Discussion Paper Series
No. 1801
March 2018

Bias Reduction by Imputation for Linear Panel Data Models with Nonrandom Missing

Goeun Lee Chirok Han

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

Bias Reduction by Imputation for Linear Panel Data Models with Nonrandom Missing*

Goeun Lee† Chirok Han‡

March 2018

Abstract

When no variables are observed for endogenous non-respondents of panel data, bias correction is available only for a limited class of instrumental variable estimators, which require strong conditions for consistency and often suffer from substantial efficiency loss. In this paper we introduce a convenient alternative method of imputing the missing explanatory variables and then using standard bias-correction procedures for sample selection. Various bias-corrected estimators are derived and their performances are compared by Monte Carlo experiments. Results verify efficiency loss by the instrumental variable estimators and suggest that the imputation method is practically useful if it is applied to first-difference regression.

Keywords: Attrition, missing, nonresponse, bias-correction, panel data, imputation

JEL Classification: C23

---

*This work was supported by the Korean Government (NRF-2014S1A2A2027803) and by Korea University (K1802771). The authors thank Professor Sangsoo Park for valuable comments and suggestions.

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

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

1 Introduction

Attrition and nonresponse are potentially a serious problem in panel studies (Baltagi, 2014). If nonresponse occurs randomly or depending on exogenous variables, then the conventional estimation methods such as the within-group and the first-difference regression using the full unbalanced panel data or a balanced subset provide consistent estimators of population characteristics. But if the nonresponse mechanism depends on incomplete endogenous variables, then the available sample of unbalanced panels loses representativeness, and estimators based on the unbalanced panel data or a balanced subset are inconsistent in general.

When the exogenous variables are completely observed and nonresponse occurs only in the dependent variable as most cross-sectional and panel sample-selection studies consider, there are multiple methods of correcting the endogenous sample-selection bias. The most popular one is probably Heckman's (1976, 1979) bias-correction by augmenting the equation with the inverse Mills ratios (IMR) in order to control for endogenous selectivity. Wooldridge (1995) extends this approach to panel data, and recently Han and Lee (2017) derive efficient estimation methods.

On the other hand, bias-correction is a challenging task due to lack of information if no variables are observed for the non-respondents. One possible method available attrition in an 'absorbing' state (that is, when nonrespondents never return to the sample) is to specify the 'response' equation as $a_{it} = I(\pi_{t0} + \mathbf{x}_{it-1}\pi_{t1} + v_{it} > 0)$ with $v_{it} \sim N(0,1)$ conditional on $a_{it-1} = 1$ and $\mathbf{x}_{it-1}$, where $\mathbf{x}_{it-1}$ is a set of *lagged* exogenous variables observed as oppposed to the incomplete current exogenous variables (see Wooldridge, 2010). The key point of using the lagged variables is that they are observed as long as unit i was in the sample in the previous period no matter whether it drops out of the sample at period t. This idea leads to a bias-corrected instrumental variable (IV) estimation as explained in Section 2 later. Although this method makes a guideline for correcting the attrition bias at least in some (limited) cases (see Section 2), its nature as an IV estimation inevitably causes the resulting estimator to lose substantial efficiency in practice. In the present paper we verify by experiments that efficiency loss by IV estimation is indeed serious.

Though not much mentioned systematically in the econometric literature, there exists another convenient and intuitive method. While the IV estimation method collects information on the response mechanism using past observations of the explanatory variables in a particular way, we can also utilize past information in panel data for filling in the missing exogenous variable

values, after which the standard bias-correction methods (see, e.g., Wooldridge, 1995, Rochina-Barranchina, 1999, and Han and Lee, 2017) are applied to the imputed data. This method is referred to as 'bias correction after imputation' (BCI, hereafter) throughout the paper.

The main purpose of the present paper is to examine the usefulness of this BCI approach for panel data models. Although imputation inevitably involves measurement error in the explanatory variables, thus introducing inconsistency to estimators, we suspect that this bias is limited because only missing $\mathbf{x}_{it}$ values are imputed, at least as long as the performance of imputation is reasonable. It would naturally be desirable to evaluate bias due to data-imputation analytically, but achieving this goal looks rather challenging, if not totally impossible. We thus employ the Monte Carlo experment method in this study for the examination of BCI and for the comparison of various estimators, while leaving algebraic analysis as an interesting future research topic.

To briefly summarize what is found in this study, the results are optimistic to the BCI estimators especially when BCI is combined with the first-difference (FD) estimation. The conventional uncorrected estimators (using the full unbalanced panel data or the maximal balanced subset) show substantial bias when missing or attrition occurs endogenously, and efficiency loss by the IV bias-correction methods is remarkable. The BCI estimators applied to FD estimation exhibit remarkable bias reduction among other considered estimators.

The rest of the paper is organized as follows. We introduce the model and various estimators in Section 2. The estimators include the conventional biased estimators using the full unbalanced panel data or using the maximal balanced subset, a few bias-corrected IV estimators, and the BCI estimators. The IV and BCI estimation methods are considered for two alternative modes of nonresponse. The first is the case where non-respondents return to the sample whenever they choose to. The second type is the 'pure attrition', in which non-respondents never return and attrition is in an absorbing state. Different assumptions are made for the two attrition modes, and corresponding different estimators are derived in Section 2. Section 3 reports experiment results and implications are discussed. Section 4 concludes.

2 Models and Estimators

The linear panel data model we consider is

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

where α_i is the unobserved individual heterogeneity and $\mathbf{x}_{it}$ is the vector of regressors which are strictly exogenous to the idiosyncratic errors u_{it}. In this paper our main concern is that both $\mathbf{x}_{it}$ and y_{it} are unobserved for some i and t, which is in counter to the case considered by conventional sample-selection studies in which $\mathbf{x}_{it}$ is completely observed and only y_{it} is possibly missing.

Let a_{it} be the indicator of being observed, i.e., $a_{it} = 1$ if and only if $(y_{it}, \mathbf{x}_{it})$ is observed. In our experiments to follow, we assume that the initial $(y_{i1}, \mathbf{x}_{i1})$ is observed for all i, that is, $a_{i1} \equiv 1$, which means that the initial sample represent the population.

Observations can be missing for $t \geq 2$ in various manners. As briefly explained in the introduction, we consider two leading cases: the 'general missing' case and the 'pure attrition'. In the 'general missing' case, units may leave and re-enter the sample at any time, while in the pure attrition case, a dropout is permanent and a non-respondent never returns. (See Little and Rubin, 2002, and Wooldridge, 2010, for related terms.)

For both types of missingness, the conventional estimators using the full unbalanced panel or the maximal balanced subset can be inconsistent if missing is endogenous in the sense that it is correlated with the regression error. For example, the pooled ordinary least squares (POLS) estimator may be inconsistent if a_{it} is correlated with $\varepsilon_{it} = (\alpha_i - \mathrm{E}\,\alpha_i) + u_{it}$ conditional on $\mathbf{x}_{it}$ because

$$\begin{aligned}\hat{\beta}_{pols} &= \left(\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}\mathbf{x}_{it}'\mathbf{x}_{it}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}\mathbf{x}_{it}'y_{it} \\ &= \beta + \left(\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}\mathbf{x}_{it}'\mathbf{x}_{it}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}\mathbf{x}_{it}'\varepsilon_{it},\end{aligned}$$

where $\mathbf{x}_{it}$ contains a constant term for convenience. Also, the first-difference (FD) estimator using the maximal balanced subset is inconsistent if the indicator $a_i = \prod_{s=1}^{T} a_{is}$ is correlated with the idiosyncratic error term conditional on $\Delta\mathbf{x}_{it}$ because

$$\begin{aligned}\hat{\beta}_{fd} &= \left(\sum_{i=1}^{n}\sum_{t=1}^{T} a_i\Delta\mathbf{x}_{it}'\Delta\mathbf{x}_{it}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} a_i\Delta\mathbf{x}_{it}'\Delta y_{it} \\ &= \beta + \left(\sum_{i=1}^{n}\sum_{t=1}^{T} a_i\Delta\mathbf{x}_{it}'\Delta\mathbf{x}_{it}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} a_i\Delta\mathbf{x}_{it}'\Delta u_{it},\end{aligned}$$

where $\mathbf{x}_{it}$ contains only the time-varying regressors. Statistical properties of other estimators such as the random-effects generalized least squares, various population-averaged model estimators, and the within-group estimators can be derived in a similar fashion. Readers are referred to Wooldridge (2010) for detailed discussions.

