Simple Tests of Random Missing for Unbalanced Panel Data Models

Do Won Kwak† and Suyong Song‡

Abstract

This paper proposes simple tests of the validity of the assumption on missing process including missing completely at random (MCAR) and missing at random (MAR) assumptions for unbalanced panel data by extending the Hausman [1978] specification test. The proposed Hausman-Type (HT) tests can be applied flexibly to estimations with complete-case, multiple imputations, sample selection correction, and inverse probability weighted methods. Monte Carlo simulations show substantial power and no size bias in the tests for the size of panel data we typically observe in applications. We illustrate the usefulness of the tests of the assumptions on missing/selection process, using a study of the effects of trade liberalization policies on trade flows. The HT test results show that conventional estimators substantially underestimate the effects of these policies on trade flows.

Keywords: Unbalanced panel data; Hausman-type specification test; Missing data; Missing completely at random; Missing at random

JEL Classification: C12, C33, F15

*We would like to thank Jeffrey M. Wooldridge for his perspective and helpful comments. We are grateful to Xuepeng Liu for kindly providing his dataset and Timothy Vogelsang, Todd Elder, Chris O’Donnell, Adrian Pagan, Alicia Rambaldi, Prasada Rao, Donggyu Sul for suggestions and comments. We also thank seminar participants at Korea University, Michigan State University, Monash University, University of Sydney, and University of Queensland, for useful comments. All remaining errors are our own.

†School of Economics, The University of Queensland, St Lucia, QLD, Australia, d.kwak@uq.edu.au
‡Department of Economics, University of Iowa, W360 Pappajohn Business Building, 21 E. Market Street, Iowa City, IA 52242, suyong-song@uiowa.edu
1 Introduction

As panel data are becoming more readily available, economic and statistical analysis using the methods of balanced data is becoming more attractive to empirical researchers. However, balanced panel data are rarely available because of missing data. For instance, Abrevaya and Donald [2010] found in their surveys that missing data occur in 40% of all published articles in four popular empirical economics journals (AER, JHR, JOLE, and QJE) from 2006-2008. If the missing completely at random (MCAR) assumption is satisfied (Rubin [1976] and Little and Rubin [2002]), we can still justify the results from the balanced data methods but, unfortunately, the MCAR assumption is not generally testable. In 70% of these publications with missing data, units with incomplete observations are all discarded based on the MCAR assumption without providing any justification for the validity of that assumption.\(^1\) Other remaining studies typically rely on the assumption that the missing and non-missing data have the same distribution conditional on observed covariates. The literature generally assumes the MAR assumption which is used with the models for imputation, inverse probability weight, and missing process and focuses on efficient estimation (Robins et al. [1994], Hahn [1998], Cattaneo [2010], Abrevaya and Donald [2010], Muris [2011], Chaudhuri [2014], Chaudhuri and Guilkey [2014], among others). Again, justification of the results for the estimates from using the MAR assumption relies on faith on imputation and inverse probability weight models rather than formal statistical analysis because these MAR assumptions have been regarded as untestable in most empirical applications.\(^2\)

For instance, Hirsch and Schumacher [2004], Bollinger and Hirsch [2006] and Bollinger and Hirsch [2013] examine the quality of imputations for the studies that use Current Population Survey data, but the MAR assumption used for imputation with the data can sometimes exacerbate bias which arises from missing data. On the other hand, Cheng and Trivedi [2015] examine the effect of attrition for the study that use Medicine in Australia: Balancing Employment and Life and find that the effect from missing data is not significant.

Although the MCAR and MAR assumptions are generally considered untestable (Horowitz and Manski [2006]), this paper identify that, for certain unbalanced linear panel data models which have at least two consistent estimators under the null hypothesis, the validity of the assumptions on missing process or MAR assumption can be tested. For instance, it is not uncommon to assume such that both fixed

---

\(^1\) Missing completely at random (MCAR) and missing at random (MAR) assumptions used in this paper are introduced by Rubin [1976] and we use in the following sense. The MCAR assumption implies that the probability of missing is independent of observed and unobserved variables. The MAR assumption implies that, conditional on observed variables and fixed effects that we are accounting for, the probability of missing does not depend on the remaining unobserved variables.

\(^2\) There is also previous studies that look at when the MAR assumption is not satisfied so they model both outcome and missing process (See Hausman and Wise [1979] and Wooldridge [2010b]). This approach also relies on correct model specification of missing process which is generally not testable.
effects and first-differencing estimators are consistent especially if main cause of bias is unobserved heterogeneity (e.g. individual fixed effects in two-dimensional panel data) under the null hypothesis in practice. In examining the validity of the assumptions in unbalanced panel data, we propose test statistics that extend the Hausman [1978] specification test. Construction of test statistics is based on the idea that, under the null hypothesis of the MCAR assumption (or the MAR assumption as well as missing process assumption), the difference in the estimates of two consistent estimators should be entirely due to sampling errors. We perform a test for the MAR assumption or modelling assumption on missing process\(^3\), using a two-stage estimation with two consistent estimators under the null hypothesis. In implementation outcome model is estimated in the second stage and imputation (or inverse probability weighting) is performed at the first stage based on the model for missing process. This is useful in practice because researchers often believe that missing occurs non-randomly (i.e. missing process depends on unobserved variables) and have information to model a missing process so they use this information to eliminate or mitigate biases from non-random missing data. However, testing for the validity of the MAR assumption (or other assumption for missing process) is rarely performed in practice although it is very simple to implement. We extend the Hausman [1978] specification test to be implemented with not just raw data analysis but also imputation based as well as inverse probability weighting (IPW) method.\(^4\) The idea of the tests for the validity of the assumption for missing process is based on the difference between two scaled, consistent MI estimators (or consistent IPW estimators). The null hypothesis implies the validity of the MAR assumption (or model for missing process) and indicates the consistency of MI (or IPW) estimators. For two imputed estimators, an assumption for the model of missing process is used to calculate missing observations at the first stage. Likewise, with two IPW estimators, a MAR assumption is used to calculate the probability of observability at the first stage. Using these two estimators from the first stage and the consistency of these two estimators under the null hypothesis, our proposed test can test the validity of the MAR assumption for the missing process.

In testing the MCAR and MAR assumptions, we construct test statistics from the difference between two scaled, consistent estimators that asymptotically follow \(\chi^2\)-distribution. There are a few notable characteristics of the proposed tests. First, no restriction is imposed on missing patterns for the justi-

---

\(^3\)This is also called as assignment mechanism in the treatment effect literature.

\(^4\)The MI method with the MAR assumption is routinely applied and the details of the MI method are well documented in Shafer [1997], Buuren et al. [1999], Schafer and Graham [2002], Little and Rubin [2002], Hahn [1998], Chen et al. [2008], and Imbens et al. [2007]. The IPW method with MAR was initially introduced by Horvitz and Thompson [1952] but has not been used much until recently because of inefficiency. This, however, has changed drastically since Robins et al. [1994] introduced an efficient version of the IPW (the Augmented IPW or AIPW) method. The IPW method with the MAR assumption has recently been studied by Tsiatis [2006], Hirano et al. [2003], Wooldridge [2007], and Graham et al. [2012].
fication of these test statistics. For instance, missing data can occur in the dependent and explanatory variables and can have any types of missing (e.g., attrition type and/or intermediate type). Second, the tests allow the researcher to focus upon the inconsistency of a single covariate of interest (e.g., “critical core covariate” in Lu and White [2014] or White and Lu [2011]) caused by non-random missing data. Focusing only upon a single/core covariate is important from a practical point of view because devising the MAR assumption such that all included covariates are conditionally random, is extremely difficult and probably unrealistic, while the MAR assumption on one variable of interest is relatively easier to work with. Third, the proposed tests of the MCAR and MAR assumptions can be easily implemented with standard statistical software such as Stata. 5 Furthermore, the tests can be generalized to using any two scaled, consistent estimators under the null hypothesis. The power of the test depends upon an alternative hypothesis that leads to at least one of estimators being inconsistent.

Finite-sample properties of the proposed tests are examined with Monte Carlo (MC) experiments and the MC results show no size distortion of the tests and substantial power for the size of panel data we typically observe in applications. For sensitivity analyses, various types of alternative hypotheses that induce bias for pooled least square (LS), fixed effects (FE) and first-differencing (FD) estimators are examined in the MC experiments.

A use of the proposed tests is illustrated using the study of the effects of trade liberalization policies – World Trade Organization (WTO) membership, Regional Trade Agreement (RTA) membership, Currency Union (CU) membership – on trade flows. A gravity equation in unbalanced panel data is often estimated with error-component models as in Wansbeek and Kapteyn [1989], Baltagi and Chang [1994], Werner [2001], and Davis [2002]. Justification of the results from these methods requires the validity of the MCAR or MAR assumptions (or the model for export market entry decision), but we find no formal testing to justify the validity of the assumption. In our sample, the dependent variable (trade flows) suffers severely from missing data, with a missing proportion of about 50%. If we can assume the other sources of biases in the gravity model are properly controlled, we can attribute the source of bias to missing data using our proposed test. A rejection of null hypothesis implies violation of the MCAR assumption (or the MAR assumption) and the rejection is induced by a statistically significant difference between the FE and FD estimates (e.g., MI-FE and MI-FD or IPW-FE and IPW-FD estimates) beyond sampling error. Furthermore, as an illustration following previous studies that used the MAR assump-

5'Stata ado’ codes to implement the HT test of the MCAR assumption and the MAR assumption with IPW and MI estimators can be downloaded at http://dwkwak.weebly.com/research.html. 'Stata do’ files with illustrative examples and 'Stata help' files for the codes are also provided at the website.
tion with imputation in the gravity model as in Linders and de Groot [2006] and Felbermayr and Kohler [2006] and two-step estimations with Heckman correction, we show the usefulness of the HT test of the MAR assumption by imputing missing observations for trade flows as well as obtaining the inverse mills ratio (IMR). We also conduct the HT tests of the MAR assumption with the IPW estimators where inverse probability weights is obtained by modeling a missing/observability process with observed variables, including exclusion restrictions.

Our illustration of the HT tests provides several interesting implications in the estimation of the effects of trade liberalization policies. First, the HT test rejects the null hypothesis. Thus, assuming that other sources of biases are properly controlled, both the differences between conventional FE and FD estimates and those between corresponding FE and FD estimates with imputations, are beyond sampling error. Or if the modelling assumption for missing process is correct, either FE and FD estimator are biased even with no missing data. Furthermore, test results imply that the effects of trade liberalization policies, such as WTO and Regional Trade Agreement (RTA hereafter), were severely underestimated by ignoring missing data, assuming that non-random missing is main source of bias. The trade creation effect of WTO membership is estimated to be 0% without imputation, but 55% with imputation; while RTA membership increases trade flows by 13% without imputation, and 105% with imputation. Therefore, the WTO and RTA membership effects are underestimated by 55 and 92 percentage points respectively. Moreover, the HT tests of the MAR assumption with imputation also imply that the standard and most popular methods of controlling for omitted variable bias are not justifiable, provided the MAR assumption used for the imputation model is valid. Assuming that our imputation model of the MAR assumption from economic argument is reasonable, we can attribute the difference between the FE and FD estimates with imputation, to omitted variable bias that is not completely removed by conventional methods of controlling omitted variables, as suggested in Magee [2003], Baldwin and Taglioni [2006], Baier and Bergstrand [2007, 2009]. Finally, two-step and inverse probability weights (IPW) estimates of the FE and FD estimators are only slightly different to conventional estimates that ignore missing data; and the qualitative conclusions remain the same with and without using IPW and self-selection corrections. This implies that the MAR assumption used to obtain IPW and self-selection correction terms, do little practically to mitigate/eliminate bias caused from non-random missing. This is not surprising in our application, because we are limited in the availability of exclusion restrictions that predict the

---

6We implement one of the most popular methods used in the gravity model that accounts for three-way error components using fixed effects as in Anderson and van Wincoop [2003], Feenstra [2004], Baldwin and Taglioni [2006], Baier and Bergstrand [2007], Magee [2008], Eicher et al. [2012], Matyas and Balazsi [2012] among others.
The rest of paper is organized as follows: in Section 2, we introduce the HT test statistics to test the MCAR assumption in an unbalanced panel data; in Section 3 we extend the HT test statistics to test the MAR assumption; in Section 4 we show the size and the power of the test statistics through Monte Carlo experiments; in Section 5 we provide application of HT tests to the study of the effect of trade liberalization on trade flows; and in Section 6 we make our concluding remarks. The appendix contains technical details, additional Monte Carlo experiments, and a “recipe” for our tests routine in Stata.

2 Linear panel data models with missing data

We consider the following linear panel data model with missing data where missing process is expressed with observability indicator which is denoted by observability indicator $s_{it}$:

$$ s_{it}y_{it} = s_{it}(d_{it}\beta + x_{1it}\delta + f_t + c_i + u_{it}), \text{ for } i = 1, 2, ..., N \text{ and } t = 1, 2, ..., T $$ (1)

where $y_{it}$ is an outcome, $d_{it}$ is a vector of regressors of primary interest, $x_{1it}$ is a vector of remaining observed covariates, $f_t$ is time effect, $c_i$ is unobserved heterogeneity and $u_{it}$ is idiosyncratic error. Using $x_{it} = (d_{it}, x_{1it}, f_t)$, $\theta = (\beta', \delta', 1)'$ and $v_{it} = c_i + u_{it}$, we rewrite (1) with selection indicator more compactly as follows

$$ s_{it}y_{it} = s_{it}x_{it} \cdot \theta + s_{it}v_{it}, \text{ for } i = 1, 2, ..., n \text{ and } t = 1, 2, ..., T. $$ (2)

2.1 HT test with missing data

Following Wooldridge [2010b], we state the assumptions in terms of conditional mean and missing indicators. We can rewrite equation (2) for all $T$ time periods as

$$ s_i y_i = s_i X_i \cdot \theta + s_i v_i $$

Specifically, we use religion and lagged dependent variables as exclusion restrictions, following Helpman et al. [2008] in the first stage estimation of the probability of observability. These excluded variables do not have enough variation to correctly predict the probability of observability and to have a major effect on the estimates for variables of trade liberalization policies.
where \( s_i \) is \( T \times T \) matrix and its \( t \)-th diagonal element is \( s_{it} \) and off-diagonal element is zero; \( y_i = (y_{i1}, y_{i2}, ..., y_{iT})' \); \( X_i = (x_{i1}', x_{i2}', ..., x_{iT})' \); and \( v_i = (v_{i1}, v_{i2}, ..., v_{iT})' \).

First, like the original test in Hausman [1978] and the test with unbalanced data in Verbeek and Nijman [1992], we consider two consistent estimators under the null hypothesis. However, unlike Verbeek and Nijman [1992] that used balanced sub-sample that throw away some observations, we only use unbalanced data in the construction of two estimates. Under Assumptions RE.1-RE.3, the original form of the Hausman test (HT) statistic can be constructed using scaled difference between RE and FE estimators where variance estimators for both RE and FE are used for scales.

**Assumption** RE.1: (a) \( \mathbf{E}(u_{it}|X_i,c_i,s_{it} = 1) = 0, t = 1,2,...,T \); (b) \( \mathbf{E}(c_i|X_i,s_{it} = 1) = 0 \) where \( X_i = (x_{i1},x_{i2},...,x_{iT}) \).

**Assumption** RE.2: \( \text{rank} \mathbf{E}(s_iX_i'\Omega^{-1}X_i) = dim(\theta) \) where \( \Omega = \mathbf{E}(s_iv_i'v_i) \).

**Assumption** RE.3: (a) \( \mathbf{E}(s_iu_iu_i'|X_i,c_i) = \sigma^2 s_iI_T \); (b) \( \mathbf{E}(c_i|X_i) = \sigma^2 \).

Compared to the assumptions in the original HT, everything remains the same except that the assumptions are applied to the sub-sample with \( s_{it} = 1 \). This is a simple replication of one case of Verbeek and Nijman [1992]'s test where \( D_4 = \begin{bmatrix} 0_k & I_k \\ -I_k & 0_k \end{bmatrix} \). As we only use unbalanced data, compared to Verbeek and Nijman [1992] which has \( 4k \) number of parameters, we have only \( 2k \) number of parameters.

We denote \( D_4 \) in Verbeek and Nijman [1992] as \( R = \begin{bmatrix} I_k & -I_k \end{bmatrix} \).

**Theorem 1** Under the Assumptions RE.1-RE.3, a Wald statistic \( W \) converges to \( \chi^2_p \) distribution under the null hypothesis of \( H_0: R \cdot \theta = 0 \), where

\[
W = |R\sqrt{N} \cdot \hat{\theta}|' [R\sqrt{\text{var}(\sqrt{N} \cdot \hat{\theta})}R']^{-1} R\sqrt{N} \cdot \hat{\theta} \sim \chi^2_p
\]

with

\[
\hat{\theta} = \left[ \begin{array}{c} I_k \\ -I_k \end{array} \right]_{k \times 2k} \left[ \begin{array}{c} \hat{\theta}_{FD}' \\ \hat{\theta}_{FE}' \end{array} \right]_{2k \times 1}
\]

As we can view a Wald statistic \( W \) as a special case of the HT test, we can use the same argument to the proof of Theorem 2.1 in Hausman [1978] except that we replace the assumptions and data generating process for \( \{y_{it},x_{it}\} \) with for \( \{s_{it},y_{it},s_{it}|x_{it}\} \) which is subset of complete sample.

For Assumption RE.1 to hold true, \( \mathbf{E}(v_{it}|X_i) = 0 \) should be satisfied \( \forall t \) and non-random missing implies \( \mathbf{E}(v_{it}|X_i) = \mathbf{E}(v_{it}|X_i,s_{it} = 1) \). Thus, under RE. 1, both the RE and FE estimators are consistent. Assuming that violation of the MCAR assumption (i.e. non-random missing data), \( \mathbf{E}(v_{it}|X_i) \neq 0 \)
\(E(v_{it}|X_i, s_{it} = 1)\), is the only source of inconsistency in the estimation, our proposed test has power against the alternative hypothesis of \(E(v_{it}|X_i) \neq E(v_{it}|X_i, s_{it} = 1)\).

