0	0.0941** (0.0457)	-0.0006** (0.0003)	0.0137 (0.0435)	–
WG0	0.0760*** (0.0263)	-0.0005*** (0.0002)	0.0209 (0.0323)	–
POLS	0.1545*** (0.0471)	-0.0008*** (0.0003)	0.0519* (0.0376)	14.19***
CW	0.1398*** (0.0528)	-0.0007*** (0.0003)	0.0287 (0.0344)	9.87*
PO-CW	0.1757*** (0.0452)	-0.0006** (0.0003)	0.0584** (0.0323)	28.02***
OPT1	0.1344*** (0.0419)	-0.0007*** (0.0003)	0.0405 (0.0323)	12.15**
FD	0.0409 (0.0646)	-0.0005** (0.0003)	0.0131 (0.0438)	14.08***
FA	0.0547* (0.0410)	-0.0005*** (0.0002)	0.0187 (0.0325)	10.39**

Note: The dependent variable is log of average hourly earnings, $\ln(wage)$. Time dummies are included but the results are suppressed. Standard errors for POLS0, FD0, and WG0 are estimated by the cluster-robust variance estimator, while those for POLS, CW, PO-CW, OPT1, FD, and FA are obtained by using the Delta method explained in Appendix A.3. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. The Wald test statistics are for joint significance of the correction terms, which are distributed as χ^2_4 under the null.