When standard estimators are inconsistent due to endogenous nonresponse, if $\mathbf{x}_{it}$ is completely observed and missingness is confined to only y_{it}, then there are available standard methods of correcting or reducing biases under some suitable distributional assumptions. For example, when a_{it} is determined by some observed exogenous variables and other unobservable factors, reseachers often assume a probit specification for a_{it} and correct estimator bias by including the IMR's (inverse Mills ratios) in the right-hand side of the main equation (see Heckman, 1976, 1979, for cross-sectional models, Wooldridge, 1995, for panel data models, and Han and Lee, 2017, for efficient estimation with panel data). This strategy is, however, unavailable if $\mathbf{x}_{it}$ is also incomplete, simply because the probit regression of the selection equation (for a_{it}) requires $\mathbf{x}_{it}$ to be available. Sometimes deterministic elements (such as age and period dummies) of $\mathbf{x}_{it}$ are used as regressors in the selection equation estimation, but practical usefulness of this strategy is rather limited due to the weak explanatory power of those determimistic factors. In the pure attrition case, instrumental variable estimation has been mentioned by Wooldridge (2010) as a possible solution, but according to the authors' simulations to be reported later, the resulting estimators are seriously inefficient especially if fixed effects are handled by differencing.

An alternative strategy is to impute the missing exogenous variable values using their lagged values. Once $\mathbf{x}_{it}$ is made complete by imputation, the issue is reduced to bias-correction. As Wooldridge's IV approach is one way of using lagged variables for bias correction, so is the imputation method another way of utilizing information contained in lagged variables for bias correction. As explained in the introduction, our main goal is to examine the statistical properties of the bias correction using the imputed data (that is, BCI) in comparison to other methods.

We now introduce in detail various estimators considered in the present study that are available in case where both y_{it} and $\mathbf{x}_{it}$ are possibly missing altogether. We especially consider strategies that can be applied to the pooled regression, the correlated random-effects (CRE) regression, which corresponds to the pooled regression with $\mathbf{x}_{it}$ replaced with $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ in the balanced panel case, and the first-difference (FD) regression. The popular within-group regression is not considered here because its bias correction is unclear. For these three classes of regression, we consider estimation without bias-correction using the full unbalanced panel data, estimation without bias-correction using the maximal balanced panel data, the bias-corrected IV estimation, and the BCI (bias-correction after imputation) estimation. Details follow below.

Estimation using full data or a balanced subset without correction

Estimation using the full (unbalanced panel) data and estimation using the maximal balanced panel data without bias correction are defined straightforwardly.

IV estimation with bias correction for pure attrition without imputation

The IV approach, implemented based on a brief remark in Wooldridge (2010), is available only for pure attrition and is derived as follows. Under pure attrition, $\mathbf{x}_{it}$ is observed only if $\mathbf{x}_{it-1}$ is, and lagged explanatory variables are used for the probit regression, based on which the IMR's are calculated. (For the general missing case, in contrast, lagged regressors may not be observed for some units observed at t, and thus lagged explanatory variables may not be used as instruments unless they are imputed. But if the missing $\mathbf{x}_{it}$ are fully imputed, then IV estimation is not required anyway.)

The way in which the correction terms are calculated is different for the pooled regression and the FD regression. For the pooled regression, consider the main equation

$$y_{it} = \alpha + \mathbf{x}_{it}\beta + \varepsilon_{it}, \quad \varepsilon_{it} = \alpha_i - \mathrm{E}\,\alpha_i + u_{it}. \tag{2}$$

Because $\mathbf{x}_{it}$ is incomplete due to attrition, it cannot be used as regressor in the selection equation, and one can consider using $\mathbf{x}_{it-1}$ as instruments for bias correction. Specifically, the selection equation to fit is specified as

$$a_{it} = I(\pi_{t0} + \mathbf{x}_{it-1}\pi_{t1} + v_{it} > 0), \quad v_{it} \sim N(0,1), \tag{3}$$

conditional on $a_{it-1} = 1$. Note that v_{it} may or may not depend on $\mathbf{x}_{it}$ conditional on $\mathbf{x}_{it-1}$. The conditional normality of v_{it} implies that $\mathrm{E}(v_{it}|\mathbf{x}_{it-1}, a_{it} = 1) = \lambda(\pi_{t0} + \mathbf{x}_{it-1}\pi_{t1})$, where $\lambda(\cdot)$ denotes the IMR, the ratio of the density of the standard normal distribution to its cumulative probability. Under the further assumption that $\varepsilon_{it} = \delta_t v_{it} + e_{it}$, where e_{it} is independent of $\mathbf{x}_{it-1}$ and v_{it}, (2) and (3) together imply that

$$\mathrm{E}(y_{it}|\mathbf{x}_{it-1}, a_{it} = 1) = \alpha + \mathrm{E}(\mathbf{x}_{it}|\mathbf{x}_{it-1}, a_{it} = 1)\beta + \delta_t\lambda_{it}, \quad \lambda_{it} = \lambda(\pi_{t0} + \mathbf{x}_{it-1}\pi_{t1}). \tag{4}$$

Thus, β can be estimated using the IV regression of y_{it} on $\mathbf{x}_{it}$ and the interaction of the period dummies and the estimated IMR's, using $\mathbf{x}_{it-1}$ and the interaction terms as instruments, for the observations with $a_{it} = 1$. Note that this IV estimator is consistent if the condition in (4)

is valid, which naturally requires that there are no fixed effects (correlated with $\mathbf{x}_{it}$) and that $\varepsilon_{it} = \delta_t v_{it} + e_{it}$ and e_{it} is independent of $\mathbf{x}_{it-1}$ among others.

Bias correction by IV explained so far does not readily extend to the CRE estimation where $\bar{\mathbf{x}}_i$ is included as extra controls for fixed effects, unless good instruments are found for $\bar{\mathbf{x}}_i$. On the other hand, the FD estimation eliminates the fixed effects, and bias correction is available under some (strong) assumptions. One such method is discussed in Wooldridge (2010) as follows. The main equation is differenced into $\Delta y_{it} = \Delta \mathbf{x}_{it}\beta + \Delta u_{it}$ in order to eliminate the fixed effects, and the selection equation is again specified as $a_{it} = I(\pi_{t0} + \mathbf{x}_{it-1}\pi_{t1} + v_{it} > 0)$, where v_{it} is distributed as $N(0,1)$ conditional on the event that $a_{it-1} = 1$ (and conditional on the exogenous variables). Importantly, Δu_{it} and v_{it} are assumed to have the relationship $\Delta u_{it} = \delta_t v_{it} + e_{it}$ this time, where e_{it} is independent of v_{it} and $\mathbf{x}_{it-1}$. Then, in a similar fashion as in the pooled regression, one can construct the IMR's for each t. The differenced main equation is accordingly augmented with the IMR's and the β parameter can be estimated by the IV regression using $\mathbf{x}_{it-1}$ and the inverse Mills ratios as instrument. If $\mathbf{x}_{it}$ does not affect attrition conditional on $\mathbf{x}_{it-1}$, then least squares can be used in the final step regression (of the equation augmented with the IMR's) instead of the IV estimation as Wooldridge (2010) notes; otherwise, IV regression is required.