### 2.2 HT test with missing data under heteroskedasticity and serial correlation

We can use other panel data estimators and allow heteroskedasticity and serial correlation, following Wooldridge [2010b]. Robust Wald statistic can be constructed using scaled difference of two consistent estimators under the null.\(^8\) Under strict exogeneity and rank conditions which are corresponding to Assumptions RE.1 and RE. 2, any two consistent estimators under the null can be used to construct robust Hausman-type Wald statistics.

#### 2.2.1 Robust Hausman Test with non-random missing

Now consider the following outcome and observability equations.

\[
y_{it} = d_{it}\beta + x_{1it}\delta + f_t + c_i + u_{it} \tag{3}
\]

where \(x_{it} = (d_{it}, x_{1it}, f_t)\) and \(v_{it} = c_i + u_{it}\).

\[s_{it} = 1(w_{it}y + \epsilon_{it} \geq 0)\]

The bias for \(\beta\) and \(\delta\) due to non-random missing is caused by the correlation between \(v_{it} = (f_t + c_i + u_{it})\) and \(\epsilon_{it}\) (i.e. \(cov(v_{it}, \epsilon_{it}) \neq 0\)).\(^9\) However, note that if the source of bias is \(cov(c_i, \epsilon_{it}) \neq 0\) and \(cov(u_{it}, \epsilon_{it}|w_{it}) = 0\), then this implies both FE and FD estimators are consistent. Thus, the HT test that use FE and FD estimators has the power against the alternative hypothesis that \(cov(u_{it}, \epsilon_{it}|w_{it}) \neq 0\). This is because, against this alternative, both FE and FD estimators are inconsistent due to non-random missing, so the power of tests rely on the different magnitude of the inconsistency of these two estimators.

If \(w_{it}\) is a subset of \(d_{it}\) and \(x_{1it}\), then the MAR assumption (i.e. selection on observables) is satisfied so pooled OLS, FE and FD estimators are all consistent for \(\beta\) and \(\delta\). In this case, we can ignore missing data and obtain consistent estimates for \(\beta\) and \(\delta\). On the other hand, if \(w_{it}\) is subset of \(d_{it}, x_{1it}, f_t, c_i\), then both FE and FD are consistent, so the MAR assumption (i.e. selection is observables and fixed effects) is

---

\(^8\)We call our test statistic allowing heteroskedasticity and serial correlation as robust HT test, in the sense that, unlike in the original Hausman test we do not make any restriction on the second moment of error term to obtain efficiency. We instead estimate a sandwich form of error covariance matrix. See the Appendix for the detail.

\(^9\)This implies \(cov(v_{it}, \epsilon_{it}|w_{it}) \neq 0\) where \(w_{it}\) include all observed variables at time \(t\).
satisfied and both FE and FD for unbalanced data are consistent. However, if \( w_{it} \) is bigger than the space spanned by \( d_{it}, x_{it}, f_i, c_i \), then the MAR assumption is violated (i.e. \( \text{cov}(v_{it}, \epsilon_{it}|d_{it}, x_{it}, f_i, c_i) \neq 0 \)) because selection process can not be accounted for all available observed variables, so both FE, and FD are inconsistent for \( \beta \) and \( \delta \). Thus, in this case, the HT test that uses FE and FD estimators has power against \( \text{cov}(v_{it}, \epsilon_{it}|d_{it}, x_{it}, f_i, c_i) \neq 0 \). If the HT test rejects the null, then we reject \( \text{cov}(v_{it}, \epsilon_{it}|d_{it}, x_{it}, f_i, c_i) = 0 \).

Thus, another HT test statistic that uses the difference between \( \hat{\theta}_{FD} \) and \( \hat{\theta}_{FE} \) could be used to test the null hypothesis of the MAR assumption (i.e. missing is at random conditional on observed variables and fixed effects). We need to adjust strict exogeneity assumption, rank condition, and regularity conditions for \( \sqrt{N} \)-asymptotically normal to invoke CLT.

**Assumption 2.3A**

1. **Strict exogeneity condition**: \( \mathbb{E}(u_{it}|X_i, c_i) = 0 \) and the MAR assumption \( \mathbb{E}(u_{it}\epsilon_{it}|w_{it}, c_i, s_{it} = 1) = 0 \) \( \forall t \).

2. **Proper rank conditions**: \( E(\sum_{t=1}^{T} s_{it}^F \Delta x'_{it} \Delta x_{it}|X_i, c_i, s_{it} = 1) \) and \( E(\sum_{t=2}^{T} s_{it}^F \Delta x'_{it} \Delta x_{it}|X_i, c_i, s_{it} = 1) \) have full rank.

3. **Regularity conditions**: conditions to guarantee positive semi-definiteness of \( R\hat{\vartheta}(\sqrt{N}\hat{\theta}_2)R' \) and \( p\lim(\hat{\vartheta}(\sqrt{N}\hat{\theta}_2)) = \text{var}(\sqrt{N}\hat{\theta}_2) \) to invoke CLT to \( \frac{1}{\sqrt{N}} \sum_{i=1}^{T} \sum_{t=2}^{T} s_{it}^F \Delta x'_{it} \Delta u_{it} \) and \( \frac{1}{\sqrt{N}} \sum_{i=1}^{T} \sum_{t=1}^{T} s_{it}^F \Delta x'_{it} \tilde{u}_{it} \) are satisfied.

We use the following stacked estimator \( \hat{\theta}_2 = (\hat{\theta}_{FD}', \hat{\theta}_{FE})' \) that uses difference between the FD and FE estimators:

\[
\sqrt{N}\hat{\theta}_2 = \left[ \begin{array}{cc}
\frac{1}{N} \sum_{i=1}^{T} s_{it}^F \Delta x'_{it} \Delta x_{it} & 0 \\
0 & \frac{1}{N} \sum_{i=1}^{T} s_{it}^F \Delta x'_{it} \tilde{x}_{it}
\end{array} \right]^{-1} \left[ \begin{array}{cc}
\frac{1}{N} \sum_{i=1}^{T} s_{it}^F \Delta x'_{it} \Delta y_{it} \\
\frac{1}{N} \sum_{i=1}^{T} s_{it}^F \Delta x'_{it} \tilde{y}_{it}
\end{array} \right]_{2k \times 2k}.
\]

We write the null hypothesis as below and the FD and FE estimators are consistent under the null hypothesis of the MAR assumption.

\[
H_0 : R\theta_2 = 0,
\]

\[
R\theta_2 = \left[ \begin{array}{c}
I_k \\
-I_k
\end{array} \right]_{k \times 2k} \left[ \begin{array}{c}
\theta'_{FD} \\
\theta'_{FE}
\end{array} \right]_{2k \times 1}.
\]
Proposition 2  Under Assumption 2.3A, a Wald statistic $W_2$ converges to $\chi^2_p$ distribution under the null hypothesis of $H_0: R\theta_2 = 0$, where

$$W_2 = [R\sqrt{N(\hat{\theta}_2)}]'[R\sqrt{\text{var}(\hat{\theta}_2)}]^{-1}R\sqrt{N(\hat{\theta}_2)} \sim \chi^2_p. \quad (7)$$

The rejection of the null hypothesis with the HT test statistic, $W_2$, implies that the FE and FD estimators are inconsistent, so conditional strict exogeneity (i.e. the MAR assumption) is violated if strict exogeneity condition hold (i.e. no endogeneity problem for balanced data).

2.2.2 Test of one parameter

Sometimes researchers are interested in the causal relationship of a small set of core parameters or often only one parameter. The practical importance of focusing only on a small number of core parameters is also examined in Lu and White [2014] and this is the case for our empirical application in this paper. One parameter HT test statistic is important because, in general, it provides a higher power and allows non-random missing process to be correlated with other included independent variables. We denote the one parameter version of test statistics that correspond to the multiple parameter version test statistics $W_q$ by $T_q$.

A HT statistic $T_q$ using the null hypothesis $H_0: R_j\theta_q = 0$ where $R_j$ is $j$th row vector of $k \times 2k$ matrix $R$ is given as follows:

$$T_q = \frac{R_j\sqrt{n}(\hat{\theta}_q)}{\sqrt{R_j\text{var}(\sqrt{n}\hat{\theta}_q)R_j'}} \sim t_{n-1} \sim z. \quad (8)$$

2.3 Variable addition tests

Simple tests of violation of the MCAR assumption for unbalanced panel data, called variable addition tests, have been introduced in the literature (Verbeek and Nijman [1996] and more recently in Wooldridge [2010a]). The idea of the tests is that under the MCAR assumption, any function of $s_{it} \cdot x_{it}$ should not appear in the expression for $E(s_{it} \cdot y_{it})$. So a test on the coefficients for the lag and lead values of $\{s_{it}, s_{it} \cdot x_{it}\}$ has been proposed accordingly. Rejection implies the violation of the MCAR assumption. For instance, a simple variable addition test for the MCAR assumption is a robust $t$-test for the coefficient on $s_{it+1}$ in a pooled linear regression model as follows:
\[ y_{it} = \gamma_1 \cdot s_{it+1} + x_{it}' \cdot \theta + u_{it}, \text{ for } s_{it} = 1. \]  \hspace{1cm} (9)

We can similarly implement variable addition tests for the FE and FD estimators as follows:

\[ \tilde{y}_{it} = \gamma_2 \cdot \tilde{s}_{it+1} + \tilde{x}_{it}' \cdot \tilde{\theta} + \tilde{u}_{it}, \text{ for } s_{it} = 1 \]  \hspace{1cm} (10)

where \( \tilde{y}_{it} = s_{it} \cdot (y_{it} - \Sigma_{k=1}^{T} s_{ik}y_{ik}) \) and \( \tilde{x}_{it}' \) and \( \tilde{u}_{it} \) are similarly defined.

We examine the relative performance of HT tests and variable addition tests by comparing the power and size bias of the tests under various alternative hypotheses using Monte Carlo simulation. We consider four data generating processes in Section 4. Although the variable addition test is quite useful sometimes, it is not easy to extend to the test of the MAR assumption, so in the test of the MAR assumption, we focus on the HT test.

### 3 HT test using Two-Step, IPW, and MI estimators

In the previous section, we have not consider estimating observability equation:

\[ s_{it} = 1(w_{it} \gamma + \epsilon_{it} \geq 0). \]  \hspace{1cm} (11)

There are numerous empirical studies that try to estimate the selection equation (11) to avoid bias from non-random missing data. The estimate of selection equation is typically used to correct for sample selection or generate probability of observability from two-step estimations. Also, there are other studies that use imputation to fill in missing observations.

We extend the proposed HT tests using the multiple imputation (MI), inverse probability weighting (IPW) and two-step estimators. \( \epsilon_{it} \) is correlated with \( u_{it} \) even after conditioning on all observed variables as well as fixed effects, \( d_{it}, x_{it}, f_t, c_i \), so the MAR assumption is violated. Under the null of correct model for selection process \( s_{it} \), estimators using IPW and selection correction can provide two consistent estimators. Alternative hypothesis is that correction model is incorrect. For instance, even after accounting for \( w_{it} \) which include \( d_{it}, x_{it}, f_t, c_i \), and exclusion restriction (i.e. additional linearly independent variables), \( \epsilon_{it} \) is correlated with \( u_{it} \).
3.1 Two-step estimator

We extend the HT test to be applicable with sample-selection (i.e. observability) correction. With the assumption on selection process (i.e. selection on observables, \( w_{it} \)) as in (11), researcher apply probit estimation to get inverse mills ratio (IMR hereafter). Under a joint normality assumption for error term in outcome, \( u_{it} \), and error term in selection model, \( \epsilon_{it} \), the IMR term can correct bias due to non-random missing. Consider the following outcome and selection model:

**Assumption 3.1** A random vector \( w_{it} \) is all available observed variables including exclusion restriction and a selection model is provided as follows:

\[
s_{it} = 1(w_{it} \gamma + \epsilon_{it} \geq 0) \quad \text{and} \quad E(u_{it}|\epsilon_{it}) = \rho \epsilon_{it}.
\]

Then, we obtain:

\[
E(y_{it}|w_{it}, s_{it} = 1) = x_{it} \theta + c_i + \rho E(\epsilon_{it}|w_{it}, \epsilon_{it} > -w_{it} \gamma) = x_{it} \theta + c_i + \rho \frac{\phi(w_{it}\gamma)}{\Phi(w_{it}\gamma)}
\]

(12)

where \( x_{it} = (d_{it}, x_{1it}, f_i) \), \( \phi(\cdot) \) and \( \Phi(\cdot) \) are the standard normal PDF and CDF respectively. Using the Probit model in (11), we can obtain \( \frac{\phi(w_{it}\gamma)}{\Phi(w_{it}\gamma)} \) (IMR) from the first-stage estimation.

For \( s_{it} = 1 \), we apply the second-stage estimation with FE and FD estimators:

\[
y_{it} = x_{it} \theta + c_i + \rho \frac{\phi(w_{it}\gamma)}{\Phi(w_{it}\gamma)} + \eta_{it}
\]

(13)

If missing is not random, we have \( \text{cov}(x_{it}, u_{it}|s_{it} = 1, x_{it}, c_i, \epsilon_{it}) \neq 0 \). But, if selection model is correctly specified, \( \text{cov}(x_{it}, \eta_{it}|s_{it} = 1, x_{it}, c_i, \epsilon_{it}) = 0 \), so both FE and FD estimations in the second stage applied to (13) are consistent.\(^{10}\)

Assuming that non-random missing is the sole source of inconsistency, the HT test statistic comparing these two-stage estimators are consistent under the null that selection model is correctly specified.

**Proposition 3** Under Assumptions 2.3A and 3.1, a test statistic \( W_{TS-1} \) converges to \( \chi^2_p \) distribution under

\(^{10}\)As IMR term in the second stage is a generated regressor, we have to adjust standard error in the second stage estimation in both FE and FD estimations. A possible adjustment is to use bootstrap standard error in the construction of HT test statistic.
the null hypothesis of \( H_0: \ R\theta_{TS-1} = 0 \), where

\[
W_{TS-1} = [R\sqrt{N(\hat{\theta}_{TS-1})}]'[R\sqrt{\text{var}(\hat{\theta}_{TS-1})}R']^{-1} R\sqrt{N(\hat{\theta}_{TS-1})} \sim \chi^2_p.
\]

3.1.1 Alternative hypothesis

From two-step estimation, we can see the difference of HT tests between the ones in Verbeek and Nijman [1992] and our proposed ones. We define hypothetical sample with no missing data as complete sample. The HT test of Verbeek and Nijman [1992] has no power, if unbalanced sample with missing data is not a representative of the complete sample. In other words, missing is non-random in the sense that missing is correlated with the outcome even after conditioning on all observed variables and fixed effects. This is because unbalanced panel that researchers have could be already not a representative of the complete sample, so transforming unbalanced data to balaced one by cutting off may not recover the original representativeness of the complete sample.

The HT test need to have power against the alternative hypothesis that we concern about. That is, missing is correlated with outcome differently according to the value of covariate of interest. For instance, for units with high value of covariate, if high proportion of missing occurs for unit with higher value of outcome, then we call this as differential missing and this is an example of non-random missing.

Consider the following example where an intervention that induces class size reduction (from 25 students in a regular classroom to 15 students in a small classroom) was introduced in public schools:

\[
Score_{it} = \text{small}_{it} \cdot \theta + u_{it}
\]

We are interested in obtaining consistent estimate of \( \theta \) (i.e. impact of small classroom on exam score). Suppose parents prefer small classrooms to regular classrooms so unsatisfied parents who have children in a regular class try to move their children to private schools. Problem is that only some of rich family can afford to move their children to private schools and students in rich family probably perform better in exam on average due to better resource input. Therefore, those outperforming students in regular classrooms are more likely to be missing so only underperforming students remain in regular classrooms. Therefore, with unbalanced data, \( \theta \) would be overestimated. Including correction term, \( \frac{\phi(\frac{w_{it}x}{\sigma_c})}{\phi(\frac{w_{it}}{\sigma_c})} \), makes to compare outcome difference only those in small and regular rooms with the same value of \( \frac{\phi(\frac{w_{it}x}{\sigma_c})}{\phi(\frac{w_{it}}{\sigma_c})} \). As a result, this adjustment is to provide these remaining outperformers in regular classrooms to get more...
weights in the estimation.

Therefore, in practice, the power of the HT test relies on the quality of the first stage estimation that produces IMR term. If information from exclusion restriction has any additional explanatory power for predicting missing process (or probability of missing), the IMR terms can mitigate or remove the source of bias from non-zero \( \text{cov}(x_{it}, u_{it}|s_{it} = 1, x_{it}, c_i) \). As we can hold \( \frac{\phi(w_{it}^T)}{\Phi(w_{it}^T)} \) fixed, we can compare the outcome difference coming from the difference in covariates for those who have the same probability of observing the unit. This removes the selection bias as long as the selection model is correctly specified and joint normal assumption on error terms is satisfied.

### 3.2 IPW estimators

We extend the HT test to the IPW estimators where the selection model is estimated to be used as the probability of observability \( (s_{it}) \). In addition to Assumption 2.3 (or 2.3A), we impose Assumption 3.2 for the probability of observability model following Wooldridge [2007].

**Assumption 3.2**

1. A random vector \( z_{it} \) is available for a correct specification of observability: 
   \[
   E(s_{it}|d_{it}, x_{1it}, z_{it}) = P(s_{it} = 1|d_{it}, x_{1it}, z_{it}) = f(z_{it}) \text{ and } f(z_{it}) > 0 \forall z_{it} \in Z \text{ where } z_{it} \text{ contains both } d_{it}, x_{1it} \text{ as well as fixed effects in the outcome model.}
   \]

2. A random vector of predictors \( z_{it} \) is always observed.

3. A parametric model \( G(z_{it}, \gamma) \) for \( f(z_{it}) \) is available so \( E(s_{it}|z_{it}) \) can be consistently estimated.