It is important to note that v_{it} should be serially independent for the consistency of the resulting bias-corrected IV estimator, because otherwise the condition that $v_{it} \sim N(0,1)$ conditional on $a_{it-1} = 1$ is likely to be violated. Also, the condition that $\Delta u_{it} = \delta_t v_{it} + e_{it}$ is a strong requirement. If $u_{it} = \delta_t^0 v_{it} + e_{it}^0$ for some independent e_{it}^0 instead, then we have $\Delta u_{it} = \delta_t^0 v_{it} - \delta_{t-1}^0 v_{it-1} + e_{it}^0 - e_{it-1}^0$, which is hardly reduced to $\delta_t v_{it} + e_{it}$ for some δ_t with $\mathrm{E}(e_{it}|\mathbf{x}_{it-1}, a_{it} = 1) = 0$ and e_{it} independent of v_{it} at the same time. For example, if $\delta_t^0 \equiv \delta^0 \neq 0$, $a_{it} = a_{it-1} I(\pi_{t0} + \mathbf{x}_{it}\pi_{t1} + v_{it} > 0)$, and all of $\mathbf{x}_{it}$, v_{it} and e_{it} are independent mutually and across all t, then $\delta_t = \delta^0$ and $e_{it} = e_{it}^0 - e_{it-1}^0 - \delta^0 v_{it-1}$ (in order for v_{it} and e_{it} to be mutually independent) so that

$$\begin{aligned}
\mathrm{E}(e_{it}|\mathbf{x}_{it-1}, a_{it} = 1) &= -\delta^0 \,\mathrm{E}(v_{it-1}|\mathbf{x}_{it-1}, a_{it} = 1) \\
&= -\delta^0 \,\mathrm{E}(v_{it-1}|\mathbf{x}_{it-1}, a_{it-1} = 1, \pi_{t0} + \mathbf{x}_{it}\pi_{t1} + v_{it} > 0) \\
&= -\delta^0 \,\mathrm{E}(v_{it-1}|\mathbf{x}_{it-1}, a_{it-1} = 1) \neq 0
\end{aligned}$$

in the pure attrition case.

Bias correction after data imputation (BCI)

The main subject of investigation in the present paper is bias correction after the imputation of $\mathbf{x}_{it}$ (BCI). There are many sophisticated data imputation methods, and in this paper we consider a simple sequential substitution starting from $t = 2$. For $t = 2$, $\mathbf{x}_{i2}$ is regressed on $\mathbf{x}_{i1}$ by OLS using the units with $a_{i2} = 1$ (note that $a_{i1} \equiv 1$ so $\mathbf{x}_{i1}$ is complete). Then the missing $\mathbf{x}_{i2}$ are replaced with the predicted values $\hat{\mathbf{x}}_{i2} = \hat{\gamma}_{20} + \mathbf{x}_{i1}\hat{\gamma}_{21}$, where $\hat{\gamma}_{2j}$ are the OLS estimates. For $t = 3$, $\mathbf{x}_{i3}$ is regressed on $\mathbf{x}_{i2}$ using the observations with $a_{i2} = a_{i3} = 1$, and then the missing $\mathbf{x}_{i3}$ values are replaced with $\hat{\mathbf{x}}_{i3} = \hat{\gamma}_{30} + [a_{i2}\mathbf{x}_{i2} + (1 - a_{i2})\hat{\mathbf{x}}_{i2}]\hat{\gamma}_{31}$, where $\hat{\gamma}_{30}$ and $\hat{\gamma}_{31}$ are the OLS estimates. Note that the regressions for imputation are run only using the observed values, and the actual imputation is conducted using the observed values if available and the imputed values otherwise. It is also possible that $\mathbf{x}_{i3}$ is regressed on $\mathbf{x}_{i2}$ and $\mathbf{x}_{i1}$ using only the observed values or also using the imputed values, but pursuing the details is not the purpose of the present study. Interested readers are referred to Little and Rubin (2002). Other t's are treated similarly.

After $\mathbf{x}_{it}$ is made complete by imputation, we apply the standard bias-correction procedures for the pooled regression, the CRE regression, and the FD regression. The specifics are different for the general missing case and the pure attrition case.

(i) *The general missing case:* For the pooled regression in the general missing case, the IMR's are obtained for each t by the probit regression of a_{it} on the imputed explanatory variables. Then y_{it} is regressed (using the observed units) on $\mathbf{x}_{it}$ and the IMR's interacted with the period dummies, the latter of which are introduced in order to account for the sample-selection bias. Extension to the CRE regression is straightforwardly done by replacing $\mathbf{x}_{it}$ with $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$. Correction for the FD regression can be done under general conditions about serial correlation in v_{it} (the selection equation error) and the relationship between u_{it} and v_{it}. To explain how, let $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$. From $\Delta y_{it} = \Delta\mathbf{x}_{it}\beta + \Delta u_{it}$, we have

$$\begin{aligned}
&\mathrm{E}(\Delta y_{it}|\mathbf{x}_i, a_{it}a_{it-1} = 1) = \Delta\mathbf{x}_{it}\beta + \mathrm{E}(\Delta u_{it}|\mathbf{x}_i, a_{it}a_{it-1} = 1), \\
&\mathrm{E}(\Delta u_{it}|\mathbf{x}_i, a_{it}a_{it-1} = 1) = \delta_t \,\mathrm{E}(v_{it}|\mathbf{x}_i, a_{it}a_{it-1} = 1) - \delta_{t-1}\,\mathrm{E}(v_{it-1}|\mathbf{x}_i, a_{it}a_{it-1} = 1),
\end{aligned}$$

where the second line is derived under the assumption that $u_{it} = \delta_t v_{it} + e_{it}$ and that $(e_{i1}, \ldots, e_{iT})$ is independent of all $\mathbf{x}_{it}$ and v_{it}. Then, under the assumption that v_{it} and v_{it-1} are jointly normal (with unit variances and correlation ρ_{vt}) and $a_{it} = I(z_{it} + v_{it} > 0)$, where z_{it} is a deterministic

function of $\mathbf{x}_i$ (e.g., $z_{it} = \pi_{t0} + \mathbf{x}_{it}\pi_{t1}$), we have

$$\mathrm{E}(v_{it}|\mathbf{x}_i, a_{it}a_{it-1} = 1) = \psi(z_{it}, z_{it-1}; \rho_{vt}),$$
$$\mathrm{E}(v_{it-1}|\mathbf{x}_i, a_{it}a_{it-1} = 1) = \psi(z_{it-1}, z_{it}; \rho_{vt}),$$

where

$$\psi(a, b; \rho) = \frac{\phi(a)\Phi(b^*)}{\Phi_2(a, b, ; \rho)} + \rho\frac{\phi(b)\Phi(a^*)}{\Phi_2(a, b; \rho)},$$