Note that \( G(z_{it}, \gamma) \) is the probability of observability. Then we get:

\[
\frac{s_{it}y_{it}}{G(z_{it}, \gamma)} = \frac{s_{it}x_{it} \cdot \theta}{G(z_{it}, \gamma)} + \frac{s_{it}v_{it}}{G(z_{it}, \gamma)} \text{ for } i = 1, 2, ..., N \text{ and } t = 1, 2, ..., T. \tag{14}
\]

\( G(z_{it}, \hat{\gamma}) \) can be estimated consistently at the first stage by using parametric models such as a Probit (or Logit) model with

\[
s_{it} = 1(z_{it} \cdot \gamma + e_{it} > 0) \tag{15}
\]
where $e_{it}$ follows either a normal or logistic distribution.\textsuperscript{11} Once we obtain $G(z_{it}, \hat{\gamma})$ consistently from the 1st stage, we can estimate (14) in the second stage using among pooled OLS, FE, and FD estimators. The IPW estimators are consistent as long as $E(s_{it}|z_{it})$ is consistently estimated, as summarized in the following theorem.\textsuperscript{12}

Following Wooldridge [2007], we can relax the assumption of complete observability of the predictors to partial observability of the predictors.

**Assumption 3.2A** A random vector $z_{it}$ is observed whenever $s_{it} = 1$.

The results we use for the proof of consistency of the IPW-LS and IPW-FD estimators remain the same except the first-stage estimation of $G(z_{it}, \hat{\gamma})$. We instead use the following equation to estimate $f(z_{it})$ because we do not have observation $z_{it}$ for unit $i$ with $s_{it} = 0$:

$$p(s_{it} = 1|z_{it}) = p(s_{it} = 1|s_{it-1} = 1, z_{it-1}) \cdot p(s_{it-1} = 1|s_{it-2} = 1, z_{it-2}) \cdots p(s_{i1} = 1) \equiv G_{it}$$

where we assume $p(s_{i1} = 1) = 1$ (i.e. random sample is assumed at the beginning of period). We estimate $\pi_{it} \equiv p(s_{it} = 1|s_{it-1} = 1, z_{it-1})$ by using parametric models such as a Probit model:

$$\pi_{it} = \Phi(c_p + z_{it-1} \cdot \gamma > 0), \text{ for } s_{it-1} = 1$$

where $z_{it-1}$ is observed and $\Phi(\cdot)$ is the standard normal CDF.

**Proposition 4** Under Assumptions 2.3A and 3.2, a test statistic $W_{IPW-1}$ converges to $\chi^2_p$ distribution under the null hypothesis of $H_0: R\theta_{IPW-1} = 0$, where

$$W_{IPW} = [R \sqrt{N(\hat{\theta}_{IPW1})}]'[R \sqrt{a_{\text{IPW}1}}(\sqrt{N(\hat{\theta}_{IPW1})})R']^{-1} R \sqrt{N(\hat{\theta}_{IPW1})} \sim \chi^2_p.$$  

where $\hat{\theta}_{IPW} = (\hat{\theta}_{IPW-FD}, \hat{\theta}_{IPW-FE})'$.

### 3.2.1 Alternative hypothesis

If researchers believe the MAR assumption is violated, they could try to estimate $E(s_{it}|d_{it}, z_{it})$. They could use $G(z_{it}, \gamma)$ as a probability weight to correct selection bias due to missing data. Note that $G(z_{it}, \gamma)$

\textsuperscript{11}As gamma is the estimated parameter in the first stage that is also used in the second stage estimator, there is so-called generated regressor problem. Thus, the standard error need to be corrected by using the bootstrap method.

\textsuperscript{12}Consistency can be shown with the law of iterated expectation. See Appendix B for the proof. The consistency for the IPW-FE and IPW-FD estimators under the assumption of correct specification for $E(s_{it}|z_{it})$ can be similarly obtained under Assumptions 2.3 and 3.2.
is estimated rather at unit level than at variable level, so the IPW may be more relevant to missing types like attrition. Consider an example in 3.1.1 again. The idea behind how inverse probability weighting corrects the bias is following. We infer that rich people in regular classrooms are more likely to be missing in the sample so there remain only very few rich people in regular classrooms. For instance, suppose there are 10 rich people in regular classroom in complete sample but 8 people left the sample, so observability probability for the rich in regular classroom is 0.2. IPW gives remaining 2 people inverse probability weight of 5 (=1/0.2). So two people play the role of 10 sample observations. If these two people represent missing 8 people well, inverse-probability-weighting could recover original random sample feature although the missing process disturbs random sampling. Therefore, the IPW method heavily rely on the correct specification of $G(z_{it}, \gamma)$. As long as non-random missing is the only source of bias, the null is that the model for $G(z_{it}, \gamma)$ is correct and both FE and FD are consistent under the null. Thus, the HT test has power against incorrect specification of $G(z_{it}, \gamma)$. The rejection of the HT test implies that either the model for selection is incorrect or bias from other sources is significant if we believe $G(z_{it}, \gamma)$ is correctly specified. Therefore, we have to suspect about the model specification on both outcome and selection if we reject the null and have a large value of the HT test statistic. In most empirical case, we can use our proposed test to examine model specifications on both outcome (e.g. omitted variable bias) and selection (e.g. non-random missing).

### 3.3 MI estimators

The HT test can be applied using the MI estimators. Imputation could be more relevant where missing occurs at variable level than at unit level. For each missing variables, researchers can define imputation models. For the sake of simplicity, we consider a case where missing occurs to dependent variable only. In addition to Assumptions 2.3 (or 2.3A), we assume a missing process of dependent variable – a correct model specification for the distribution and predictors of missing observations for the dependent variable conditional on the availability of predictors.

**Assumption 3.1** Conditional distribution of $y_{m,it}$ given $z_{it}$ and $c_i$ which include all observed variables and individual specific-fixed effects, is correctly specified as

$$D(y_{m,it}|z_{it}, c_i), \quad (18)$$

where $y_{m,it}$ denotes a missing dependent variable for individual $i$ at time $t$ and $z_{it}$ and $c_i$ (i.e. fixed
effects) are valid predictors for missing observation.

Under the validity of the assumption (i.e. the model for missing observations, \( y_{m,it} \), conditional on predictors, \( z_{it} \), is correct), we obtain a consistent estimator for \( \theta \). Our proposed HT test of the MAR assumption is constructed by using the difference between two consistent estimators under the null hypothesis that the model for imputation in the equation (18) is correctly specified.

The MI estimators are obtained from three steps. First, the missing values, \( y_{m,it} \), are filled with predicted values of \( y_{m,it} \) from (18) \( M \) times so we obtain \( M \) complete data. Second, standard balanced panel data methods are applied to the \( M \) complete data sets. Therefore, we can obtain \( M \) number of the MI-FE and MI-FD estimates using the original and imputed observations. For instance, for the MI-FE estimator, we get coefficient estimates on \( d_{it}, \hat{\beta}_{j,fe} \), and its covariance matrix \( \hat{V}_{j} \) where \( j = 1, 2, ..., M \).

Finally, the results from the \( M \) analyses are combined into a single analysis by applying the Rubin’s formula for the MI-FE estimator as follows:

\[
\hat{\theta}_{MI-1} = \frac{1}{M} \sum_{j=1}^{M} \hat{\theta}_{j,fe}; \quad \hat{V}_{MI-1} = \frac{1}{M} \sum_{j=1}^{M} \hat{V}_{j,fe} + \frac{M+1}{M} B_{fe}
\]  

(19)

where \( B_{fe} = \frac{1}{M-1} \sum_{j=1}^{M} (\hat{\beta}_{j,fe} - \bar{\hat{\beta}}_{fe})(\hat{\beta}_{j,fe} - \bar{\hat{\beta}}_{fe})' \). \( \hat{\beta}_{MI-1} \) and \( \hat{V}_{MI-1} \) can be similarly obtained.

We can extend the formula in (19) to a HT test statistic, \( W_{MI-1} \). We stack two \( k \)-dimensional estimators \( \hat{\beta}_{j,MI-1} \) and \( \hat{\beta}_{j,MI-1} \) and define this as \( \hat{\theta}_j = (\hat{\beta}'_{j,MI-1}, \hat{\beta}'_{j,MI-1})' \) and its variance estimate as \( \bar{\text{var}}(\hat{\theta}_j) \).

Using \( \hat{\theta}_j \) and \( \bar{\text{var}}(\hat{\theta}_j) \) and applying (19) we obtain

\[
\hat{\theta}_{MI-1} = \frac{1}{M} \sum_{j=1}^{M} \hat{\theta}_j; \quad \bar{\text{var}}(\hat{\theta}_{MI-1}) = \frac{1}{M} \sum_{j=1}^{M} \bar{\text{var}}(\hat{\theta}_j) + \frac{M+1}{M} B,
\]

(20)

where \( B = \frac{1}{M-1} \sum_{j=1}^{M} (\hat{\theta}_j - \hat{\theta}_{MI})(\hat{\theta}_j - \hat{\theta}_{MI})' \). Then \( W_{MI-1} \) is obtained as follows:

\[
W_{MI-1} = [R\sqrt{N}(\hat{\theta}_{MI-1})]'[RN\bar{\text{var}}(\hat{\theta})_{MI-1}R']^{-1}R\sqrt{N}(\hat{\theta}_{MI-1}).
\]

Under conditional strict exogeneity, rank condition, regularity conditions, and the correct model specification for observability, it could be shown that \( W_{MI-1} \) converges to \( \chi^2_p \) distribution under the null hypothesis of \( R\theta_{MI-1} = 0 \). The result is summarized in the following proposition.

**Proposition 5** Under Assumptions 2.3 and 3.1, a test statistic \( W_{MI-1} \) converges to \( \chi^2_p \) distribution under
the null hypothesis of $H_0 : R\theta_{MI-1} = 0$, where

$$W_{MI-1} = [R\sqrt{N(\hat{\theta}_{MI-1})}]'[RN\sqrt{\text{var}(\hat{\theta}_{MI-1})}R']^{-1}R\sqrt{N(\hat{\theta}_{MI-1})} \sim \chi^2_p.$$  

### 3.3.1 Alternative hypothesis

In the MI case, under the null, the model for $D(y_{m,it}|z_{it}, c_i)$ is correctly specified. As long as non-random missing is the only source of bias, both FE and FD obtained from multiple imputations are consistent under the null. Thus, the HT test has power against incorrect specification of $D(y_{m,it}|z_{it}, c_i)$. The rejection of the HT test implies that either imputation model is incorrect or bias from other sources is significant if we believe $D(y_{m,it}|z_{it}, c_i)$ to be correctly specified.

### 4 Monte Carlo Experiment

Finite-sample performances of the HT tests for the MCAR and MAR assumptions with the IPW and two-step estimators are provided. The results show no size distortion and high power of the HT tests in various data generating processes (DGPs).

#### 4.1 Tests of the MCAR assumption

We consider the following linear panel data model where the response variable is continuous and explanatory variables can be either binary or continuous:

$$y_{it} = x_{it} \cdot \beta + c_{1i} + u_{it}, \text{ for } i = 1, 2, ..., n \text{ and } t = 1, 2, ..., T$$  

where $x_{it}$ includes binary (treatment dummy) variables and continuous variables. We are interested in performing statistical inference on $\beta$.

Selection processes reduce the sample size ranging from 40 to 70 percent of full samples. We consider observations of cross-sectional dimension from $n=100$ to $n=1,000$ and of time-dimension from $T=4$ to $T=8$. Note that

$$s_{i1} = 1; \quad s_{it} = 1(c_p + z_{it} \gamma + c_{2i} + v_{it} > 0)$$  

where $c_p$ is chosen to make the missing fraction be between 40 and 70 percent. The violation of the MCAR assumption is induced by the correlation between $u_{it}$ (or $c_i$) and $x_{it}$ conditional on observability.
(s_{it} = 1).

### 4.1.1 Binary explanatory variable \( d_{it} \) and non-random missing

We consider following two data generating processes (DGPs) as a benchmark where we are interested in estimating \( \beta \) in equation (21). The following DGPs are chosen to mimic our empirical example in a simple way and to satisfy ideal conditions for desirability of estimators except endogeneous missing data.\(^{13}\)

**DGP I**

1. \( c_{1i} = i.i.d.N(0, \frac{1}{16}) + w_{1i}; c_{2i} = q_{i} + w_{2i}; q_{i} \sim i.i.d.N(0, 1). \)

2. \((w_{1i}, w_{2i}) \sim BN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)\) where BN is bivariate normal.

3. \( d_{it}^* = q_{i} + i.i.d.N(0, \frac{1}{16}); x_{1it} = 1 \text{ if } d_{it}^* > 0; \beta_1 = 1. \)

4. \( x_{2it} \sim i.i.d.N(0, 1); \beta_2 = -0.2. \)

5. \( z_{it} \sim i.i.d.N(0, 1); \gamma = 1. \)

6. \( u_{it} \sim i.i.d.N(0, 1); v_{it} \sim i.i.d.N(0, 1). \)

First, in the DGP I, the violation of the MCAR assumption is induced by the correlation between the time-invariant unobserved factor \((w_{1})\) and \(x_{1it}\) conditional on observability. The correlation is indirect as it appears between error in selection model and \(x_{1it}\), and between error in selection model and time-invariant unobserved factor in \(c_{1i}\). Therefore, although \(x_{1it}\) and \(c_{1i}\) are not directly correlated, they are correlated conditional on \(s_{it} = 1\). The pooled LS estimator is inconsistent but the FE and FD estimators are consistent conditional on observability because the unobserved factor in \(c_{1i}\) causes inconsistency on the coefficient of \(x_{1it}\). In DGP I, the null hypothesis of the MCAR assumption is satisfied only if \(\rho = 0\). In addition, ideal conditions for the RE model are satisfied under the null hypothesis. Thus, we can replace the LS estimator with the RE estimator in the construction of the HT test. However, in the DGP I under alternative of \(\rho \neq 0\), the LS and RE estimators are inconsistent while the FE and FD estimators are consistent. Thus, we perform tests of the MCAR assumption using the HT statistics of \(W_{1}\).

\(^{13}\)To save space, we only report the HT tests for the MCAR assumption based on the coefficient of a binary covariate, \(d_{it}\), while we report the HT tests for the MAR assumption in estimating continuous covariate. We provide DGPs for the HT tests for the MCAR assumption based on a continuous covariate in Appendix B. The qualitative implications from simulations remain exactly the same.
(based on difference between the LS and FE estimators), $W_2$ (based on difference between the FD and FE estimators), $W_3$ (based on difference between the LS and FD estimators).

**DGP II**

1. $c_{1t} = \text{i.i.d.}\ N(0, 1)$; $c_{2t} = \text{i.i.d.}\ N(0, 1)$.
2. $z_{it} \sim \text{i.i.d.}\ N(0, \frac{1}{4})$, $\gamma = 1$.
3. $d_{it}^* = z_{it} \alpha + \text{i.i.d.}\ N(0, \frac{1}{16})$; $x_1_{1t} = 1$ if $d_{it}^* > 0$; $\beta_1 = 1$.
4. $x_{it} \sim \text{i.i.d.}\ N(0, 1)$; $\beta_2 = -0.2$.
5. $(u_{it}, v_{it}) \sim BN \begin{pmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \end{pmatrix}$.

Second, in the DGP II, the violation of the MCAR assumption is induced by the correlation between time-varying unobserved factor $(u_{it})$ and $x_1_{1t}$ conditional on observability. In this case, all three estimators – the pooled LS, FE, and FD estimators – are inconsistent. The null hypothesis is satisfied if $\rho = 0$. Under the null, ideal conditions for these estimators are satisfied. All three estimators are inconsistent, under the alternative of $\rho \neq 0$, but the magnitude of inconsistency differs among estimators.

**4.1.2 Simulation results**

MC simulations are performed with 2,000 replications for $T=4$ and $T=8$ and $n$ ranging from 100 to 1,000 and cluster robust standard errors are used in all simulations. The magnitude of correlation because of non-random missing and its effect on the inconsistency of $\beta_1$ in (21) are determined by $\rho$. When $\rho = 0$, no differential missing across $x_1_{1t}$ occurs, so no inconsistency is caused by non-random missing. Finite-sample performance of the HT tests when covariate of interest is binary is examined in Tables 1 and 2 for DGP I and Tables 3 and 4 for DGP II.

The true value of $\beta_1$ is 1 and the rejection rate of the null hypothesis should be nominal size, 0.05, under the true null hypothesis. When $\rho$ is zero, no inconsistency on $\beta_1$ is caused by non-random missing. Five HT tests and three variable addition tests using four panel data estimators such as the LS, RE, FE, and FD estimators are examined. The mean of the MC point estimates for $\beta_1$ is close to true value 1 for all four estimators under $\rho = 0$ regardless of missing proportion. In the DGP I with $\rho = 0$, the null hypothesis is true.$^{14}$ Table 1 shows that, for DGP I with missing proportion of 40%, coverages of rejection

---

$^{14}$In this paper, all simulations of tests are based on 95% confidence interval.
probability contain 0.05 for all five tests. As missing proportion increases from 40% to 60%, slight over-
rejections of null hypothesis appear for some of the HT tests when \( N = 100 \) and \( N = 500 \), but coverages
of rejection probability of all five HT tests contain 0.05 when \( N = 1,000 \). Proposed five HT tests show
no size bias. Finite-sample behavior of the HT tests improves as \( N \) increases. The last three columns
report rejection probability for variable addition tests using the pooled LS, FE and FD estimators. Over-
rejection appears for the pooled LS estimator and size bias is larger when missing proportion is higher.
Coverages of rejection probability contain 0.05 in cases for the FE and FD estimators regardless of miss-
ing proportion. Variable addition tests with the FE and FD estimators show no size bias while the pooled
LS estimator shows slight over-rejection.