$a^* = (a-\rho b)(1-\rho^2)^{-1/2}$, $b^* = (b-\rho a)(1-\rho^2)^{-1/2}$, and $\Phi_2(a, b; \rho)$ is the cumulative distribution function of the bivariate standard normal distribution with correlation ρ (see Rosenbaum, 1961, Maddala, 1983, and Han and Lee, 2017). To estimate z_{it}, one can run the probit regression of a_{it} on a set of variables, which often contain $\mathbf{x}_i$ and other exogenous variables (all of which should be observed or imputed), at every t. The estimation of ρ_{vt} is more challenging. A practically useful method is to construct the likelihood function for (a_{it-1}, a_{it}) jointly as a function of ρ_{vt} for every $t = 2, \ldots, T$ after z_{it} and z_{it-1} are constructed as the fitted indices from individual probit regressions. Specifically, letting $p_{it}^{jk} = \Pr(a_{it-1} = j, a_{it} = k|z_{it}, z_{it-1}; \rho_{vt})$, we have $p_{it}^{11} = \Phi_2(z_{it-1}, z_{it}; \rho_{vt})$, $p_{it}^{10} = \Phi(z_{it-1}) - p_{it}^{11}$, $p_{it}^{01} = \Phi(z_{it}) - p_{it}^{11}$ and $p_{it}^{00} = 1 - p_{it}^{11} - p_{it}^{10} - p_{it}^{01}$. From this, the log-likelihood function

$$\ln L_t(\rho_{vt}) = \sum_{i=1}^{n}\Big[I(a_{it-1} = 1, a_{it} = 1)\ln p_{it}^{11} + I(a_{it-1} = 1, a_{it} = 0)\ln p_{it}^{10} + I(a_{it-1} = 0, a_{it} = 1)\ln p_{it}^{01} + I(a_{it-1} = 0, a_{it} = 0)\ln p_{it}^{00}\Big]$$

is constructed as a function of ρ_{vt} given z_{it} and z_{it-1}. For each t, ρ_{vt} is estimated by maximizing $\ln L_t(\rho_{vt})$. Note that $\ln L_t(\rho)$ is singular at $\rho = \pm 1$, and the inverse hyperbolic tangent (arctanh) transformation of ρ to $\frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$ is useful (Fisher, 1915, 1921). Alternatively, bivariate probit regression can be run for each pair of $t-1$ and t, but numerical optimization often fails.

Once all z_{it} and ρ_{vt} are estimated as explained above, the bias-correction terms $\psi(z_{it}, z_{it-1}; \rho_{vt})$ and $\psi(z_{it-1}, z_{it}; \rho_{vt})$ can be estimated in a straightforward fashion. Finally, one can estimate β by pooling the differenced equations augmented with the bias-correction terms. One method of pooling is to regress

$$\text{(5)} \qquad \begin{pmatrix} a_{i1}a_{i2}\Delta y_{i2} \\ a_{i2}a_{i3}\Delta y_{i3} \\ \cdots \\ a_{iT-1}a_{iT}\Delta y_{iT} \end{pmatrix} \text{ on } \begin{bmatrix} a_{i1}a_{i2}(\Delta\mathbf{x}_{i2} & \hat{\lambda}_{i,2} & 0 & \cdots & 0) \\ a_{i2}a_{i3}(\Delta\mathbf{x}_{i3} & -\hat{\psi}_{i,23} & \hat{\psi}_{i,32} & \cdots & 0) \\ \vdots & \vdots & \vdots & & \vdots \\ a_{iT-1}a_{iT}(\Delta\mathbf{x}_{iT} & 0 & 0 & \cdots & \hat{\psi}_{i,T,T-1}) \end{bmatrix}$$

using the pooled OLS (Han and Lee, 2017). Note the IMR $\hat{\lambda}_{i,2}$ in (5), which appears because $a_{i1} \equiv 1$. Importantly, pooling as in (5) requires that $u_{it} = \delta_t v_{it} + e_{it}$, where v_{is} and e_{it} are independent. This condition is not always satisfied. It is violated if, for example, $\Delta u_{it} = \delta_t v_{it} + e_{it}$, which is assumed by the IV bias correction of the FD regression. In order to encompass $\Delta u_{it} = \delta_t v_{it} + e_{it}$ as a special case, we change the right-hand side of (5) to

$$\begin{bmatrix} a_{i1}a_{i2}(\Delta \mathbf{x}_{i2} & \hat{\lambda}_{i,2} & 0 & 0 & \cdots & 0 & 0) \\ a_{i2}a_{i3}(\Delta \mathbf{x}_{i3} & 0 & \hat{\psi}_{i,23} & \hat{\psi}_{i,32} & \cdots & 0 & 0) \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{iT-1}a_{iT}(\Delta \mathbf{x}_{iT} & 0 & 0 & 0 & \cdots & \hat{\psi}_{i,T-1,T} & \hat{\psi}_{i,T,T-1}) \end{bmatrix}, \tag{6}$$

where the multiplication of $\hat{\psi}_{i,t-1,t}$ by -1 is not required because the -1 factor is handled by the regression.

(ii) *The pure attrition case:* Caution is needed for the pure attrition case, where the non-respondents never return. In contrast to the general missing case where a unit is observed whenever it selects to enter the sample, in the pure attrition case a_{it} is always zero if $a_{it-1} = 0$. To analyze this case, we assume that $a_{it} = a_{it-1} I(z_{it} + v_{it} > 0)$, where $v_{it} \sim N(0, 1)$ conditional on z_{it} and $a_{it-1} = 1$, so a probit model is specified conditional on the event that a unit is observed in the previous period, in contrast to the general missing case where the probit model is specified unconditionally.

When $\mathbf{x}_{it}$ is fully observed, bias correction is possible for pure attrition as follows. Consider, first, the pooled regression. For the model $y_{it} = \alpha + \mathbf{x}_{it}\beta + \varepsilon_{it}$, $a_{it} = a_{it-1} I(z_{it} + v_{it} > 0)$ and $\varepsilon_{it} = \delta_t v_{it} + e_{it}$, we have

$$\mathrm{E}(y_{it} | \mathbf{x}_i, a_{it} = 1) = \alpha + \mathbf{x}_{it}\beta + \delta_t \, \mathrm{E}(v_{it} | \mathbf{x}_i, a_{it} = 1),$$

where $\mathrm{E}(v_{it} | \mathbf{x}_i, a_{it} = 1) = \lambda(z_{it})$ as before. For the estimation of z_{it}, one can fit a probit model, not using all the observations but using only the units with $a_{it-1} = 1$. All the remaining procedure is identical to the general missing case. The final step is augmenting the equation with the IMR's and interaction with period dummies, pooling the data for $a_{it} = 1$ and then running the OLS regression. Extension to the CRE regression is naturally done by replacing $\mathbf{x}_{it}$ with $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$.

Bias correction of the FD estimation for the pure attrition is more challenging because the procedure involves the estimation of (pairwise) serial correlation in v_{it}. Obviously, estimation of a serial correlation coefficient requires a pair of variables, not just one. In the general missing

case, a_{it} and $\mathbf{x}_{it}$ are fully observed, where $\mathbf{x}_{it}$ is imputed if missing. In the pure attrition case, in contrast, $(1,1)$, $(1,0)$ and $(0,0)$ are observed for (a_{it-1}, a_{it}) but $(0,1)$ is never. In addition, if $a_{it-2} = 0$, then a_{it-1} and a_{it} are automatically zero, and thus $(0,0)$ delivers no useful information about the true possible (a_{it-1}, a_{it}) values in case $a_{it-2} = 0$. This is a nuisance problem, which we overcome by conditioning on the event that $a_{it-2} = 1$.

Conditional on $a_{it-2} = 1$, there are three possible sets of values for (a_{it-1}, a_{it}): $(1,1)$, $(1,0)$ and $(0,0)$; the excluded one $(0,1)$ is subsumed in $(0,0)$. Given the z_{it} and z_{it-1} values, a likelihood for ρ_{vt} can be constructed as follows:

$$\ln L_t(\rho_{vt}) = \sum_{i=1}^{n} a_{it-2}\Big[I(a_{it-1}=1, a_{it}=1)\ln p_{it}^{11} + I(a_{it-1}=1, a_{it}=0)\ln p_{it}^{10} + I(a_{it-1}=0)\ln(1-p_{it}^{11}-p_{it}^{10})\Big], \tag{7}$$

where a_{it-2} is multiplied in order to condition on the event that $a_{it-2} = 1$, and which can be maximized to estimate ρ_{vt}. Now, with z_{it}, z_{it-1} and ρ_{vt} estimated as explained so far, the correction terms $\psi(z_{it-1}, z_{it}; \rho_{vt})$ and $\psi(z_{it}, z_{it-1}; \rho_{vt})$ are obtained for the observed units ($a_{it} = 1$), and the pooled regression (5), or more preferably its modification to (6) can be conducted to estimate β. When $\mathbf{x}_{it}$ is incomplete as well as y_{it}, these procedures are make possible because $\mathbf{x}_{it}$ is imputed.

Summary on estimators

To summarize, we consider the following estimators for both the general missing case and the pure attrition case. (A1) The POLS estimator using the full unbalanced panel data is obtained by regressing y_{it} on $\mathbf{x}_{it}$ using the observed units. (A2) The CRE estimator using the full unbalanced panel data is obtained by regressing y_{it} on $\mathbf{x}_{it}$ and $\bar{\mathbf{x}}_i$ using the observed units, where $\bar{\mathbf{x}}_i$ is introduced in order to control for correlated individual effects. (A3) The FD estimator using the full unbalanced panel data is obtained by regressing Δy_{it} on $\Delta \mathbf{x}_{it}$ using all possible observations. (A4) The POLS estimator using the maximal balanced panel data is obtained by regressing y_{it} on $\mathbf{x}_{it}$ using the units observed for all t. (A5) The CRE estimator using the maximal balanced panel data is obtained by regressing y_{it} on $\mathbf{x}_{it}$ and $\bar{\mathbf{x}}_i$ using the units observed for all t. (A6) The FD estimator using the maximal balanced panel data is obtained by regressing Δy_{it} on $\Delta \mathbf{x}_{it}$ using the units observed for all t. These estimators do not correct the bias due to nonresponse.

For the general missing case, we consider the following BCI methods, together with the A1–A6 estimators. (B1) For the pooled regression with correction, the missing $\mathbf{x}_{it}$ are imputed, and

the bias-correction procedure is applied to the pooled equation. That is, the IMR's are obtained from the probit regression of a_{it} on $\mathbf{x}_{it}$ (imputed) using the observations with $a_{it-1} = 1$, and then the main pooled equation is augmented with the IMR's interacted with the period dummies. This estimation does not involve the estimation of the serial correlation coefficients in the selection equation error v_{it}. (B2) The B1 procedure is modified to CRE by substituting $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ for every occurrance of $\mathbf{x}_{it}$. (B3) The FD estimation with bias correction after imputation for the general missing case is obtained by using the 'ψ' correction terms in place of the IMR's and then pooling the data as in the regression of (5), preferably with modification to (6).

For the pure attrition case, A1–A6 are again considered, and B1–B3 are modified so that all the probit regressions are conducted conditional on $a_{it-1} = 1$, and the ρ_{vt} parameters in B3 are estimated conditional on $a_{it-2} = 1$ by maximizing (7). Let us name the resulting procedures B4, B5 and B6, respectively. The IV bias correction is also available in this case as follows. (C1) For the pooled regression, the IMR's are calculated for each t by the probit regression of a_{it} on $\mathbf{x}_{it-1}$ using the units with $a_{it-1} = 1$. The interaction terms of these IMR's and the period dummies are added to the right-hand side of the main equation and then a pooled IV regression is conducted using $\mathbf{x}_{it-1}$ and the interaction terms as instruments. (C2) For the FD regression, the bias-correction terms are obtained using the same method as in C1, but this time, they are added to the right-hand side of the FD equation rather than of the pooled levels equation.

3 Experiment Results and Discussions

In this section, we compare the alternative estimators introduced in Section 2 using Monte Carlo simulation. Comparison is made for both the general missing case and the pure attrition case. The main focus is put on the performance of the bias-correction methods after imputing $\mathbf{x}_{it}$. In what follows, the notation $e_{it} \sim AR_1(\rho)$ means that $e_{i1} \sim iid\ N(0,1)$ and $e_{it} = \rho e_{it-1} + (1 - \rho^2)^{1/2} e^0_{it}$, $t = 2, \ldots, T$, where $(e_{i1}, e^0_{i2}, \ldots, e^0_{iT}) \sim iid\ N(0, I_T)$ so that e_{it} is stationary AR(1) with serial correlation equal to ρ over t and iid across i.

Table 1: Average survival rate (general missing case, %)

t	$T=2$	$T=5$	$T=10$
1	100.00	100.00	100.00
2	79.28	79.23	79.32
3		79.25	79.33
4		79.25	79.32
5		79.31	79.32
6			79.34
7			79.34
8			79.35
9			79.36
10			79.34

Note. Data are generated by (8). Survival rate for each year is computed as a ratio to the first year. The average rates are obtained from 2,000 replications.

3.1 The General Missing Case

For the general missing case, data are generated as follows:

$$\text{(8a)} \qquad x_{it} = \xi_i^0 + x_{it}^0, \quad \xi_i^0 \sim N(0,1), \;\; x_{it}^0 \sim AR_1(0.75),$$

$$\text{(8b)} \qquad v_{it} = (\sigma_\eta^2 + 1)^{-1/2}(\sigma_\eta \eta_i + v_{it}^0), \quad \sigma_\eta = 1, \;\; \eta_i \sim N(0,1), \;\; v_{it}^0 \sim AR_1(0.65),$$

$$\text{(8c)} \qquad a_{it} = I(1 + 0.5x_{it} + v_{it} > 0), \;\; t \geq 2, \quad a_{i1} \equiv 1,$$

$$\text{(8d)} \qquad u_{it} = \delta_t v_{it} + e_{it}, \;\; t \geq 2, \quad u_{i1} \equiv e_{i1}^0, \;\; (e_{i1}^0, e_{i2}, \ldots, e_{iT}) \sim iid \; N(0, I_T),$$

$$\text{(8e)} \qquad y_{it} = \alpha_i + \beta x_{it} + u_{it}, \quad \beta = 1, \;\; \alpha_i \sim iid \; N(0,1) + \theta_\alpha \bar{x}_i, \;\; \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it},$$

where α_i, ξ_i^0, x_{it}^0, η_i, v_{it}^0, and $(e_{i1}^0, e_{i2}, \ldots, e_{iT})$ are mutually independent. Data (y_{it}, x_{it}) are observed if $a_{it} = 1$ in (8c), where v_{it} is serially dependent due to (8b). Missing is endogenous if $\delta_t \neq 0$ in (8d), where δ_t measures the degree of endogeneity. The α_i term in (8e) is random effects if $\theta_\alpha = 0$ and fixed effects if $\theta_\alpha \neq 0$.

The average survival rates are summarized in Table 1 when data are generated by (8). Every year, approximately 80% of the units in the initial sample are observed, and approximately 20% drop out of the sample.