Under DGP I where inconsistency on an estimator of \( \beta_1 \) is caused by conditional correlation between
covariate \( x_{1it} \) and unobserved time-invariant factors, the pooled LS is inconsistent when \( \rho \) is not equal
to zero. In simulation, we choose \( \rho = 0.9 \) to induce inconsistency of 50% for a missing proportion of 40% and inconsistency of 100% for a missing proportion of 60%.\(^{15} \) No inconsistency arises for the FE and FD estimators because FE and FD transformations eliminate unobserved time-invariant factors that cause inconsistency because of non-random missing data. In Table 1, the last six rows report results when \( \rho \) is 0.9. The sample means of the MC point estimates for \( \beta \) are about 0.47 and 0.07 for the pooled LS estimator with missing proportions of 40% and 60%, respectively. Finite-sample mean estimates for \( \beta_1 \) do not depend on cross-sectional dimension, \( n \), for the pooled LS estimator. On the other hand, the
mean values of the MC point estimates of \( \beta_1 \) for the FE and FD estimators are 1 for all \( n \) regardless of the
missing proportion. Therefore, the Pooled LS estimator is inconsistent while the FE and FD estimators
are consistent. Among three HT tests, two HT tests compare a consistent estimator with an inconsistent
estimator, while one HT test compares two consistent estimators. As \( n \) increases over 500, the mean of
rejection probability becomes 1 for all four HT tests that compare a consistent estimator with an incon-
sistent estimator. In an empirical application in this paper, observations in cross-sectional dimension
are 46,440. Given 50% inconsistency (i.e. mean estimate is \( \hat{\beta}_1 = 0.5 \) when true \( \alpha = 1 \)), simulation results
indicate that the powers of all four tests are 1 as long as \( n \geq 500 \). The last three columns in Table 1 re-
port variable addition tests with \( \rho = 0.9 \) for the pooled LS, FE and FD estimators. The power is 1 for the
pooled LS estimator as long as \( n \geq 500 \). For variable addition tests of the FE and FD estimators, the null
is true because the source of non-random missing is removed by FE and FD transformations. The last
two columns show that coverages of rejection probability contain 0.05 for the FE and FD estimators.

\(^{15}\)Inconsistency increases for the Pooled LS as \( \rho \) increases.
Table 2 reports the results for DGP I with increased time dimension $T = 8$. We obtain the same qualitative results as in $T = 4$. As there is no over-time complication such as serial correlation introduced in DGP I, not many differences appear in the MC results with increased time dimension.

Table 3 reports the MC results where $T = 4$ and data are generated by the DGP II for which non-random missing is caused by the correlation between covariate, $x_{1it}$, and unobserved time-varying factors conditional on observability if $\rho \neq 0$. The first six rows report results with $\rho = 0$ and the last six rows report results with $\rho = 0.9$. For $\rho = 0$, the means of the MC estimates for $\beta_1$ are the same as true value, 1, for all four estimators, so all four estimators are consistent. The null hypothesis of the MCAR assumption is true so the size of tests should be 0.05. All five HT test statistics are based on two consistent estimators. Table 3 shows that the means of rejection probability are close to 0.05 and coverages of rejection probability of all five HT tests contain 0.05. It is true for all $n$ when missing proportion is 40% and for $n \geq 1,000$ when missing proportion is 60%. Table 3 also reports results for variable addition tests with the Pooled LS, FE and FD estimators. The means of rejection probability are close to 0.05 (ranging from 0.38 to 0.75) although coverages sometimes do not contain 0.05.

The last six rows in Table 3 report the results with $\rho = 0.9$ so all four estimators are inconsistent. In contrast to the cases with DGP I, the FE and FD transformations cannot remove the source of inconsistency because it is time-varying. The means of the MC estimates for $\beta_1$ for the pooled LS, FE and FD estimators are 0.57, 0.66, and 0.89, respectively. Induced finite-sample inconsistency is 43%, 34% and 11% for the pooled LS, FE and FD estimators, respectively. All three HT test statistics are based on two inconsistent estimators. The magnitudes of inconsistency differ among the pooled LS, FE and FD estimators, and the HT test statistics using difference in two estimates can have substantial power for the null of the MCAR assumption. With DGP II and non-zero $\rho$, three HT tests – that compare between the pooled LS and FD and between the FE and FD – show high power of tests. This is because the FD estimate is quite different from other estimates. In particular, these three HT tests show their powers are 1 when $n$ is 1,000. The power of all HT tests increases with sample size. Three variable addition tests show relatively low power when $n$ is 100 and when the pooled LS estimator is used for all $n$. Table 4 provides results from the DGP II with increased time dimension $T = 8$. Qualitative implications remain the same regardless of the size of the time dimension.

In sum, correlation between a covariate and unobserved factor conditional on observability leads to discrepancies among the pooled LS, FE, and FD estimates. The proposed HT tests use these discrepancies to determine whether they are beyond sampling errors. The simulation results show that (i) the
power of the HT tests is substantial, (ii) the power of HT tests increases with \( n \), and (iii) there is no size bias for HT tests when no correlation exists that causes non-random missing, especially with \( n \geq 500 \).

The MC simulation results where covariate is continuous are provided in Appendix B and they show that the HT tests with continuous covariates perform as well in terms of size and power as those with binary covariate.

[Insert Tables 1, 2, 3, and 4]

### 4.2 Tests of the MAR assumption

The size and power of the HT tests based on Inverse Probability Weighted (IPW) and two-step estimators are examined. These estimators use the MAR assumption (i.e. selection on observables) in obtaining probability weights and the Inverse Mills Ratio.

#### 4.2.1 DGP for IPW estimators

Outcome and selection are determined by the following two equations:

\[
y_{it} = x_{it} \cdot \beta + c_{1i} + u_{it}, \text{ for } i = 1, 2, \ldots, n \text{ and } t = 1, 2, \ldots, T
\]

\[
s_{it} = 1(a^{*} + \gamma_{1} z_{1it} + \gamma_{2} z_{2it} + v_{it} > 0), \text{ for } i = 1, 2, \ldots, n \text{ and } t = 1, 2, \ldots, T
\]

where the value of \( a^{*} \) is chosen to set the missing proportion of data, in a way that missing proportion increases as value of \( a^{*} \) increases. We consider DGPs to be similar to those in the previous section.

#### DGP III

1. \( c_{1i} \sim \gamma_{2} z_{2i} + \text{i.i.d.} N(0, \frac{1}{4}) \) where \( \gamma_{2} = 0 \).\(^{16}\)

2. \( z_{2i} \sim \text{i.i.d.} N(0, 1); z_{1it} = \text{i.i.d.} N(0, 1) \) and \( \gamma_{1} = 1 \).

3. \( x_{it} \sim z_{2i} + \text{i.i.d.} N(0, 1); \beta = 1 \).\(^{17}\)

4. \( u_{it} \sim \gamma_{1} z_{1it} + \text{i.i.d.} N(0, 1); v_{it} \sim \frac{z_{1it}}{3} + \frac{z_{2i}}{3} + \text{i.i.d.} N(0, 1) \).

\(^{16}\)Variance of \( \frac{1}{4} \) is chosen to generate the inconsistency of 20% for a pooled LS estimator. Variance of 4 gives inconsistency of 60% for a pooled LS estimator.

\(^{17}\)We test the power of HT tests with \( \gamma_{1} = 0 \) and \( \gamma_{2} = 1 \) and the size of HT tests with \( \gamma_{1} = 1 \) and \( \gamma_{2} = 0 \).
The MAR assumption is used to model selection/missing process determined by observed \( z_{it} = (z_{1it}, z_{2it}, x_{it}) \). We assume that these variables are available for \( s = 1 \) and \( s = 0 \). The probability weight is estimated from

\[
E(s = 1|z_{it}) = \Phi(\gamma_1 z_{1it} + \gamma_2 z_{2it}) \equiv G(z_{it}, \gamma) \tag{23}
\]

where \( \Phi(\cdot) \) is the standard normal CDF and (23) is estimated by a Probit model. For balanced data, all pooled LS, FE and FD estimators for \( \beta \) are consistent and these estimators become inconsistent only conditional on \( s = 1 \). A covariate \( x \) and unobserved factor \( u \) are correlated indirectly through unobserved components \( z_1 \) and \( z_2 \) in \( s \). As long as (23) is a correct specification for the selection/observability model, the pooled LS, FE and FD estimators with probability weights can remove inconsistency from non-random missing and deliver consistent estimates because, according to DGP I, the inconsistency only arises because of non-random missing.

We adjust DGP III by replacing gammas and \( v \) with \( \gamma_2 = 1, \gamma_1 = 0 \) and \( v_{it} \sim i.i.d. N(0,1) \). We denote this by DGP III-B. In this DGP, the pooled IPW-LS for \( \beta \) is inconsistent but the IPW-FE and IPW-FD estimators are consistent because the source of inconsistency, \( z_{2it} \), is time-invariant. Thus, the HT test statistics using the pooled IPW-LS and IPW-FE estimators (or using the pooled IPW-LS and IPW-FD estimators) should reject the null hypothesis of \( H_0 : \beta_{LS} = \beta_{FE} \) (or \( H_0 : \beta_{LS} = \beta_{FD} \)) while those using the IPW-FE and IPW-FD estimators should not reject the null.

We further adjust DGP III with the following changes and denote it by DGP III-C:

1. \( c_{1it} \sim i.i.d. N(0,1) \).
2. \( z_{1it} \sim i.i.d. N(0, \frac{100}{(T-t)^2}) \); \( z_{2it} \sim i.i.d. N(0, \frac{100}{t^2}) \).
3. \( x_{it} \sim z_{1it} + i.i.d. N(0,1) \); \( \beta = 1 \).
4. \( u_{it} \sim z_{2it} + i.i.d. N(0,1) \); \( \gamma_1 = 1 \); \( v_{it} \sim z_{1it} + z_{2it} + i.i.d. N(0,1) \).

According to DGP III-C, the source of non-random missing is the time-varying component. So the FE and FD estimators are inconsistent but the magnitude of inconsistency of the two estimators is different.

### 4.2.2 DGP for sample-selection correction by inverse mills ratio

To examine the size of HT tests based on estimators with Inverse Mills Ratio (IMR) correction, we consider DGPs that satisfy ideal conditions for the pooled LS, FE and FD estimators to be consistent. We

\[ z_{1it} \text{ and } z_{2it} \text{ are the sources of inconsistency and large variances are chosen to generate large enough discrepancy between FE and FD estimators.} \]
assume multivariate normal distribution for the errors from outcome and selection equations. The se-
lection is determined by

\[ s_{it} = 1(a^* + \gamma_1 z_{1it} + v_{it} > 0), \] for \( i = 1, 2, ..., n \) and \( t = 1, 2, ..., T, \)

where we set the value of \( a^* \) to select the missing proportion of data. We consider the following DGPs.

**DGP IV**

1. \( c_{1i} \sim i.i.d. N(0, 1) \).
2. \( z_{1it} \sim i.i.d. N(0, 1); \gamma_1 = 1 \).
3. \[
\begin{pmatrix}
    x_{it} \\
    u_{it} \\
    v_{it}
\end{pmatrix}
\sim MV
\begin{pmatrix}
    1 & 0 & \rho_1 \\
    0 & 1 & \rho_2 \\
    \rho_1 & \rho_2 & 1
\end{pmatrix},
\]

where \( MV \) is multivariate normal with \( \rho_1 = 0.7 \) and \( \rho_2 = 0.5 \).

Again non-random missing is the only source of inconsistency. \( x_{it} \) and \( u_{it} \) are not directly correlated but \( x_{it} \) is correlated with \( u_{it} \) through \( v_{it} \). With proper correction by the IMR, all Two-Step (TS) estimators – TS-LS, TS-FE and TS-FD – are consistent. In two-step estimation, we can obtain IMR from the first-stage Probit estimation of selection model:

\[ E(s = 1|z_{1t}) = \Phi(a + \gamma_1 z_{1it}). \] (24)

At the second step, we apply the pooled LS, FE and FD estimations\(^{20}\) to the following equation to estimate \( \beta \):

\[ y_{it} = x_{it} \cdot \beta + \rho \frac{\phi(\hat{a} + \hat{\gamma}_1 z_{1it})}{\Phi(\hat{a} + \hat{\gamma}_1 z_{1it})} + c_{1i} + u_{it}, \] for \( s_{it} = 1 \). (25)

To examine the power of HT tests, we need to induce inconsistency for the pooled LS estimator by adjusting DGP IV as follows: \( c_{1i} \sim z_{2i} + i.i.d. N(0, 1) \), \( x_{it} \sim z_{2i} + i.i.d. N(0, 1) \), \( z_{2i} \sim i.i.d. N(0, \frac{1}{4}) \) and \[
\begin{pmatrix}
    u_{it} \\
    v_{it}
\end{pmatrix}
\sim MV
\begin{pmatrix}
    1 & \rho_1 \\
    \rho_1 & 1
\end{pmatrix}
\]

where \( \rho_1 = 0.5 \).\(^{21}\) We denote this by DGP IV-B. Thus, the TS-FE and TS-FD estimators are consistent while the TS-LS estimator is inconsistent. We examine the power of two HT tests using that compares TS-LS and TS-FE estimators and using that compares the TS-LS and TS-FD

\(^{19}\) \( a = -0.5 \) is chosen for the missing proportion of 36%.

\(^{20}\) Standard errors are obtained by bootstrap to avoid first-stage estimation error of \( \gamma_1 \).

\(^{21}\) The values for the variance of \( z_{2i} \) and \( \rho_1 \) are chosen to induce inconsistency of about 20%.
estimators. We examine the size of the HT test that uses the comparison between the TS-FE and TS-FD estimators.

Finally, to examine the power of the HT test that uses the TS-FE and TS-FD estimators, we further adjust DGP IV as the following DGP IV-C. DGP IV-C induces inconsistency for the TS-FE and TS-FD estimators, but the magnitudes of inconsistency are different.

**DGP IV-C**

1. $c_{1i} \sim i.i.d. N(0,1)$.

2. $z_{1it} \sim i.i.d. N(0,1); \gamma_1 = 1$.

3. \[
\begin{pmatrix}
  tx_{it} \\
  tu_{it} \\
  tv_{it}
\end{pmatrix} \sim MV
\begin{pmatrix}
  1 & 0 & \rho_1 \\
  0 & 1 & \rho_2 \\
  \rho_1 & \rho_2 & 1
\end{pmatrix},
\]
   where $MV$ is multivariate normal with $\rho_1 = 0.7$ and $\rho_2 = 0.7$.

4. $x_{it} = c_{1i} + \frac{tx_{it}}{T}; u_{it} = c_{1i} + T \cdot \frac{tu_{it}}{T}; v_{it} = T \cdot \frac{tv_{it}}{T-1}$.

### 4.2.3 Simulation results

MC simulations are performed with 2,000 replications for $T=4, 8, 16$ and $n=100, 200, 400$ in all simulations. The size distortion of the HT tests of the MAR using IPW and two-step methods is examined in Table 5. The power of the tests is provided in Table 6 for the DGP where source of inconsistency is time-invariant and in Table 7 for the DGP where the source of inconsistency is time-varying. True value of $\beta$ is 1 and significance level of test is 0.05 in all estimations. The first three columns in Table 5 report the point estimates of the MC mean of $\beta$ and their coverage rate of 95% confidence interval. The first three panels show the results for the IPW method and the last three panels show the results for the two-step method. Each panel represents different time periods. In all LS, FE and FD estimations, the IPW and two-step methods are consistent. We perform MC experiments with various missing proportions but we only report on a missing proportion of 36% to save space. The size of test is correct for HT tests based on the IPW and two-step methods.

Table 6 shows the results for the DGP for which the source of non-random missing is a time-invariant unobserved factor. So, under correct specification of the MAR assumption for IPW and IMR, the FE and FD estimators are consistent while the pooled LS is inconsistent. This is verified from the point estimates of the MC mean of $\beta$ in the first three columns of Table 6. The first three panels show the results for the
IPW method and the last three panels show the results for the two-step method. For the sample size of \( n \) greater than 400 and for all \( T = 4, 8, 16 \), the power of tests using the difference between the LS and FE (or LS and FD) estimators is close to 1. The power of tests increases with sample size \( n \), \( T \) and missing proportion. We report results from the DGP with a conservative missing proportion of about 35% and inconsistency of 20%. Given that our empirical application has a sample size of \( n = 46,440 \) and \( T = 33 \) and \( a^* = 47\% \), it is safe to say that the power should be close to 1 for HT tests with IPW and IMR. In this DGP, the FE and FD estimators are consistent. As shown in the last column, the HT tests that use the difference between the FE and FD estimates show correct size of 0.05. For all sample sizes, coverage rate of 95% confidence intervals contains 0.05.

Table 7 reports the results for the DGP for which the source of non-random missing is a time-varying unobserved factor. According to DGP III-C and DGP IV-C, the FE and FD estimators are inconsistent but the magnitudes of inconsistency differ for the FE and FD estimators. The magnitudes of difference between the FE and FD estimates are 0.12 (12%) for the IPW estimator and 0.52 (52%) for the IMR estimator. As for \( n \geq 600, T > 8 \) and \( a^* \approx 40\% \), the power of test approaches 1 for HT tests based on IPW and IMR methods.

[Insert Tables 5, 6, and 7]

5 Empirical Application: The effect of trade liberalization on trade flows

We investigate the effects of trade liberalization policies, including being a member of GATT/WTO (denoted by WTO membership hereafter), RTA (Regional Trade Agreement membership), and CU (Currency Union membership), on bilateral trade flows using a gravity model. We focus on the endogeneity problem caused by missing data using the HT tests. An early empirical study (Rose [2004]) that treated missing data as reliant on a selection on observables assumption, failed to find evidence of positive trade creation effect of WTO membership. Most of following studies emphasize that the results in Rose [2004] could be tainted because of omitted variable bias (Subramanian and Wei [2007], Liu [2009], Eicher and Henn [2011], Roy [2011], Chang and Lee [2011]) or heteroskedastic error (Silva and Tenreyro [2006]) than because of missing data. The proposed HT test could provides a formal test of whether the bias for the

---

22 Data and model specification may also contribute to Rose [2004]'s unexpected results. Subramanian and Wei [2007] emphasize heterogeneity in the GATT/WTO system, e.g. developing countries vs. developed countries; old GATT members vs. new members; they find heterogeneous effect of WTO memberships on trade flows. Furthermore, Tomz et al. [2007]
effect of WTO membership remains even after account for omitted variables as well as heteroskedastic error. Assuming that omitted variables, heteroskedastic error, and non-random missing data are the only sources of bias, the rejection of the null implies that bias due to non-random missing data is not correctly properly.

5.1 Gravity model

Recent studies on the effects of WTO membership usually ignore missing data or rely on the assumption of 'selection on observables' to treat missing data, and focus on removing omitted variable bias by accounting for the multilateral resistance terms (MRTs) in e.g. Anderson and van Wincoop [2003] and unobserved country-pair heterogeneity in e.g. Baldwin and Taglioni [2006]. MRTs capture the effects of trade barriers having a larger impact on relatively small countries, and unobserved country-pair heterogeneity captures cultural and political ties of pair countries. Following these studies, we account for these two unobserved factors – country-pair heterogeneity (UCPH) (e.g., Magee [2003], Baldwin and Taglioni [2006], Baier and Bergstrand [2007]) and unobserved country-time heterogeneity (UCTH) (e.g., Anderson and van Wincoop [2003], Magee [2008], Eicher et al. [2012]) – in the gravity model using the following methods below. Fixed effects (FE i.e. within transformation) and first-differencing (FD) are adopted to account for UCPH; and approximation of MRTs from Baier and Bergstrand [2009] is used to account for UCTH. Our baseline gravity model can be written as the following equation:

\[
\ln(TF_{ijt}) = \alpha_0 + \alpha_1 \ln Y_{it} + \alpha_2 \ln Y_{jt} + \alpha_3 \ln y_{it} + \alpha_4 \ln y_{jt} + X_{ijt} \beta + \gamma_1 WTO_{ijt} + \gamma_2 RTA_{ijt} + \gamma_3 CU_{ijt} + \mu_t + \omega_{ij} - \ln P_{it}^{1-\sigma} - \ln P_{jt}^{1-\sigma} + \epsilon_{ijt} \tag{26}
\]

and is subject to the following \( n \) nonlinear market-equilibrium conditions:

\[
\ln P_{it}^{1-\sigma} = \sum_{k=1}^n \ln P_{kt}^{1-\sigma}(Y_{kt}/Y_{Wt})e^{X_{ijt} \beta + \gamma_1 RTA_{ijt}}, i = 1, \ldots, n, \tag{27}
\]

where the \( WTO_{ijt} \) is a dummy variable equal to one if importer \( i \) and exporter \( j \) are WTO members in year \( t \); \( TF_{ijt} \) is trade flow; \( RTA_{ijt} \) is the regional trade agreement dummy variable which is equal to one if importer \( i \) and exporter \( j \) are members of either a free trade agreement or customs union; \( CU_{ijt} \) is a re-define WTO membership using actual participation rather than formal membership and find evidence of positive WTO effect. More recently, literature also argues that omission of zero observations causes sample selection bias and shows that inclusion of zero bilateral trade flows results in a positive WTO effect [Liu, 2009, Konya et al., 2011, Chang and Lee, 2011].
currency unions dummy variable; \( Y \) is real GDP measured in purchasing power parity (PPP) terms; \( y \) is real per capita GDP; \( X \) is a vector of other determinants usually included in a gravity model, such as the general system of preferences (\( GSP_{ijt} \) and \( GSP_{jiti} \)), hostility, alliance, distance, area, border, landlock, island, common language, common religion, colony, colonizer, current colony, current colonizer, common colony, and remoteness; \( \mu_t \) represents any unobserved global trend in trade and aggregate shocks in each year; \( \omega_{ij} \) is country-pair heterogeneity; \( P^1-\sigma_k \) is an MRT where \( P_k \) is the price level of country \( k \) and \( \sigma \) is the elasticity of substitution in consumption; and \( \epsilon_{ijt} \) is an idiosyncratic error term. Specifically, we account for unobserved MRTs in the equation (26) by using a linear approximation of the MRTs, as in the equation (28) which is suggested in Baier and Bergstrand [2009]:

\[
\ln(TF_{ijt}) = a_0 + a_1 \ln Y_{it} + a_2 \ln Y_{jt} + a_3 \ln y_{it} + a_4 \ln y_{jt} + (X_{ijt} - MRX_{ijt}) \cdot \beta + \\
\gamma_1 (WTO_{ijt} - MRWTO_{ijt}) + \gamma_2 (RTA_{ijt} - MRRTA_{ijt}) + \\
\gamma_3 (CU_{ijt} - MRCU_{ijt}) + \mu_t + \omega_{ij} + \epsilon_{ijt}
\]

and

\[
MRz_{ijt} = \sum_{k=1}^{N} \left( \frac{Y_{kt}}{Y_{Wt}} \right) z_{ikt} + \sum_{m=1}^{N} \left( \frac{Y_{mt}}{Y_{Wt}} \right) z_{jmt} - \sum_{k=1}^{N} \sum_{m=1}^{N} \left( \frac{Y_{kt}}{Y_{Wt}} \right) \left( \frac{Y_{mt}}{Y_{Wt}} \right) z_{kmt}
\]

where \( Y_{kt} \) is \( k \)-th country’s GDP at time \( t \), \( Y_{Wt} \) is world GDP at time \( t \), \( MRz_{ijt} \) is a linearized MRT of \( z_{ijt} \), with \( z_{ijt} \in Z_{ijt} = (X_{ijt}, WTO_{ijt}, RTA_{ijt}, CU_{ijt}) \). Therefore, in the equation (28), the effects of trade liberalization policies – WTO, RTA and CU – are decomposed into two parts: the first term is to capture the direct effect of policy and the second term is to capture the indirect effect induced by the change in MRT terms on \( \ln(TF) \) where our primary interest is direct effect. WTO trade effects remain to be measured by \( \hat{\gamma}_1 \) and RTA and CU effects are measured by \( \hat{\gamma}_2 \) and \( \hat{\gamma}_3 \).

### 5.1.1 Data

The primary dataset is obtained from Liu [2009]. It covers 216 countries from 1948 to 2003 so the number of observations is more than 2 million. The GDP and population data are obtained from the PWT6.1, PWT5.6, WDI2003, Maddison Historical Statistics, the IMF International Financial Statistics (IFS) and the United Nations Statistical Yearbooks (UNSYB). The US consumer price index (CPI) is used to convert
these GDP measures into 1995 real dollar terms. All GDP data used in this paper, except those from the IFS, are measured by purchasing power parity (PPP) methods. The PWT6.1 dataset is taken as the base source for GDP data and any missing data are filled with data from other datasets after being multiplied by a ratio calculated from the overlapped GDP data in the two datasets. GATT/WTO formal membership and RTA and CU data are obtained from the WTO website.

5.1.2 Biases from omitted variable and non-random missing

Most empirical studies in gravity model literature rely on random missing (i.e. the validity of the MCAR assumption conditional on observed variables) to justify ignoring missing data. However, as shown in Table 8, the missing proportion is too substantial to simply ignore without providing any justification. In our example, missing proportions are about 47% in the full sample for the test of the MCAR assumption and 43% in the restricted sample for the test of the MAR assumption. The sample is reduced to allow maximum availability of observations for exclusion restriction variables.

[Insert Table 8]

Although there are a few early studies that treat non-random missing data with imputation (Linders and de Groot [2006] and Felbermayr and Kohler [2006]), the assumption used for imputation has been rarely tested. Moreover, to our knowledge, the existing papers that treat non-random missing are limited in the way they account for omitted variable bias caused by not controlling MRTs and UCPH. MRTs and UCPH turn out to be major sources of omitted variable bias so proper control of these factors is critical for a consistent estimation and previous literature such as Magee [2003], Baldwin and Taglioni [2006], Baier and Bergstrand [2007], Anderson and van Wincoop [2003], Feenstra [2004] accounted for these factors by fixed effects. This study considers both omitted variables bias and non-random missing bias simultaneously using the HT test. In particular, we control for MRTs and UCPH in all our estimations following the popular methods that are widely used in previous studies such as Magee [2008], Eicher et al. [2012] while we examine whether missing data are conditionally random using the HT tests of the MCAR assumption and MAR assumptions. The missing data problem is important because missing mostly occurs for countries with specific attributes such as least developed or small in size – two characteristics also known to affect dependent variable (i.e. ‘trade flows’ is missing if a country pair do not trade and these countries are less likely to engage in international trade) – or a newly-independent country, which implies that missing data is not random. However, if, for instance, income and size of countries included in equation (26) completely determine missing data, it is possible to obtain consistent estimates using
the MAR assumption (also called ‘selection on observables’ assumption) with imputation, IPW or/and IMR methods. Therefore, the HT tests of the MAR assumption could be useful to determine whether the correction made by using the MAR assumption with these three methods can mitigate/remove bias from non-random missing.

5.1.3 The HT test of the MCAR assumption

We consider two panel data estimators, namely the FE and FD estimators, which differ only in accounting for UCPH ($\omega_{ij}$) in application of the HT test. The FE and FD estimators in the equations (26) and (28) are consistent as long as either the MCAR assumption or the MAR assumptions for imputation, IPW, or IMR are valid.

First, we perform an HT test using the FD and FE estimators to examine whether the MCAR assumption is valid in Table 6. Columns (4) and (5) report the estimation results for equations (26) and (28) where MRTs are controlled by the method in Baier and Bergstrand [2009] and UCPH is controlled by within-transformation and first-differencing for the FE and FD estimators, respectively. Assuming that unobserved factors, $\omega_{ij}$, $u_{it}$ and $v_{jt}$ (i.e. UCPH and MRTs) are properly controlled, we have:

$$0 = Cov(\epsilon_{it}, WTO_{ijt}|RTA_{ijt}, CU_{ijt}, x_{ijt}, \omega_{ij}, u_{it}, v_{jt}).$$

So, if missing is not correlated with the variable of our interests (i.e. trade liberalization policy) conditional on observed variables and $\omega_{ij}, u_{it}, v_{jt}$ (i.e. we refer to this conditional random as ‘the MCAR’ hereafter), we have:

$$Cov(\epsilon_{it}, WTO_{ijt}|RTA_{ijt}, CU_{ijt}, x_{ijt}, \omega_{ij}, u_{it}, v_{jt}) = Cov(\epsilon_{it}, WTO_{ijt}|RTA_{ijt}, CU_{ijt}, x_{ijt}, \omega_{ij}, u_{it}, v_{jt}, s_{it} = 1)$$

Thus, under conditional random assumption, any difference between the FE and FD estimates should be due to sampling errors; so differences beyond sampling errors should be caused entirely by non-random missing. In other words, the rejection of the HT test in column (6) implies violation of the MCAR assumption and that the estimates in columns (4) and (5) are inconsistent; additionally, the non-random missing must be the source of bias, assuming that the omitted variable bias is properly controlled. In columns (4) and (5), the estimated effects of the WTO membership on trade flows for the FE and FD estimators are 0.002% and -0.05%, respectively. The difference between the estimates for the WTO ef-
effects is not statistically significant, with a p-value of 0.49, so the HT test cannot reject the null of random missing for WTO membership variable. It implies that

\[ 0 = \text{Cov}(\epsilon_{it}, \text{WTO}_{ijt}|\text{RTA}_{ijt}, \text{CU}_{ijt}, \omega_{ij}, u_{it}, v_{jt}) \]

\[ = \text{Cov}(\epsilon_{it}, \text{WTO}_{ijt}|\text{RTA}_{ijt}, \text{CU}_{ijt}, \omega_{ij}, u_{it}, v_{jt}, s_{it} = 1) \]

where \( x_{ijt} \) includes all other explanatory variables. Likewise, for other trade liberalization policy variables, namely RTA and CU, we perform the HT tests using the estimates in columns (4) and (5). The estimated positive trade creation effects of the RTA membership on trade flows for the FE and FD estimators are 90% and 13%, respectively. The difference is statistically significant, with a p-value of zero, so the HT test rejects the MCAR assumption. This implies that non-random missing causes inconsistency for the RTA effect:

\[ 0 = \text{Cov}(\epsilon_{it}, \text{RTA}_{ijt}|\text{WTO}_{ijt}, \text{CU}_{ijt}, \omega_{ij}, u_{it}, v_{jt}) \]

\[ \neq \text{Cov}(\epsilon_{it}, \text{RTA}_{ijt}|\text{WTO}_{ijt}, \text{CU}_{ijt}, \omega_{ij}, u_{it}, v_{jt}, s_{it} = 1) \].

Finally, the estimated positive trade creation effects of the currency union (CU) membership on trade flows for the FE and FD estimators are 868% and 4,044%, respectively. The difference is highly statistically significant and this implies that

\[ 0 = \text{Cov}(\epsilon_{it}, \text{CU}_{ijt}|\text{WTO}_{ijt}, \text{RTA}_{ijt}, \omega_{ij}, u_{it}, v_{jt}) \]

\[ \neq \text{Cov}(\epsilon_{it}, \text{CU}_{ijt}|\text{WTO}_{ijt}, \text{RTA}_{ijt}, \omega_{ij}, u_{it}, v_{jt}, s_{it} = 1) \].

The HT tests imply non-random missing for RTA and CU variables and these are identified from the different magnitudes of bias for the FE and FD estimators, assuming that we properly control for omitted variables bias.

[Insert Table 9]
5.1.4 The HT tests of the MAR assumption

In our data, missing occurs mostly due to the dependent variable, bilateral trade flows, for least developed and small countries. Two explanatory variables, GDP and GDP per capita also have some missing observations but only a small proportion of them. Following previous studies such as Linders and de Groot [2006] and Felbermayr and Kohler [2006], we impute missing observations for trade flows. Note that those countries with missing trade flows are typically (i) small and less developed, (ii) closed, or (iii) least developed countries. Therefore, first, for the case of less developed and small countries, missing observations typically imply either little trade or no trade. Using the fact that rounding exercises are common for trade flow data, regarding little trade flow as zero flows is unlikely to be an additional source of bias. Second, by definition, missing observations of closed countries are zero trade flows. Least developed countries typically have not maintained proper records and their trade flows are either small or zero, at least for most of our sample period. Therefore, we argue that no serious additional bias comes from this rounding exercise. It is not unreasonable to assume that missing observations are either zero or very close to zero trade flows. For GDP and GDP per capita data, we use a simple method for imputation.

We also perform the HT test with the IPW or Heckman’s two-step methods as an illustration of using the HT tests of the MAR assumptions. The MAR assumptions are used to model the probability of observability and self-selection of trade. In the estimation of the probability of observability/selection, we use all available explanatory variables in the trade flows equation and two exclusion restrictions – religious proximity (as in Helpman et al. [2008]) and lagged/mean trade flows – as predictors. The choice of exclusion restrictions is limited and even problematic because of the limitation on observed variables. Thus, our main purpose here is to illustrate the use of the HT tests that examine whether the MAR assumption (the model for observability/selection) is valid so it is effective to reduce/remove bias from non-random missing data.

Assuming proper control of omitted variables, if the MAR assumptions with imputation, IPW, and two-step methods are valid, these proposed correction methods remove the inconsistency of both FE and FD estimators caused by non-random missing. Then, no rejection of the HT test indicates both the FE and FD estimators are consistent. On the other hand, the rejection of the HT test implies that either

\[23\text{After imputation, data becomes balanced. Therefore, controlling for MRTs and UCPH by the dummy variables approach only requires the availability of } ijt\text{-varying variables because the FE and FD estimators with imputation can remove all } ij\text{-varying, } it\text{-varying, and } jt\text{-varying factors. Thus, there is no need to impute GDP and GDP per capita data, which are } it\text{-varying and } jt\text{-varying.}\]
omitted variables bias is not properly handled or that missing is not random, even after conditioning on observed variables (i.e. a violation of the MAR assumption).

5.1.5 Implementation

First, for the first-stage estimation of the IPW and two-step methods, the following observability/selection model is estimated:

\[
Pr(s_{ijt} = 1|z_{ijt}) = \Phi(\alpha_0 + \alpha_1 \ln Y_{it} + \alpha_2 \ln Y_{jt} + \alpha_3 \ln y_{it} + \alpha_4 \ln y_{jt} + \beta \ln T_{ijt}^* + \gamma_1 WTO_{ijt} + \gamma_2 RTA_{ijt} + \gamma_3 CU_{ijt} + \delta \cdot religion),
\]

(30)

where average trade flows are \(T_{ijt}^* = \frac{1}{T_{ij}} \sum_{t=1}^{T} s_{ijt} T_{ijt}\) (or other function of past and future values of trade flows) and \(T_{ij} = \sum_{t=1}^{T} s_{ijt}\). Note that predictors in the selection model include all explanatory variables in the outcome equation and exclusion restrictions, namely lagged/mean trade flows and religious proximity variables. Intuitively, the majority of missing observations for trade flows occur in three forms – poor data keeping, no trade flows, and the rounding of small trade flows – and religious proximity variable used as an exclusion restriction in Helpman et al. [2008] can help explain, at least partially, why observations are missing for those pairs that do not trade.

Finally, for our imputation method, we simply replace missing trade flows with zero flows. The MAR assumption we use is:

\[
E(TF_{mt}, TF_{mt-1} | z_{mt}, TF_{mt-1}) = 0.
\]

(31)

Given that most of the missing trade flows are from small and least developed countries, it is not unreasonable to assume that those trade flows are zero or close to zero. For GDP and GDP per capita, we apply simple imputation by extrapolating from a linear trend model, based on the idea that GDP and GDP per capita are probably persistent, at least in the short run, so any available observations from other periods can help predict missing data.\(^{24}\)

Under the validity of the MAR assumption used either to model selection/missing process of \(Pr(S_{ijt} = 1|z_{ijt})\) in equation (30) or to impute missing observations with zero flows, the IPW, two-step, and impu-

\(^{24}\)Variable is the only variable excluded in the gravity model but is added as an exclusion restriction to the imputation model. We estimate MI model with various \(T_{mt,ijt}\) including minimum and mean values for each pair over time. But the MI estimates change little and statistical conclusions of the HT tests do not change depending on specific choices of \(T_{mt,ijt}\).
The following exogeneity conditions hold for the IPW and two-step methods, respectively:

\begin{equation}
0 = E(\epsilon^*_i | CU^*_ijt, WTO^*_ijt, RTA^*_ijt, x^*_ijt, \omega^*_ij, u^*_it, v^*_jt, s_{ijt} = 1) = E(\epsilon_i | CU_{ijt}, WTO_{ijt}, RTA_{ijt}, x_{ijt}, \omega_{ij}, u_{it}, v_{jt}),
\end{equation}

where variables with an asterisk denote inverse probability weighted variables, and

\begin{equation}
0 = E(\epsilon_{it} | CU_{ijt}, WTO_{ijt}, RTA_{ijt}, x_{ijt}, \omega_{ij}, u_{it}, v_{jt}, IMR_{ijt}, s_{ijt} = 1) = E(\epsilon_{it} | CU_{ijt}, WTO_{ijt}, RTA_{ijt}, x_{ijt}, \omega_{ij}, u_{it}, v_{jt}),
\end{equation}

The following exogeneity condition (34) holds for observed and imputed observations combined:

\begin{equation}
0 = E(\epsilon_{it} | CU^c_{ijt}, WTO^c_{ijt}, RTA^c_{ijt}, z^c_{ijt}, \omega_{ij}, u_{it}, v_{jt}),
\end{equation}

where $c$ denotes combined data.