Table 2 presents the simulated biases, the standard deviations, and the root mean squared errors of various estimators for $\theta_\alpha = 0$ (random effects) with $\delta_t = 0.25$ and $\delta_t = 0.75$, for $n = 1{,}000$ and for $T = 2, 5, 10$, obtained from 2,000 replications. The considered estimators

Table 2: The general missing case with random effects ($\theta_\alpha = 0$)

		$\delta_t = 0.25$			$\delta_t = 0.75$		
		Not corrected		Corrected	Not corrected		Corrected
T	Name	Full	Balanced	Imputation	Full	Balanced	Imputation
2	POLS	0.0131	0.0189	-0.0009	0.0407	0.0588	-0.0019
		(0.0284)	(0.0321)	(0.0316)	(0.0292)	(0.0331)	(0.0317)
		[0.0312]	[0.0373]	[0.0316]	[0.0501]	[0.0675]	[0.0318]
	CRE	0.0141	0.0141	0.0014	0.0390	0.0390	-0.0003
		(0.0719)	(0.0719)	(0.0743)	(0.0797)	(0.0797)	(0.0802)
		[0.0733]	[0.0733]	[0.0743]	[0.0887]	[0.0887]	[0.0802]
	FD	0.0141	0.0141	0.0043	0.0390	0.0390	0.0097
		(0.0719)	(0.0719)	(0.0723)	(0.0797)	(0.0797)	(0.0788)
		[0.0733]	[0.0733]	[0.0724]	[0.0887]	[0.0887]	[0.0794]
5	POLS	0.0267	0.0340	-0.0011	0.0820	0.1046	-0.0021
		(0.0252)	(0.0309)	(0.0317)	(0.0273)	(0.0335)	(0.0328)
		[0.0367]	[0.0460]	[0.0317]	[0.0865]	[0.1098]	[0.0329]
	CRE	0.0150	0.0100	-0.0100	0.0450	0.0290	-0.0307
		(0.0283)	(0.0318)	(0.0358)	(0.0304)	(0.0346)	(0.0379)
		[0.0320]	[0.0333]	[0.0371]	[0.0543]	[0.0451]	[0.0487]
	FD	0.0126	0.0102	0.0025	0.0349	0.0278	0.0062
		(0.0380)	(0.0412)	(0.0581)	(0.0398)	(0.0431)	(0.0613)
		[0.0400]	[0.0424]	[0.0581]	[0.0529]	[0.0513]	[0.0616]
10	POLS	0.0337	0.0403	0.0044	0.1017	0.1210	0.0122
		(0.0227)	(0.0306)	(0.0310)	(0.0246)	(0.0331)	(0.0338)
		[0.0407]	[0.0507]	[0.0313]	[0.1047]	[0.1254]	[0.0359]
	CRE	0.0190	0.0119	-0.0082	0.0567	0.0348	-0.0263
		(0.0160)	(0.0197)	(0.0280)	(0.0178)	(0.0224)	(0.0312)
		[0.0249]	[0.0230]	[0.0292]	[0.0594]	[0.0414]	[0.0408]
	FD	0.0114	0.0077	0.0031	0.0332	0.0212	0.0049
		(0.0261)	(0.0307)	(0.0509)	(0.0272)	(0.0324)	(0.0538)
		[0.0284]	[0.0317]	[0.0510]	[0.0429]	[0.0387]	[0.0540]

Note. Data are generated by (8) with $\theta_\alpha = 0$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

Table 3: The general missing case with fixed effects ($\theta_\alpha = 1$)

		$\delta_t = 0.25$			$\delta_t = 0.75$		
		Not corrected		Corrected	Not corrected		Corrected
T	Name	Full	Balanced	Imputation	Full	Balanced	Imputation
2	CRE	0.0141	0.0141	0.0059	0.0390	0.0390	0.0043
		(0.0719)	(0.0719)	(0.0743)	(0.0797)	(0.0797)	(0.0803)
		[0.0733]	[0.0733]	[0.0745]	[0.0887]	[0.0887]	[0.0804]
	FD	0.0141	0.0141	0.0043	0.0390	0.0390	0.0097
		(0.0719)	(0.0719)	(0.0723)	(0.0797)	(0.0797)	(0.0788)
		[0.0733]	[0.0733]	[0.0724]	[0.0887]	[0.0887]	[0.0794]
5	CRE	0.0150	0.0100	-0.0020	0.0450	0.0290	-0.0227
		(0.0283)	(0.0318)	(0.0359)	(0.0304)	(0.0346)	(0.0380)
		[0.0320]	[0.0333]	[0.0359]	[0.0543]	[0.0451]	[0.0442]
	FD	0.0126	0.0102	0.0025	0.0349	0.0278	0.0062
		(0.0380)	(0.0412)	(0.0581)	(0.0398)	(0.0431)	(0.0613)
		[0.0400]	[0.0424]	[0.0581]	[0.0529]	[0.0513]	[0.0616]
10	CRE	0.0190	0.0119	-0.0090	0.0567	0.0348	-0.0272
		(0.0160)	(0.0197)	(0.0282)	(0.0178)	(0.0224)	(0.0313)
		[0.0249]	[0.0230]	[0.0296]	[0.0594]	[0.0414]	[0.0414]
	FD	0.0114	0.0077	0.0031	0.0332	0.0212	0.0049
		(0.0261)	(0.0307)	(0.0509)	(0.0272)	(0.0324)	(0.0538)
		[0.0284]	[0.0317]	[0.0510]	[0.0429]	[0.0387]	[0.0540]

Note. Data are generated by (8) with $\theta_\alpha = 1$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

are A1–A3 (the POLS, CRE and FD estimators using the full sample), A4–A6 (those using the maximal balanced subset), and B1–B3 (bias correction after imputation). When $\theta_\alpha = 0$, the POLS, CRE and FD estimators are all consistent if there are no missing observations or if missing is not endogenous.

In Table 2, when the degree of endogeneity is small ($\delta_t = 0.25$), the conventional uncorrected estimators using the full sample ("Full") or using the maximal balanced subset ("Balanced") are only slightly biased, and BCI performs well. Under more severe endogeneity ($\delta_t = 0.75$), the uncorrected estimators are more seriously biased, and again the BCI estimators handle most of the biases except for CRE. It is notable that the CRE estimators (with or without bias correction) do not perform well except for $T = 2$. This poor performance is somewhat unexpected, and we have done vast extra simulations (unreported). It seems that the performance of the corrected estimators heavily depends on how well the missing values of x_{it} are imputed, naturally, and the inclusion of $\bar{\mathbf{x}}_i$ is especially harmful.

Table 3 reports the results when the individual effects are fixed effects ($\theta_\alpha = 1$). In the presence of fixed effects, the POLS estimator is inconsistent anyway, and we do not report the results for it. The CRE estimator is close to the within-group estimator (they are the same if the panel data is balanced), and the reported biases of the uncorrected CRE estimator are due to the endogenous missing. The performance of the imputation estimator for CRE is arguable but its bias is notable for larger T values. Combination of BCI and CRE is not recommended. BCI performs quite well for the FD estimator, on the other hand.

Simulation results for the general missing case suggest that BCI can be useful when applied to the FD estimation. Most of the bias of the uncorrected estimators is corrected by the BCI procedure. It is again naturally true that the performance of the corrected estimator depends on the accuracy of the data imputation according to the unreported vast simulations.