Table 10 reports estimation results for the HT tests of the MAR assumption. In contrast to the test of the MCAR assumption which uses a sample from 1948 to 2003, we restrict sample periods to post-1970 periods in order to increase the availability of the religious proximity variable (i.e. an exclusion restriction) that is used as a predictor to model a missing process for the first stage estimations of the IPW and two-step methods.\(^{25}\) Missing data mostly occurs in trade flows. The first three columns from (1) to (3) report results for the MCAR assumption. The difference is within the sampling errors for the RTA variable coefficients only. For the coefficients on WTO and CU variables, the differences between the FE and FD estimates are large enough to reject the null hypothesis. The columns from (4) to (6) report the results for two-step estimators with Heckman’s selection correction terms and the columns from (7) to (9) report results for the IPW estimators. The results for the HT tests of the MAR assumption using two-step and IPW estimators are qualitatively the same as the tests of the MCAR assumption in Table 6. In fact, the estimates from conventional methods are not significantly different from those from \(^{25}\)This sample restriction is not necessary for imputation model.
IPW and two-step estimators. Assuming the omitted variable bias is removed, the results of the tests on WTO and CU variables show evidence for violation of the MAR assumption. This implies that bias from non-random missing is not removed or mitigated noticeably for either FE or FD estimators, because the estimates of three coefficients on WTO, RTA and CU change little, compared to the estimates in the first three columns of Table 10. The columns from (10) to (12) report the HT test results for the imputation method based on replacing missing trade flows entries with zero flows. The estimated FE and FD estimates with imputation are noticeably different from the estimates obtained by conventional, IPW and two-step estimators. This implies that if our assumption used to impute missing trade flows is correct, the estimated effect of WTO and RTA derived from conventional (i.e. ignoring missing data), IPW and two-step estimators are severely underestimated.

The HT test rejects the null hypothesis and this implies that the MAR assumptions for the coefficients on WTO and RTA variables are violated. However, if we believe the MAR assumption used for imputation is valid, we can infer that omitted variable bias for the coefficients on WTO and RTA variables is not completely removed by the control of MRTs and UCPH using fixed effects. For instance, if we assume the MAR assumption for imputation is valid, the HT tests of columns (1) and (3) capture evidence of the bias caused by omitted variables and non-random missing, while those from columns (10) and (12) capture the bias from omitted variable bias only.

By comparing the estimates of the coefficients on WTO and RTA in the columns (1) and (10) (or (2) and (11)), the direction of biases from non-random missing can be obtained. WTO and RTA effects are dramatically underestimated for the FE and FD estimators, while CU effects are underestimated for the FE estimator and overestimated for the FD estimator if the MAR assumption with imputation is valid and missing data is ignored. Intuitively, underestimation of conventional estimators is quite natural because including zero trade flows changes the composition of control groups in the effects of WTO and RTA treatment estimations. Missing data mostly occurs to non-member countries for WTO and RTA and they are removed from the comparison group due to missing data. For instance, using more conservative results for FD estimates in columns (2) and (11), we obtain that the WTO membership increases trade flows by 0% without imputation and 55% with imputation, while RTA membership increases them by 13% without imputation and 105% with imputation. Therefore, the WTO and RTA effects are underestimated by 55 and 92 percentage points, respectively, if we ignore missing data. Furthermore, the differences of estimates for WTO and RTA variables are shown in columns (10) and (11) and the HT tests of the equality of estimates in columns (11) and (12) are shown in column (12). Assuming that
our imputations for missing trade flows are valid and FE and FD estimators only differ in accounting for unobserved factors, the omitted variables bias may not be completely removed. Finally, the HT test results for the coefficient on CU in (12) imply that, for the CU effect there is no evidence of bias caused by non-random missing or omitted variables.

[Insert Table 10]

6 Conclusion

We extend the Hausman specification test to test whether the MCAR assumption (i.e. random missing) and the MAR assumption (i.e. random missing conditional on observables) are satisfied for unbalanced panel data. The test for the validity of the MAR assumption is particularly useful because it is generally regarded as an untestable assumption. The proposed HT tests are based on the difference between two consistent estimators under the null hypothesis. The idea of the test rests on the fact that the difference should be entirely because of sampling error if the MCAR assumption (or the MAR assumption) is satisfied. Therefore, if the null hypothesis is rejected, the difference beyond sampling error should be attributed to non-random missing or invalid methods used with the MAR assumption to correct the bias from non-random missing. We provide a Stata ado code to implement the proposed HT tests so that they can be widely applied to any unbalanced panel data where two consistent estimators are available under the null hypothesis. Finite-sample performances of the proposed tests are examined for the size and power in a Monte Carlo experiment and the results show correct size and high power of the tests in various data generating processes, and various forms of the MAR assumption (e.g. the MAR assumptions are used to obtain imputed observations, inverse probability weights, and the Inverse Mills Ratio).

We illustrate the empirical usefulness of the test with a study for the effects of trade liberalization policies such as WTO, RTA, and CU membership on trade flows where units with missing observations are composed of 47% of the full sample. Extensive empirical studies exist for the effects of trade liberalization policies but the missing data problem has not received much attention generally because of the limited information for missing process. We adopt gravity equation with error component models that are widely used in this literature. The proposed tests provide several important implications about potential biases from omitted variable and non-random missing in gravity model estimations. First, the HT test results indicate violation of the MCAR assumption. Second, the results also imply that using the
MAR assumption with inverse probability weight or Inverse Mills Ratio correction helps little in reducing biases originating from non-random missing data, because exclusion restriction variables we used have very little variation (or explaining power) with observability status. However, in the application of imputation, the validity of imputation for trade flows can be justified because missing trade flows, at least in our sample, mostly occur in the data from very poor and small sized countries where imports are rare or in very small amounts so that missing observations for trade flows could be replaced with zero. The results from imputation imply that severe biases exist for the coefficients on WTO and RTA variables if missing is ignored. The effects of WTO and RTA on trade flows are substantially underestimated. We can also infer that, if our method of imputation can be justified, the differences in the coefficients on WTO and RTA can be attributed to omitted variable bias.
References


S. Chaudhuri and D. Guilkey. GMM with multiple missing variables. mimeo, 2014.


C. Muris. Efficient GMM estimation with a general missing data pattern. mimeo, 2011.


Table 1: The size and power of the test where $T = 4$ and DGPI

<table>
<thead>
<tr>
<th>$T$</th>
<th>$n$</th>
<th>$DGP$</th>
<th>40 percent of data is missing and $\rho = 0$</th>
<th>60 percent of data is missing and $\rho = 0$</th>
<th>40 percent of data is missing and $\rho = 0.9$</th>
<th>60 percent of data is missing and $\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>coverage</td>
<td>coverage</td>
<td>coverage</td>
<td>coverage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>1.00</td>
<td>.99</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FE</td>
<td>FE</td>
<td>FE</td>
<td>FE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.038</td>
<td>.038</td>
<td>.038</td>
<td>.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.064</td>
<td>.064</td>
<td>.064</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.046</td>
<td>.046</td>
<td>.046</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.044</td>
<td>.044</td>
<td>.044</td>
<td>.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.035</td>
<td>.035</td>
<td>.035</td>
<td>.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.055</td>
<td>.055</td>
<td>.055</td>
<td>.055</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.064</td>
<td>.064</td>
<td>.064</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.062</td>
<td>.062</td>
<td>.062</td>
<td>.062</td>
</tr>
</tbody>
</table>

Note: The number of replications is 2,000. When $\rho = 0$, observability is not correlated with an unobserved factor so that the MCAR assumption is satisfied. On the other hand, a positive value of $\rho$ induces violation of the MCAR assumption as observability is correlated with an unobserved factor so that it is chosen to induce bias that is not so small.
Table 2: The size and power of the test where $T = 8$ and DGP I

<table>
<thead>
<tr>
<th>point estimates($\alpha$)</th>
<th>The rate of rejecting the null hypothesis</th>
<th>variable addition test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LS$</td>
<td>$RE$</td>
<td>$FE$</td>
</tr>
<tr>
<td>$FD$</td>
<td>$W_{1,LS-FE}$</td>
<td>$W_{2,FE-FD}$</td>
</tr>
<tr>
<td>$W_{1,FD}$</td>
<td>$W_{2,FD}$</td>
<td>$W_{3,FE-FD}$</td>
</tr>
<tr>
<td>$W_{1,FD}$</td>
<td>$W_{2,FD}$</td>
<td>$W_{3,FD}$</td>
</tr>
</tbody>
</table>

40 percent of data is missing and $\rho = 0$

| $n = 100$ | 1.00 | 1.00 | 1.00 | 1.00 | .064 | .068 | .067 | .055 | .061 | .170 | .047 | .047 |
| coverage  | (.99,1.01) | (1.00,1.00) | (1.00,1.00) | (.053,0.74) | (.057,0.79) | (.056,0.77) | (.045,0.65) | (.050,0.71) | (.153,1.86) | (.037,0.56) | (.038,0.56) |
| $n = 500$ | 1.00 | 1.00 | 1.00 | 1.00 | .057 | .056 | .058 | .037 | .048 | .158 | .048 | .052 |
| coverage  | (.99,1.00) | (1.00,1.00) | (1.00,1.00) | (.047,0.67) | (.045,0.66) | (.048,0.68) | (.029,0.45) | (.038,0.57) | (.142,1.74) | (.038,0.57) | (.042,0.61) |
| $n = 1000$| 1.00 | 1.00 | 1.00 | 1.00 | .052 | .049 | .052 | .052 | .053 | .167 | .043 | .056 |
| coverage  | (1.00,1.00) | (1.00,1.00) | (1.00,1.00) | (.042,0.61) | (.039,0.58) | (.042,0.62) | (.042,0.62) | (.043,0.63) | (.151,1.83) | (.034,0.51) | (.045,0.66) |

70 percent of data is missing and $\rho = 0$

| $n = 100$ | 1.00 | 1.00 | 1.00 | 1.00 | .102 | .100 | .097 | .056 | .088 | .199 | .043 | .046 |
| coverage  | (.98,1.01) | (1.00,1.00) | (1.00,1.00) | (.089,115) | (.087,113) | (.084,110) | (.045,0.66) | (.075,100) | (.181,216) | (.034,0.51) | (.036,0.45) |
| $n = 500$ | 1.00 | 1.00 | 1.00 | 1.00 | .056 | .054 | .063 | .044 | .053 | .170 | .040 | .047 |
| coverage  | (.99,1.00) | (1.00,1.00) | (1.00,1.00) | (.046,0.66) | (.044,0.64) | (.052,0.73) | (.035,0.52) | (.043,0.63) | (.154,1.87) | (.031,0.49) | (.038,0.56) |
| $n = 1000$| 1.00 | 1.00 | 1.00 | 1.00 | .059 | .050 | .058 | .051 | .049 | .212 | .039 | .051 |
| coverage  | (0.99,1.00) | (1.00,1.00) | (1.00,1.00) | (.048,0.69) | (.040,0.59) | (.048,0.68) | (.041,0.60) | (.040,0.58) | (.194,230) | (.031,0.47) | (.041,0.60) |

40 percent of data is missing and $\rho = 0.9$

| $n = 100$ | .53 | .94 | 1.00 | 1.00 | .738 | .064 | .720 | .996 | .234 | 1 | .046 | .053 |
| coverage  | (.53,54) | (.94,95) | (1.00,1.00) | (.718,718) | (.502,0.75) | (.720,741) | (.993,999) | (.214,253) | (1.1) | (.036,0.55) | (.043,0.63) |
| $n = 500$ | .53 | .94 | 1.00 | 1.00 | 1 | .052 | 1 | 1 | .739 | 1 | .048 | .051 |
| coverage  | (.52,53) | (.94,94) | (1.00,1.00) | (.11) | (.042,0.62) | (1.1) | (.720,759) | (1.1) | (.038,0.57) | (.041,0.61) |
| $n = 1000$| .53 | .94 | 1.00 | 1.00 | 1 | .057 | 1 | 1 | .959 | 1 | .047 | .051 |
| coverage  | (.52,53) | (.94,94) | (1.00,1.00) | (.11) | (.047,0.67) | (1.1) | (.950,967) | (1.1) | (.037,0.56) | (.041,0.60) |

70 percent of data is missing and $\rho = 0.9$

| $n = 100$ | .22 | .82 | 1.00 | 1.00 | .860 | .097 | .834 | .987 | .481 | .999 | .038 | .050 |
| coverage  | (.21,24) | (.81,82) | (1.00,1.00) | (.844,876) | (.804,110) | (.817,851) | (.981,992) | (.458,504) | (.997,1) | (.029,0.46) | (.040,0.6) |
| $n = 500$ | .20 | .81 | 1.00 | 1.00 | 1 | .055 | 1 | 1 | .962 | 1 | .040 | .055 |
| coverage  | (.20,21) | (.81,81) | (1.00,1.00) | (.11) | (.045,0.66) | (1.1) | (.953,970) | (1.1) | (.031,0.49) | (.045,0.66) |
| $n = 1000$| .20 | .81 | 1.00 | 1.00 | 1 | .062 | 1 | 1 | .999 | 1 | .038 | .056 |
| coverage  | (.20,21) | (.81,81) | (1.00,1.00) | (.11) | (.051,0.73) | (1.1) | (.997,1) | (1.1) | (.030,0.46) | (.046,0.66) |

Note: The number of replications is 2,000. When $\rho = 0$ observability is not correlated with an unobserved factor so that the MCAR assumption is satisfied. On the other hand, a positive value of $\rho$ induce violation of the MCAR assumption as observability is correlated with an unobserved factor so that it is chosen to induce bias that is not so small.
Table 3: The size and power of the test where $T = 4$ and DGP II

<table>
<thead>
<tr>
<th>mean point estimates($\alpha$)</th>
<th>The rate of rejecting the null hypothesis</th>
<th>variable addition test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$W_{1,LS-FE}$</td>
<td>$W_{2,FE-FD}$</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>40 percent of data is missing and $\rho = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 percent of data is missing and $\rho = 0.9$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The number of replications is 2,000. When $\rho = 0$ observability is not correlated with an unobserved factor so that the MCAR assumption is satisfied. On the other hand, a positive value of $\rho$ induce violation of the MCAR assumption as observability is correlated with an unobserved factor so that it is chosen to induce bias that is not so small.
Table 4: The size and power of the test where $T = 8$ and DGP II

<table>
<thead>
<tr>
<th>point estimates ($\alpha$)</th>
<th>The rate of rejecting the null hypothesis</th>
<th>variable addition test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$LS$ $RE$ $FE$ $FD$ $W_{1,LS-RE}$ $W_{2,RE-FD}$ $W_{3,LS-FD}$ $W_{4,FE}$ $W_{5,FE-FD}$ pooled $FD$ $FE$</td>
<td></td>
</tr>
<tr>
<td>40 percent of data is missing and $\rho = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99,1.01)</td>
<td>(1.00,1.01)</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
</tr>
<tr>
<td>70 percent of data is missing and $\rho = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99,1.01)</td>
<td>(1.00,1.01)</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00,1.02)</td>
<td>(1.00,1.01)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
</tr>
<tr>
<td>40 percent of data is missing and $\rho = 0.9$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 100$</td>
<td>.66</td>
<td>.67</td>
</tr>
<tr>
<td>coverage</td>
<td>(.65,0.67)</td>
<td>(.66,0.67)</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>.66</td>
<td>.67</td>
</tr>
<tr>
<td>coverage</td>
<td>(.66,0.66)</td>
<td>(.67,0.67)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>.66</td>
<td>.67</td>
</tr>
<tr>
<td>coverage</td>
<td>(.66,0.66)</td>
<td>(.67,0.67)</td>
</tr>
<tr>
<td>70 percent of data is missing and $\rho = 0.9$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 100$</td>
<td>.39</td>
<td>.39</td>
</tr>
<tr>
<td>coverage</td>
<td>(.38,0.40)</td>
<td>(.38,0.40)</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>.39</td>
<td>.40</td>
</tr>
<tr>
<td>coverage</td>
<td>(.39,0.39)</td>
<td>(.39,0.40)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>.39</td>
<td>.39</td>
</tr>
<tr>
<td>coverage</td>
<td>(.38,0.39)</td>
<td>(.39,0.40)</td>
</tr>
</tbody>
</table>

Note: The number of replications is 2,000. When $\rho = 0$ observability is not correlated with an unobserved factor so that the MCAR assumption is satisfied. On the other hand, a positive value of $\rho$ induce violation of the MCAR assumption as observability is correlated with an unobserved factor so that it is chosen to induce bias that is not so small.
Table 5: The size of HT tests of the MAR assumption using DGP III and DGP IV