3.2 The Pure Attrition Case

The data generating process for the pure attrition case is modified from Section 3.1 in that $a_{it} = a_{it-1} I(1 + 0.5x_{it} + v_{it} > 0)$ instead of (8c). Everything else is identical to the corresponding

Table 4: Average survival rate (pure attrition case, %)

	$T=2$		$T=5$		$T=10$	
t	First	Last	First	Last	First	Last
1	100.00	100.00	100.00	100.00	100.00	100.00
2	79.28	79.28	79.23	79.23	79.32	79.32
3			72.78	91.85	72.89	91.90
4			67.97	93.40	68.09	93.41
5			64.05	94.22	64.13	94.19
6					60.76	94.74
7					57.80	95.14
8					55.20	95.49
9					52.85	95.75
10					50.73	95.99

Note. Data are generated by (9). "First" and "Last" indicate the survival rates (from 2,000 replications) in comparison to the period 1 and the previous period, respectively.

elements of (8). We reiterate the full data generating processes for future reference as follows:

$$\text{(9a)} \qquad x_{it} = \xi_i^0 + x_{it}^0, \quad \xi_i^0 \sim N(0,1), \;\; x_{it}^0 \sim AR_1(0.75),$$

$$\text{(9b)} \qquad v_{it} = (\sigma_\eta^2 + 1)^{-1/2}(\sigma_\eta \eta_i + v_{it}^0), \quad \sigma_\eta = 1, \;\; \eta_i \sim N(0,1), \;\; v_{it}^0 \sim AR_1(0.65),$$

$$\text{(9c)} \qquad a_{it} = a_{it-1} I(1 + 0.5 x_{it} + v_{it} > 0), \;\; t \geq 2, \quad a_{i1} \equiv 1,$$

$$\text{(9d)} \qquad u_{it} = \delta_t v_{it} + e_{it}, \;\; t \geq 2, \quad u_{i1} \equiv e_{i1}^0, \;\; (e_{i1}^0, e_{i2}, \ldots, e_{iT}) \sim iid \; N(0, I_T),$$

$$\text{(9e)} \qquad y_{it} = \alpha_i + \beta x_{it} + u_{it}, \quad \beta = 1, \;\; \alpha_i \sim iid \; N(0,1) + \theta_\alpha \bar{x}_i, \;\; \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it},$$

where α_i, ξ_i^0, x_{it}^0, η_i, v_{it}^0, and $(e_{i1}^0, e_{i2}, \ldots, e_{iT})$ are mutually independent.

When data are generated according to (9), Table 4 summarizes the simulated average survival rates. The dropout rate is approximately 80% in period 2 and less than 10% in very subsequent years in comparison to the previous year. After 10 waves, approximately a half of the initial units remain in the sample.

The estimators in this section include the conventional estimators using the full sample (A1–A3) and those using the maximal balanced subset (A4–A6), the bias-corrected estimators using data imputation (B4–B6), and two IV estimators C1–C2 (for the pooled regression and FD, respectively).

Tables 5 and 6 report the results for $\theta_\alpha = 0$ (random effects), for $\delta_t = 0.25$ and $\delta_t = 0.75$, respectively. The biases of the conventional estimators using the full sample ("Full") are similar

Table 5: The pure attrition case with random effects ($\theta_\alpha = 0$, $\delta_t = 0.25$)

T	Name	Not corrected		Corrected	
		Full	Balanced	IV	Imputation
2	POLS	0.0131	0.0189	-0.0010	-0.0009
		(0.0284)	(0.0321)	(0.0487)	(0.0316)
		[0.0312]	[0.0373]	[0.0487]	[0.0316]
	CRE	0.0141	0.0141	–	0.0014
		(0.0719)	(0.0719)		(0.0743)
		[0.0733]	[0.0733]		[0.0743]
	FD	0.0141	0.0141	0.0049	0.0043
		(0.0719)	(0.0719)	(0.3326)	(0.0723)
		[0.0733]	[0.0733]	[0.3326]	[0.0724]
5	POLS	0.0252	0.0340	0.0093	0.0011
		(0.0260)	(0.0309)	(0.0352)	(0.0314)
		[0.0362]	[0.0460]	[0.0364]	[0.0314]
	CRE	0.0112	0.0100	–	-0.0168
		(0.0305)	(0.0318)		(0.0384)
		[0.0325]	[0.0333]		[0.0419]
	FD	0.0118	0.0102	0.0068	0.0023
		(0.0390)	(0.0412)	(0.0977)	(0.0613)
		[0.0407]	[0.0424]	[0.0979]	[0.0613]
10	POLS	0.0316	0.0403	0.0080	0.0070
		(0.0247)	(0.0306)	(0.0315)	(0.0323)
		[0.0401]	[0.0507]	[0.0325]	[0.0330]
	CRE	0.0132	0.0119	–	-0.0174
		(0.0181)	(0.0197)		(0.0319)
		[0.0224]	[0.0230]		[0.0364]
	FD	0.0096	0.0077	0.0062	0.0043
		(0.0281)	(0.0307)	(0.0541)	(0.0512)
		[0.0297]	[0.0317]	[0.0545]	[0.0514]

Note. Data are generated by (9) with $\theta_\alpha = 0$ and $\delta_t = 0.25$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

Table 6: The pure attrition case with random effects ($\theta_\alpha = 0$, $\delta_t = 0.75$)

T	Name	Not corrected		Corrected	
		Full	Balanced	IV	Imputation
2	POLS	0.0407	0.0588	-0.0010	-0.0019
		(0.0292)	(0.0331)	(0.0487)	(0.0317)
		[0.0501]	[0.0675]	[0.0487]	[0.0318]
	CRE	0.0390	0.0390	–	-0.0003
		(0.0797)	(0.0797)		(0.0802)
		[0.0887]	[0.0887]		[0.0802]
	FD	0.0390	0.0390	-0.0004	0.0097
		(0.0797)	(0.0797)	(0.3672)	(0.0788)
		[0.0887]	[0.0887]	[0.3671]	[0.0794]
5	POLS	0.0779	0.1046	0.0310	0.0049
		(0.0283)	(0.0335)	(0.0389)	(0.0330)
		[0.0828]	[0.1098]	[0.0497]	[0.0334]
	CRE	0.0328	0.0290	–	-0.0521
		(0.0330)	(0.0346)		(0.0409)
		[0.0465]	[0.0451]		[0.0663]
	FD	0.0330	0.0278	0.0189	0.0060
		(0.0408)	(0.0431)	(0.1065)	(0.0649)
		[0.0525]	[0.0513]	[0.1081]	[0.0652]
10	POLS	0.0950	0.1210	0.0233	0.0196
		(0.0267)	(0.0331)	(0.0357)	(0.0357)
		[0.0987]	[0.1254]	[0.0426]	[0.0407]
	CRE	0.0391	0.0348	–	-0.0543
		(0.0206)	(0.0224)		(0.0356)
		[0.0442]	[0.0414]		[0.0649]
	FD	0.0277	0.0212	0.0154	0.0081
		(0.0295)	(0.0324)	(0.0587)	(0.0540)
		[0.0405]	[0.0387]	[0.0606]	[0.0546]

Note. Data are generated by (9) with $\theta_\alpha = 0$ and $\delta_t = 0.75$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

Table 7: The pure attrition case with fixed effects ($\theta_\alpha = 1$, $\delta_t = 0.25$)

T	Name	Not corrected		Corrected	
		Full	Balanced	IV	Imputation
2	CRE	0.0141	0.0141	–	0.0059
		(0.0719)	(0.0719)		(0.0743)
		[0.0733]	[0.0733]		[0.0745]
	FD	0.0141	0.0141	0.0049	0.0043
		(0.0719)	(0.0719)	(0.3326)	(0.0723)
		[0.0733]	[0.0733]	[0.3326]	[0.0724]
5	CRE	0.0112	0.0100	–	0.0078
		(0.0305)	(0.0318)		(0.0387)
		[0.0325]	[0.0333]		[0.0395]
	FD	0.0118	0.0102	0.0068	0.0023
		(0.0390)	(0.0412)	(0.0977)	(0.0613)
		[0.0407]	[0.0424]	[0.0979]	[0.0613]
10	CRE	0.0132	0.0119	–	0.0210
		(0.0181)	(0.0197)		(0.0325)
		[0.0224]	[0.0230]		[0.0387]
	FD	0.0096	0.0077	0.0062	0.0043
		(0.0281)	(0.0307)	(0.0541)	(0.0512)
		[0.0297]	[0.0317]	[0.0545]	[0.0514]

Note. Data are generated by (9) with $\theta_\alpha = 1$ and $\delta_t = 0.25$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

to the results in Table 2 (identical for $T = 2$). The results for the conventional estimators using the maximal balanced subset ("Balanced") are identical to the corresponding part in Table 2. With $\delta_t = 0.25$ in Table 5, bias due to attrition is limited, but the biases of the imputation estimator (BCI) are remarkably small (except for CRE). Table 6 presents the results for the case with a higher degree of endogeneity ($\delta_t = 0.75$), where the conventional uncorrected estimators suffer from larger biases. Notably the IV bias correction is invalid for $T > 2$ because the way u_{it} is generated in (9d) does not allow for the proper $\Delta u_{it} = \delta_t v_{it} + e_{it}$, which is crucial for the IV bias correction. As a result, the IV correction estimators are more or less biased. Among the estimators the BCI (imputation) estimators show least biases except for CRE. The preformance of BCI is especially remarkable for FD.

Tables 7 and 8 report the results with fixed effects ($\theta_\alpha = 1$). Again, due to the fixed effects, the POLS estimator is biased regardless of attrition, so we do not report the results for POLS.

Table 8: The pure attrition case with fixed effects ($\theta_\alpha = 1$, $\delta_t = 0.75$)

T	Name	Not corrected		Corrected	
		Full	Balanced	IV	Imputation
2	CRE	0.0390	0.0390	–	0.0043
		(0.0797)	(0.0797)		(0.0803)
		[0.0887]	[0.0887]		[0.0804]
	FD	0.0390	0.0390	-0.0004	0.0097
		(0.0797)	(0.0797)	(0.3672)	(0.0788)
		[0.0887]	[0.0887]	[0.3671]	[0.0794]
5	CRE	0.0328	0.0290	–	-0.0276
		(0.0330)	(0.0346)		(0.0409)
		[0.0465]	[0.0451]		[0.0494]
	FD	0.0330	0.0278	0.0189	0.0060
		(0.0408)	(0.0431)	(0.1065)	(0.0649)
		[0.0525]	[0.0513]	[0.1081]	[0.0652]
10	CRE	0.0391	0.0348	–	-0.0160
		(0.0206)	(0.0224)		(0.0353)
		[0.0442]	[0.0414]		[0.0388]
	FD	0.0277	0.0212	0.0154	0.0081
		(0.0295)	(0.0324)	(0.0587)	(0.0540)
		[0.0405]	[0.0387]	[0.0606]	[0.0546]

Note. Data are generated by (9) with $\theta_\alpha = 1$ and $\delta_t = 0.75$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

The CRE estimators do not work well regardless of bias correction but the bias-corrected FD estimators (with imputation) show remarkable performance in terms of bias correction. Again, the IV correction estimator is supposed to be biased due to the way Δu_{it} and v_{it} are related each other.

As explained in Section 2, the assumption that $\Delta u_{it} = \delta_t v_{it} + e_{it}$ is not satisfied when $u_{it} = \delta_t v_{it} + e_{it}$, and thus the IV correction estimators are inconsistent. In the rest of this section, we modify the data generating process so that the corrected IV estimators are consistent. Specifically, we conduct another set of experiments where u_{it} are generated by $\Delta u_{it} = \delta_t v_{it} + e_{it}$ instead of (9d). For future reference, again, we fully describe the considered data generating processes:

$$\text{(10a)} \quad x_{it} = \xi_i^0 + x_{it}^0, \quad \xi_i^0 \sim N(0,1), \;\; x_{it}^0 \sim AR_1(0.75),$$

$$\text{(10b)} \quad v_{it} = (\sigma_\eta^2 + 1)^{-1/2}(\sigma_\eta \eta_i + v_{it}^0), \quad \sigma_\eta = 0, \;\; \eta_i \sim N(0,1), \;\; v_{it}^0 \sim AR_1(0),$$

$$\text{(10c)} \quad a_{it} = a_{it-1} I(1 + 0.5 x_{it} + v_{it} > 0), \;\; t \geq 2, \quad a_{i1} \equiv 1,$$

$$\text{(10d)} \quad u_{it} = \delta_t v_{it} + e_{it} + u_{it-1}, \;\; t \geq 2, \quad u_{i1} \equiv e_{i1}^0, \;\; (e_{i1}^0, e_{i2}, \ldots, e_{iT}) \sim iid \; N(0, I_T),$$

$$\text{(10e)} \quad y_{it} = \alpha_i + \beta x_{it} + u_{it}, \quad \beta = 1, \;\; \alpha_i \sim iid \; N(0,1) + \theta_\alpha \bar{x}_i, \;\; \bar{x}_i = \frac{1}{T} \sum_{t=1}^{T} x_{it},$$

where α_i, ξ_i^0, x_{it}^0, η_i, v_{it}^0, and $(e_{i1}^0, e_{i2}, \ldots, e_{iT})$ are mutually independent. Above, (10d) is modified from (9d), and (10b) is also revised so that v_{it} is serially independent. For data generated by (10), the bias-corrected IV estimators are consistent, so is the BCI estimator when the differenced equations are pooled by the more flexible (6) instead of (5).

Table 9 presents the results for data generated by (10) with $\theta_\alpha = 1$ (fixed effects). Both the bias-corrected IV estimator and the BCI estimator show little bias (except for CRE), and the BCI estimator is substantially more efficient than the IV estimator. As for the FD estimators, Wooldridge's IV estimation works well.

4 Conclusion

This study considers the method of correcting bias after imputing missing exogenous variables (BCI) for panel data models with missing observations and attrition, together with conventional estimators without correction, and the instrumental-variable (IV) bias-correction estimators. When no variables are observed for the non-respondents, Monte Carlo experments suggest that

Table 9: The pure attrition case with fixed effects under the new assumption ($\theta_\alpha = 1, \delta_t = 0.75$)

T	Name	Not corrected		Corrected	
		Full	Balanced	IV	Imputation
2	CRE	0.0380 (0.0613) [0.0721]	0.0380 (0.0613) [0.0721]	–	0.0032 (0.0602) [0.0603]
	FD	0.0380 (0.0613) [0.0721]	0.0380 (0.0613) [0.0721]	0.0008 (0.2866) [0.2865]	0.0090 (0.0601) [0.0608]
5	CRE	0.0423 (0.0492) [0.0649]	0.0258 (0.0576) [0.0631]	–	-0.0875 (0.0553) [0.1035]
	FD	0.0555 (0.0337) [0.0650]	0.0432 (0.0386) [0.0579]	-0.0020 (0.1508) [0.1508]	0.0052 (0.0516) [0.0518]
10	CRE	0.0628 (0.0549) [0.0834]	0.0197 (0.0783) [0.0807]	–	-0.1638 (0.0726) [0.1792]
	FD	0.0603 (0.0263) [0.0658]	0.0403 (0.0358) [0.0539]	0.0022 (0.1088) [0.1088]	-0.0010 (0.0467) [0.0467]

Note. Data are generated by (10) with $\theta_\alpha = 1$ and $\delta_t = 0.75$. The simulated bais, simulated standard deviation (in parentheses), and root-mean-squared errors (in square brackets) are reported.

bias-correction after imputation can be useful when combined with the FD regression. Bias-correction of the CRE estimator does not show desirable properties, on the other hand. When both the IV bias-correction and the imputation bias-correction show tiny biases, the imputation method seems utilize information in a considerably more efficient way.

The results reported in this paper are based on Monte Carlo experiments, and results may vary across data and the accuracy of imputation. But the results look stable according to vast (unreported) simulations we have conducted. More experments and possibly analytical studies are called for.

References

Baltagi, B. H. (2014). Econometric Analysis of Panel Data, 5th edition, Wiley.

Han, C., and G. Lee (2017). Efficient estimation of linear panel data models with sample selection and fixed effects, Working paper, Korea University.

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.

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.

Little, R. J. A., and D. B. Rubin (2002). Statistical Analysis with Missing Data, 2nd edition, Wiley.

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

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.

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.