<table>
<thead>
<tr>
<th>mean</th>
<th>The rate of rejecting the null hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>point estimates(β)</td>
</tr>
<tr>
<td></td>
<td>IPW - LS</td>
</tr>
<tr>
<td>T = 4 and 36 percent of data is missing</td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>0.98</td>
</tr>
<tr>
<td>coverage</td>
<td>(.98, .99)</td>
</tr>
<tr>
<td>n = 200</td>
<td>0.99</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99, 1.00)</td>
</tr>
<tr>
<td>n = 400</td>
<td>0.99</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99, 1.00)</td>
</tr>
<tr>
<td>n = 1000</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>T = 8 and 36 percent of data is missing</td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 200</td>
<td>0.99</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99, 1.00)</td>
</tr>
<tr>
<td>n = 400</td>
<td>0.99</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99, 1.00)</td>
</tr>
<tr>
<td>n = 1000</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>T = 16 and 36 percent of data is missing</td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 200</td>
<td>0.99</td>
</tr>
<tr>
<td>coverage</td>
<td>(.99, 1.00)</td>
</tr>
<tr>
<td>n = 400</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 1000</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>T = 4 and 36 percent of data is missing</td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 200</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 400</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>T = 8 and 36 percent of data is missing</td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 200</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 400</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>T = 16 and 36 percent of data is missing</td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 200</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>n = 400</td>
<td>1.00</td>
</tr>
<tr>
<td>coverage</td>
<td>(1.00, 1.00)</td>
</tr>
</tbody>
</table>

Note: The number of replications is 2,000. The 95% coverage of the mean values is reported in parenthesis.
<table>
<thead>
<tr>
<th>$n$</th>
<th>PW−LS</th>
<th>PW−FE</th>
<th>PW−FD</th>
<th>$W_{1,IPW−LS,IPW−FE}$</th>
<th>$W_{2,IPW−LS,IPW−FD}$</th>
<th>$W_{3,IPW−FE,IPW−FD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>.52</td>
<td>.41</td>
<td>.058</td>
</tr>
<tr>
<td>coverage (1.17,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(.50,55)</td>
<td>(.39,43)</td>
<td>(.048,.068)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>.81</td>
<td>.68</td>
<td>.049</td>
</tr>
<tr>
<td>coverage (1.18,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(.80,83)</td>
<td>(.66,70)</td>
<td>(.039,.058)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.94</td>
<td>.054</td>
</tr>
<tr>
<td>coverage (1.18,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(.98,99)</td>
<td>(.93,95)</td>
<td>(.044,.063)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>.99</td>
<td>.053</td>
</tr>
<tr>
<td>coverage (1.18,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(1,1)</td>
<td>(.99,1)</td>
<td>(.043,.063)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>PW−LS</th>
<th>PW−FE</th>
<th>PW−FD</th>
<th>$W_{1,IPW−LS,IPW−FE}$</th>
<th>$W_{2,IPW−LS,IPW−FD}$</th>
<th>$W_{3,IPW−FE,IPW−FD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>.81</td>
<td>.66</td>
<td>.049</td>
</tr>
<tr>
<td>coverage (1.17,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(.80,83)</td>
<td>(.64,68)</td>
<td>(.040,.058)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>.98</td>
<td>.91</td>
<td>.057</td>
</tr>
<tr>
<td>coverage (1.18,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(.98,99)</td>
<td>(.90,92)</td>
<td>(.047,.067)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
<td>.051</td>
</tr>
<tr>
<td>coverage (1.18,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(1,1)</td>
<td>(.99,1)</td>
<td>(.041,.061)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>1.18</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
<td>.069</td>
</tr>
<tr>
<td>coverage (1.18,1.18)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(.050,.070)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>PW−LS</th>
<th>PW−FE</th>
<th>PW−FD</th>
<th>$W_{1,IPW−LS,IPW−FE}$</th>
<th>$W_{2,IPW−LS,IPW−FD}$</th>
<th>$W_{3,IPW−FE,IPW−FD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.20</td>
<td>1.00</td>
<td>1.00</td>
<td>.94</td>
<td>.86</td>
<td>.056</td>
</tr>
<tr>
<td>coverage (1.20,1.20)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(.93,95)</td>
<td>(.84,87)</td>
<td>(.046,.066)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.20</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
<td>.047</td>
</tr>
<tr>
<td>coverage (1.20,1.20)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(1,1)</td>
<td>(.99,99)</td>
<td>(.037,.056)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>1.20</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
<td>.045</td>
</tr>
<tr>
<td>coverage (1.20,1.20)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(.035,.054)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>1.20</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
<td>.053</td>
</tr>
<tr>
<td>coverage (1.20,1.20)</td>
<td>(1.00,1.00)</td>
<td>(1.00,1.00)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(.043,.062)</td>
<td></td>
</tr>
</tbody>
</table>

Note: The number of replications is 2,000. The 95% coverage of the mean values is reported in parenthesis. The powers of HT tests from the simulations with a bias of 60% for the LS estimator are 1 for all $T$ from 4 to 16 and $n$ from 100 to 600 with missing proportion of about 35%.
Table 7: The power of HT tests of the MAR assumption using DGP III-C and DGP IV-C

<table>
<thead>
<tr>
<th>mean</th>
<th>The rate of rejecting the null hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>point estimates((\hat{\beta}))</td>
</tr>
<tr>
<td></td>
<td>(IPW - FE)</td>
</tr>
</tbody>
</table>

**\(T = 4\) and 60 percent of data is missing**

<table>
<thead>
<tr>
<th>(n)</th>
<th>()</th>
<th>()</th>
<th>()</th>
<th>()</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.56</td>
<td>.71</td>
<td>.35</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.55,.56)</td>
<td>(.71,.72)</td>
<td>(.33,.37)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>.56</td>
<td>.71</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.56,.56)</td>
<td>(.70,.71)</td>
<td>(.44,.48)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>.56</td>
<td>.70</td>
<td>.62</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.56,.57)</td>
<td>(.70,.70)</td>
<td>(.60,.64)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>.56</td>
<td>.70</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.56,.56)</td>
<td>(.69,.70)</td>
<td>(.67,.71)</td>
<td></td>
</tr>
</tbody>
</table>

**\(T = 8\) and 46 percent of data is missing**

<table>
<thead>
<tr>
<th>(n)</th>
<th>()</th>
<th>()</th>
<th>()</th>
<th>()</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.72</td>
<td>.83</td>
<td>.43</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.72,.72)</td>
<td>(.83,.84)</td>
<td>(.41,.45)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>.72</td>
<td>.83</td>
<td>.66</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.72,.72)</td>
<td>(.83,.84)</td>
<td>(.63,.68)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>.72</td>
<td>.83</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.72,.72)</td>
<td>(.83,.84)</td>
<td>(.85,.88)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>.72</td>
<td>.83</td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.72,.72)</td>
<td>(.83,.83)</td>
<td>(.93,.96)</td>
<td></td>
</tr>
</tbody>
</table>

**\(T = 16\) and 35 percent of data is missing**

<table>
<thead>
<tr>
<th>(n)</th>
<th>()</th>
<th>()</th>
<th>()</th>
<th>()</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.87</td>
<td>.94</td>
<td>.48</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.87,.87)</td>
<td>(.93,.94)</td>
<td>(.46,.50)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>.87</td>
<td>.93</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.87,.87)</td>
<td>(.93,.94)</td>
<td>(.75,.79)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>.87</td>
<td>.93</td>
<td>.96</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.87,.87)</td>
<td>(.93,.94)</td>
<td>(.95,.97)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>.87</td>
<td>.93</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(.87,.87)</td>
<td>(.93,.94)</td>
<td>(1,1)</td>
<td></td>
</tr>
</tbody>
</table>

**\(T = 4\) and 32 percent of data is missing**

<table>
<thead>
<tr>
<th>(n)</th>
<th>()</th>
<th>()</th>
<th>()</th>
<th>()</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-.75</td>
<td>-1.27</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-.77,-.72)</td>
<td>(-1.29,-1.25)</td>
<td>(.32,.36)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>-.75</td>
<td>-1.27</td>
<td>.58</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-.77,-.74)</td>
<td>(-1.29,-1.26)</td>
<td>(.56,.60)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>-.74</td>
<td>-1.27</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-.76,-.73)</td>
<td>(-1.28,-1.25)</td>
<td>(.85,.88)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>-.75</td>
<td>-1.27</td>
<td>.96</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-.76,-.74)</td>
<td>(-1.28,-1.26)</td>
<td>(.95,.97)</td>
<td></td>
</tr>
</tbody>
</table>

**\(T = 8\) and 36 percent of data is missing**

<table>
<thead>
<tr>
<th>(n)</th>
<th>()</th>
<th>()</th>
<th>()</th>
<th>()</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-1.57</td>
<td>-2.76</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-1.61,-1.54)</td>
<td>(-2.80,-2.72)</td>
<td>(.48,.52)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>-1.56</td>
<td>-2.73</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-1.58,-1.53)</td>
<td>(-2.76,-2.71)</td>
<td>(.75,.79)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>-1.58</td>
<td>-2.74</td>
<td>.97</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-1.59,-1.56)</td>
<td>(-2.75,-2.72)</td>
<td>(.96,.98)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>-1.56</td>
<td>-2.72</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-1.57,-1.54)</td>
<td>(-2.74,-2.71)</td>
<td>(1,1)</td>
<td></td>
</tr>
</tbody>
</table>

**\(T = 16\) and 40 percent of data is missing**

<table>
<thead>
<tr>
<th>(n)</th>
<th>()</th>
<th>()</th>
<th>()</th>
<th>()</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-2.71</td>
<td>-5.03</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-2.77,-2.65)</td>
<td>(-5.11,-4.96)</td>
<td>(.48,.53)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>-2.70</td>
<td>-5.02</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-2.74,-2.65)</td>
<td>(-5.07,-4.97)</td>
<td>(.76,.80)</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>-2.72</td>
<td>-5.02</td>
<td>.97</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-2.75,-2.69)</td>
<td>(-5.06,-4.99)</td>
<td>(.97,.98)</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>-2.71</td>
<td>-5.00</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>(-2.73,-2.68)</td>
<td>(-5.03,-4.97)</td>
<td>(1,1)</td>
<td></td>
</tr>
</tbody>
</table>

Note: The number of replications is 2,000. The 95% coverage of the mean values is reported in parenthesis.
Table 8: Description of missing data for trade flows

<table>
<thead>
<tr>
<th></th>
<th>available y</th>
<th>missing y</th>
<th>all possible cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive y</td>
<td>y = 0</td>
<td>y ≥ 0</td>
</tr>
<tr>
<td>num of obs</td>
<td>613,579</td>
<td>760,041</td>
<td>1,373,620</td>
</tr>
<tr>
<td>% of total</td>
<td>24%</td>
<td>29%</td>
<td>53%</td>
</tr>
<tr>
<td>full sample of y</td>
<td>1948-2003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>num of obs</td>
<td>495,084</td>
<td>410,111</td>
<td>905,195</td>
</tr>
<tr>
<td>% of total</td>
<td>31%</td>
<td>26%</td>
<td>57%</td>
</tr>
<tr>
<td>restricted sample of y</td>
<td>1970-2003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: WTO/RTA/CU membership effects on Trade - Test of the MCAR

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>FD</td>
<td>HT-Test</td>
<td>FE</td>
<td>FD</td>
<td>HT-Test</td>
</tr>
<tr>
<td>WTO</td>
<td>1.10***</td>
<td>0.22**</td>
<td>335.48***</td>
<td>0.002</td>
<td>-0.05</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>[0.00]</td>
<td>(0.06)</td>
<td>(0.05)</td>
<td>[0.49]</td>
</tr>
<tr>
<td>RTA</td>
<td>0.71***</td>
<td>0.22**</td>
<td>41.85***</td>
<td>0.64***</td>
<td>0.12**</td>
<td>34.04***</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>[0.00]</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>[0.00]</td>
</tr>
<tr>
<td>CU</td>
<td>3.38***</td>
<td>4.34**</td>
<td>5.68***</td>
<td>2.27***</td>
<td>3.70***</td>
<td>20.42***</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.30)</td>
<td>[0.02]</td>
<td>(0.28)</td>
<td>(0.28)</td>
<td>[0.00]</td>
</tr>
</tbody>
</table>

|                  | TFE     | UCPH    | MRTs    | Yes     | Yes     | Yes     |
|                  | No      | No      | No      | Yes     | Yes     | Yes     |

<table>
<thead>
<tr>
<th></th>
<th>No. of obs</th>
<th>sample year</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTO</td>
<td>1,184,525</td>
<td>1948-2003</td>
</tr>
<tr>
<td>RTA</td>
<td>1,147,781</td>
<td>1,147,781</td>
</tr>
<tr>
<td>CU</td>
<td>1,147,781</td>
<td>1,147,781</td>
</tr>
</tbody>
</table>

Notes: TFE, UCPH, and MRTs denote time fixed effects, unobserved country-pair heterogeneity and BB approximation, respectively. Cluster (country-pairs) robust standard errors are reported in parentheses. ***, ** and * indicate that the coefficient is statistically significant at 1%, 5% and 10% levels, respectively. RTA represents regional trade agreements which include free trade agreements and customs union. In column (4), HT test statistics are reported for each covariate and p-values are reported in brackets. The number of observations for y is 1,373,620 so 189,095 (14%) observations are missing for covariates, mostly GDP.
Table 10: WTO/RTA/CU trade effects - Test of the MAR

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
<th>(11)</th>
<th>(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>FD</td>
<td>HT-Test</td>
<td>FE</td>
<td>FD</td>
<td>HT-Test</td>
<td>FE</td>
<td>FD</td>
<td>HT-Test</td>
<td>FE</td>
<td>FD</td>
<td>HT-Test</td>
</tr>
<tr>
<td><strong>WTO</strong></td>
<td>0.21***</td>
<td>-0.05</td>
<td>11.88***</td>
<td>0.28***</td>
<td>-0.02</td>
<td>11.11***</td>
<td>0.20**</td>
<td>0.02</td>
<td>2.90*</td>
<td>1.66***</td>
<td>0.44***</td>
<td>568.05***</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>[0.00]</td>
<td>(0.07)</td>
<td>(0.03)</td>
<td>[0.00]</td>
<td>(0.09)</td>
<td>(0.08)</td>
<td>[0.09]</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>[0.00]</td>
</tr>
<tr>
<td><strong>RTA</strong></td>
<td>0.18**</td>
<td>0.12***</td>
<td>0.39</td>
<td>0.28***</td>
<td>0.15**</td>
<td>2.44</td>
<td>0.39***</td>
<td>0.10</td>
<td>7.03***</td>
<td>1.43***</td>
<td>0.72***</td>
<td>89.96***</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>[0.53]</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>[0.12]</td>
<td>(0.08)</td>
<td>(0.06)</td>
<td>[0.01]</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>[0.00]</td>
</tr>
<tr>
<td><strong>CU</strong></td>
<td>2.44***</td>
<td>3.79***</td>
<td>8.16***</td>
<td>2.13***</td>
<td>3.63***</td>
<td>10.27</td>
<td>1.14***</td>
<td>3.18***</td>
<td>11.96***</td>
<td>3.23***</td>
<td>3.25***</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.36)</td>
<td>(0.27)</td>
<td>[0.004]</td>
<td>(0.36)</td>
<td>(0.29)</td>
<td>[0.02]</td>
<td>(0.30)</td>
<td>(0.38)</td>
<td>[0.00]</td>
<td>(0.35)</td>
<td>(0.39)</td>
<td>[0.95]</td>
</tr>
<tr>
<td><strong>TFE</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>UCPH</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>MRTs</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>MAR</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>IMR</td>
<td>IMR</td>
<td>IMR</td>
<td>IPW</td>
<td>IPW</td>
<td>IPW</td>
<td>Imputation</td>
<td>Imputation</td>
<td>Imputation</td>
</tr>
<tr>
<td><strong>No. of obs.</strong></td>
<td>835,726</td>
<td>800,457</td>
<td>800,457</td>
<td>835,726</td>
<td>800,457</td>
<td>800,457</td>
<td>835,726</td>
<td>800,457</td>
<td>800,457</td>
<td>1,578,960</td>
<td>1,532,520</td>
<td>1,532,520</td>
</tr>
</tbody>
</table>

Notes: TFE, UCPH, and MRTs denote time fixed effects, unobserved country-pair heterogeneity and BB approximation, respectively. Cluster (country-pairs) robust standard errors are reported in parentheses. ***, ** and * indicate that the coefficient is statistically significant at 1%, 5% and 10% levels, respectively. RTA represents regional trade agreements which include free trade agreements and customs union. In column (4), HT test statistics are reported for each covariate and p-values are reported in brackets. The number of observations for y is 905,195 so 69,469 (8%) observations are missing for covariates, mostly GDP.
A Proofs of Asymptotic Results

A.1 Proof of Theorem 2

We start with rewriting the stacked estimator to obtain asymptotic variance. We can rewrite \( \sqrt{N}(\hat{\theta}_1 - \theta_1) \) as follows.

\[
\hat{\theta}_1 = \left[ \begin{array}{cc} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' x_{it} & 0 \\ 0 & \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{x}_{it} \end{array} \right]^{-1} \left[ \begin{array}{c} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' y_{it} \\ \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{y}_{it} \end{array} \right]_{2k \times 2k}
\]

\[
\sqrt{N}(\hat{\theta}_1 - \theta_1) = \left[ \begin{array}{cc} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' x_{it} & 0 \\ 0 & \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{x}_{it} \end{array} \right]^{-1} \left[ \begin{array}{c} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' u_{it} \\ \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{u}_{it} \end{array} \right]_{2k \times 2k}
\]

\[
\sqrt{N}(\hat{\theta}_1 - \theta_1) = \left[ \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' x_{it} \right]^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' u_{it} \right) = \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{x}_{it} \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{u}_{it} \right)
\]

where \( \hat{\theta}_1 \) is a consistent estimator for \( \theta_1 \) under the null.

Firstly, we will show \( \hat{\theta}_1 \) is a consistent estimator for \( \theta_1 \) by showing \( (\hat{\theta}_1 - \theta_1) \overset{p}{\rightarrow} 0. \)

Let’s denote \( \hat{A}_{11} \) and \( \hat{A}_{22} \) as follows.

\[
\hat{A}_{11} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' x_{it}, \quad \hat{A}_{22} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{x}_{it}.
\]

Then, using the Assumption 2.3.2, we can obtain the following well defined inverses.

\[
(p \lim \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}' x_{it})^{-1} = E(\sum_{t=1}^{T} s_{it} x_{it}' x_{it})^{-1} \equiv A_{11}^{-1}
\]

\[
(p \lim \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{x}_{it})^{-1} = E(\sum_{t=1}^{T} s_{it} \tilde{x}_{it}' \tilde{x}_{it})^{-1} \equiv A_{22}^{-1}.
\]

We could also denote as

\[
\hat{A}_{11}^{-1} \overset{p}{\rightarrow} A_{11}^{-1}, \quad \hat{A}_{22}^{-1} \overset{p}{\rightarrow} A_{22}^{-1}
\]

where \( \overset{p}{\rightarrow} \) denotes convergence in probability.

Then, the consistency result can be obtained as follows.
\[ \text{plim}(\hat{\theta}_1 - \theta_1) = \left( \text{plim} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}x_{it}'x_{it} \right)^{-1} \left( \text{plim} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}x_{it}'v_{it} \right) = \left( A_{11}^{-1}(E \sum_{t=1}^{T} (s_{it}x_{it}'u_{it})) \right) \]
\[ \left( A_{22}^{-1}(E \sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it}) \right) \]
\[ = \left( O_p(1) \cdot o_p(1) = 0 \right) \]

We get \( A_{11}^{-1}(E \sum_{t=1}^{T} (s_{it}x_{it}'u_{it})) = 0 \) and \( A_{22}^{-1}(E \sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it}) = 0 \) because the inverses are well-defined by the Assumption 2.3.2 and \( E(v_{it}|s_{it},x_{it}) = 0 \) is sufficient for both \( E(\sum_{t=1}^{T} (s_{it}x_{it}'u_{it})) \) and \( E(\sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it}) \) are zeros by the Assumption 2.3.1.

Now, we want to show that \( \sqrt{N}(\hat{\theta}_1 - \theta_1) \) converges to a normal density with well-defined asymptotic covariance as in (54).

\[ \sqrt{N}(\hat{\theta}_1 - \theta_1) \Rightarrow N \left( 0, \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} \right) \]

where \( \Rightarrow \) denotes weak convergence (i.e. convergence in distribution).

We use the following notations for the asymptotic variance estimator as in (36):

\[ \hat{G}_1^{-1} = \begin{bmatrix} \hat{A}_{11} & 0 \\ 0 & \hat{A}_{22} \end{bmatrix}^{-1}, \hat{D}_1 = \begin{bmatrix} \hat{D}_{11} & \hat{D}_{12} \\ \hat{D}_{21} & \hat{D}_{22} \end{bmatrix} \]

\[ \hat{D}_{11} = \frac{1}{N} \sum_{i=1}^{N} (T \sum_{t=1}^{T} s_{it}x_{it}'\hat{v}_{it})(T \sum_{t=1}^{T} s_{it}x_{it}'\hat{v}_{it})', \hat{D}_{12} = \frac{1}{N} \sum_{i=1}^{N} (T \sum_{t=1}^{T} s_{it}x_{it}'\hat{v}_{it})(T \sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it})', \hat{D}_{21} = \frac{1}{N} \sum_{i=1}^{N} (T \sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it})(T \sum_{t=1}^{T} s_{it}x_{it}'\hat{v}_{it})', \hat{D}_{22} = \frac{1}{N} \sum_{i=1}^{N} (T \sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it})(T \sum_{t=1}^{T} s_{it}x_{it}'\tilde{u}_{it})' \]

where \( \tilde{u}_{it} \) is residual from FE estimation and \( \hat{v}_{it} \) residual from pooled LS estimation. Then, we can denote

\[ \hat{\text{var}}(\sqrt{N}(\hat{\theta}_1 - \theta_1)) = \begin{bmatrix} \hat{A}_{11} & 0 \\ 0 & \hat{A}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \hat{D}_{11} & \hat{D}_{12} \\ \hat{D}_{21} & \hat{D}_{22} \end{bmatrix} \begin{bmatrix} \hat{A}_{11} & 0 \\ 0 & \hat{A}_{22} \end{bmatrix}^{-1} = \hat{G}_1^{-1}\hat{D}_1\hat{G}_1^{-1} \]

where

\[ \hat{G}_1^{-1} = \begin{bmatrix} \hat{A}_{11} & 0 \\ 0 & \hat{A}_{22} \end{bmatrix}^{-1}, \hat{D}_1 = \begin{bmatrix} \hat{D}_{11} & \hat{D}_{12} \\ \hat{D}_{21} & \hat{D}_{22} \end{bmatrix}. \]
We further assume that \( \lim p(\hat{D}_1) = D_1 \), then we have the following.

\[
\hat{D}_{11} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{t=1}^{T} s_{it} x_i' t \hat{v}_{it})(\sum_{t=1}^{T} s_{it} x_i' t \hat{v}_{it})' \xrightarrow{p} E[(\sum_{t=1}^{T} s_{it} x_i' t v_{it})(\sum_{t=1}^{T} s_{it} x_i' t v_{it})'] = D_{11}
\]

\[
\hat{D}_{12} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{t=1}^{T} s_{it} x_i' t \hat{v}_{it})(\sum_{t=1}^{T} s_{it} x_i' t \hat{u}_{it})' \xrightarrow{p} E[(\sum_{t=1}^{T} s_{it} x_i' t v_{it})(\sum_{t=1}^{T} s_{it} x_i' t u_{it})'] = D_{12}
\]

\[
\hat{D}_{21} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{t=1}^{T} s_{it} x_i' t \hat{u}_{it})(\sum_{t=1}^{T} s_{it} x_i' t \hat{v}_{it})' \xrightarrow{p} E[(\sum_{t=1}^{T} s_{it} x_i' t u_{it})(\sum_{t=1}^{T} s_{it} x_i' t v_{it})'] = D_{21}
\]

\[
\hat{D}_{22} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{t=1}^{T} s_{it} x_i' t \hat{u}_{it})(\sum_{t=1}^{T} s_{it} x_i' t \hat{u}_{it})' \xrightarrow{p} E[(\sum_{t=1}^{T} s_{it} x_i' t u_{it})(\sum_{t=1}^{T} s_{it} x_i' t u_{it})'] = D_{22}.
\]

Using regularity conditions, it could be shown that

\[
\overline{\text{var}}(\sqrt{N}(\hat{\theta}_1 - \theta_1)) = \hat{\sigma}_1^{-1} \hat{D}_1 \hat{\sigma}_1^{-1} = G_1^{-1} D_1 G_1^{-1} + o_p(1), \text{ (i.e. } \hat{\sigma}_1^{-1} \hat{D}_1 \hat{\sigma}_1^{-1} \xrightarrow{p} G_1^{-1} D_1 G_1^{-1},)\]

so we have the normality result as follows.

\[
\sqrt{N}(\hat{\theta}_1 - \theta_1) \Rightarrow N \left( 0, \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} \equiv G_1^{-1} D_1 G_1^{-1} \right)
\]

Furthermore, using the following matrix operation, we can simplify the calculation for variance covariance matrix.

\[
\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix}
\]

\[
= \begin{bmatrix} A_{11}^{-1} D_{11} A_{11}^{-1} & A_{11}^{-1} D_{12} A_{22}^{-1} \\ A_{22}^{-1} D_{21} A_{11}^{-1} & A_{22}^{-1} D_{22} A_{22}^{-1} \end{bmatrix}
\]

Therefore, using \( \sqrt{N} \)-normal convergence result and restriction matrix \( R \), we have

\[
R \sqrt{N}(\hat{\theta}_1 - \theta_1) \Rightarrow N \left( 0, \begin{bmatrix} A_{11}^{-1} D_{11} A_{11}^{-1} & A_{11}^{-1} D_{12} A_{22}^{-1} \\ A_{22}^{-1} D_{21} A_{11}^{-1} & A_{22}^{-1} D_{22} A_{22}^{-1} \end{bmatrix} R' \right)
\] (37)
\[ R \sqrt{N} (\hat{\theta}_1 - \theta_1) \overset{d}{\sim} N(0, R G_1^{-1} D_1 G_1^{-1} R') \]

We can obtain a Wald statistic as in (57):

\[ \hat{W}_1 = [R \sqrt{N} (\hat{\theta}_1 - \theta_1)]' [R \hat{G}_1^{-1} \hat{D}_1 \hat{G}_1^{-1} R']^{-1} R \sqrt{N} (\hat{\theta}_1 - \theta_1) \]

which converges to chi-squared distribution under the null hypothesis as follows.

\[ [R \sqrt{N} \hat{\theta}_1]' [R \hat{G}_1^{-1} \hat{D}_1 \hat{G}_1^{-1} R']^{-1} R \sqrt{N} \hat{\theta}_1 \sim \chi^2_q \]

A.2 Proof of Lemma 4

By Assumptions 2.1 and 3.2 and the Law of Iterated Expectation, we have

\[ \text{plim} \hat{\theta}_{1PW-pls} = \theta + (\text{plim} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}^' x_{it})^{-1} \text{plim} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} x_{it}^' v_{it} \]
\[ = \theta + E \{ E(\sum_{t=1}^{T} s_{it} x_{it}^' x_{it} | \mathbf{w}_{it}) \}^{-1} E \{ E(\sum_{t=1}^{T} s_{it} x_{it}^' v_{it} | \mathbf{w}_{it}) \} \]
\[ = \theta + E \{ E(\sum_{t=1}^{T} f(\mathbf{z}_{it}) x_{it}^' x_{it} | \mathbf{z}_{it}, \gamma) \}^{-1} E \{ E(\sum_{t=1}^{T} f(\mathbf{z}_{it}) x_{it}^' v_{it} | \mathbf{z}_{it}, \gamma) \} \]
\[ = \theta + \sum_{t=1}^{T} E(\mathbf{x}_{it}^' \mathbf{x}_{it})^{-1} \sum_{t=1}^{T} E(\mathbf{x}_{it}^' v_{it}) = \theta. \]

B Additional Monte Carlo Experiments

We investigate finite-sample performances of the proposed test in a linear panel data model where both response variable and explanatory variables are continuous:

\[ y_{it} = x_{it} \cdot \beta + a \cdot c_{i1} + b \cdot u_{it}, \text{ for } i = 1, 2, ..., n \text{ and } t = 1, 2, ..., T \]

where \( x_{it} \) is a continuous variable and the interest is in \( \beta \). Selection processes reduce full samples to sample sizes ranging from 40 to 70 percent. We consider observations from \( N = 100 \) to \( N = 1,000 \) for the cross-section and from \( T = 4 \) to \( T = 8 \) for the time dimension. \( s_{i1} = 1 \) and \( s_{it} = 1(c_p + z_{it} \cdot \gamma + c \cdot c_{2t} + d \cdot v_{it} > 56) \]
0), \( \forall t \geq 2 \) where \( c_p \) is chosen to make missing fraction be between 40 and 60 percent.

First, we consider a case where the violation of the MCAR assumption is induced by the correlation between time-invariant unobserved factor and \( x_{it} \).

**DGP B-I**

1. \( c_{1i} = i.i.d. N(0,1) + 5 \cdot w_{1i}; \ a = 10; \ c_{2i} = 2 \cdot q_i + w_{2i}; \ c = 1; \ q_i \sim i.i.d. N(0,1). \)

2. \( (w_{1i}, w_{2i}) \sim B N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \).

3. \( i t = 10 \cdot q_i + i.i.d. N(0,100); \ \beta = 1. \)

4. \( z_{it} \sim i.i.d. N(0,1); \ \gamma = 0.1. \)

5. \( u_{it} \sim i.i.d. N(0,1); \ b = 1; \ v_{it} \sim i.i.d. N(0,1); \ d = 0.1. \)

The violation of the MCAR assumption is induced by the correlation between time-invariant unobserved factor \( (w_{1i}) \) and \( x_{it} \) conditional on \( s_{it} = 1 \). The correlation is indirect.

Second, we consider a case where the violation of the MCAR assumption is induced by the correlation between time-varying unobserved factor and \( x_{it} \).

**DGP B-II**

1. \( c_{1i} = i.i.d. N(0,1); \ a = 1; \ c_{2i} = i.i.d. N(0,1); \ c = 1. \)

2. \( x_{it} = 10 \cdot z_{it} + i.i.d. N(0,25), \ \beta = 1. \)

3. \( z_{it} \sim i.i.d. N(0,1), \ \gamma = 1. \)

4. \( (u_{it}, v_{it}) \sim B N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right); \ b = 5; \ d = 1. \)

The violation of the MCAR assumption is induced by the correlation between time-varying unobserved factor \( (u_{it}) \) and \( x_{it} \). The correlation is indirect.

The qualitative implications from the simulations are very similar to the case for a binary covariate. Thus we omit the results for brevity.\(^{26}\)

\(^{26}\)Detailed test results for the experiments are available from the authors upon request.
C Implementing the HT tests in Stata

C.1 HT test of the MCAR test

An HT test of the MCAR assumption is implemented by the following four steps in Stata. First, for unbalanced panel data, define cross-section and time identifiers (i) and (j) by typing 'xtset i j'. Second, provide dependent variable (y), independent variables (x), and independent variables of primary interest (z \( \subseteq x \)) by typing 'y x_1 x_2 ... x_k z_1 z_2 ... z_m'. Third, identify two panel data estimators such as pooled OLS and fixed effects by typing 'pols fe'. Fourth, if more than one alternative hypothesis is considered, one can repeat the first, second and third steps with another two panel data estimators such as fixed effects (FE) and first differencing (FD). Moreover, this command for the test of the MCAR assumption is written in a very flexible way so that it also allows comparing two consistent estimators under the null while at least one of them is inconsistent under alternative.

C.2 HT test of the MAR test

The HT tests based on two-step approaches with IMR and the IPW methods are examined together because their procedures for implementation are very close to each other. An HT test of the MAR assumption is implemented by the following steps in Stata. Probability of an observation for unit \( i \) and time \( t \) is first estimated by the Probit model using all covariates in an outcome equation and exclusion restrictions (\( w_1 ... w_l \)). Probability of observing unit \( i \) at \( t \) is estimated as predicted value, \( \hat{s}_{it} = 1(w_{it} \hat{\gamma}) \), where \( w = (x_1 x_2 ... x_k w_1 ... w_l) \). Then, \( \hat{s}_{it} \) will be obtained, so, using inversely weighted dependent and explanatory variables, the outcome equation is estimated for the test using the IPW estimators and \( \hat{s}_{it} \) and \( w_{it} \hat{\gamma} \) can be used to generate the IPW and IMR terms. Therefore, implementation can be achieved by following these steps. First, define the cross-section and time identifiers. Second, define the dependent variable, independent variables and independent variable of primary interest in the outcome equation and exclusion restrictions for the first stage estimation. Third, define two panel data estimators that are consistent under the null, but where at least one of which is inconsistent under the alternative.

---

27 Although we restrict our attention to the tests of the MCAR and MAR assumptions in the paper, our proposed Stata code can be applied to comparing two estimates from any panel data model where two consistent estimators are available under the null. For instance, we can use the code to test whether non-compliance of the control and treatment is random. In this case, the pooled OLS estimator and the pooled IV estimator, for which initial assignment is used as an IV for actual assignment, can be used to construct a test statistic. Under the null of random non-compliance, both the pooled OLS and pooled IV are consistent. But only the pooled IV is consistent under the alternative of non-random non-compliance. Therefore, our proposed tests can be very flexibly used beyond testing random missing.

28 \( k \geq m \) and in practice typically \( m = 1 \) as in our application.

29 If \( w \) changes over time, the Probit estimation can be done by each \( t \).
An HT test of the MAR assumption with the MI method starts with an estimation of missing observations. Missing observations can usually occur in a dependent variable or covariates (or a small number of covariates). Practitioners need to model missing processes for all variables with missing observations. In our empirical example, missing observations occur in a dependent variable and one of covariates. Models for missing processes of missing variables could be different for each variable, so they should be modeled accordingly. Benchmark imputation is performed using the linear projection for continuous variables and the Probit model for binary variables using all available variables. Missing observations for a dependent variable and a covariate are imputed multiple times and the estimates are obtained from using the formula in Section 3.1. Finally, using both original and imputed observations, the procedure for the test of the MCAR assumption could be performed.

C.3 Empirical example

Using the same data in Section 5, we now illustrate how to implement the HT tests of the MCAR assumption and of the MAR assumption with imputation. As an illustration of the method, we reduce the size of data and restrict sample periods to 1996-2000 and the equation (39) is estimated by the FE and FD estimators:

\[
\ln(TF_{ijt}) = \alpha_0 + \alpha_1 \ln Y_{it} + \alpha_2 \ln Y_{jt} + \alpha_3 \ln y_{it} + \alpha_4 \ln y_{jt} + \\
\gamma_1 W T O_{ijt} + \gamma_2 R T A_{ijt} + \gamma_3 C U_{ijt} + \omega_{ij} + \epsilon_{ijt}. \tag{39}
\]

First, we test the validity of the MCAR assumption on the coefficients \( \gamma_1 \), \( \gamma_2 \), and \( \gamma_3 \) using the FE and FD estimations. The steps are as follows:

1. Load data in Stata: use "E:\Liu_HT_test.dta", clear
2. Define cross-section (e.g. pair_id) and time (e.g. year) indexes: xtset pair_id year
3. Type in HT test command: htfefd “dependent variable” “covariates of interests” (“other remaining observed covariates”).

In step (3), if one wants to use LS and FE estimators instead, he/she can replace 'htfefd' with 'htlsfe'. And standard errors used in the test are clustered at the cross-section unit.

Second, similarly, we can test the validity of the MAR assumption with imputation using the tests on the coefficients \( \gamma_1 \), \( \gamma_2 \), and \( \gamma_3 \) which are obtained from the MI-FE and MI-FD estimations. The first two
steps are the same as in the test of the MCAR assumption. However, we replace (3) with (3)’:

(3)’ Type in HT test command: htfefd2_mi “dependent variable” “covariates of interests” (“covariates with missing observations” | “other remaining covariates and exclusion restrictions”).

Likewise, if one wants to use MI-LS and MI-FE estimators instead, all that we need is replacing 'htfefd2_mi' with 'htlsfe_mi' in (3)’.

Furthermore, we also provide the HT tests that use two-step estimators such as IPW estimators, IPW-FE and IPW-FD estimators. In an implementation of an HT test that uses the IPW estimators, all we need to do is replace 'htfefd2_mi' with 'htfefd_ipw' in (3)’.

Both data and do file codes for empirical examples and ado files to implement the HT tests can be downloaded at http://dwkwak.weebly.com/research.html.