

# Finite-Population and Partially-Matched-Sample Corrections in Pseudo Panel Minimum Distance Estimation

Fei Jia<sup>1</sup>

<sup>1</sup>Department of Economics, Chaifetz School of Business  
Saint Louis University

2024 Asian Meeting of the Econometric Society in China

# Outline

- 1 Motivation
- 2 A Pseudo Panel Model
- 3 Sampling-Based Finite-Population Correction
- 4 Partially-Matched-Sample Correction

# Outline

- 1 Motivation
- 2 A Pseudo Panel Model
- 3 Sampling-Based Finite-Population Correction
- 4 Partially-Matched-Sample Correction

## Population of interest is not infinite

- Classical survey sampling: a random sample is drawn from an infinite super-population.
  - ▶ The super-population may be the right target: one may be more concerned with a conceptual population of people like those at present living in China. See Barnard (1973).
  - ▶ Simplifies asymptotic analysis.
- But there are cases where the infinite super population assumption is not appropriate.
  - ▶ “The Current Population Survey (CPS) must estimate such parameters as unemployment rates for the population of the U.S.A. during each particular month and super-population parameters are irrelevant to the main purpose of the C.P.S.” Hartley (1975).
  - ▶ The infinite super population may still be good enough approximation if the sampling rate  $\lambda$  is small.
  - ▶ The finite-population correction (fpc) is needed only if the sample is nontrivial relative to the finite population of interest.
- fpc can become more prominent as sampling technology advances.

## Sampling may not be independent over time

- A standard pseudo panel data set is composed of repeated cross sections that are independent over time.
- Depending on the sampling design, samples from different time period may not be independent. Example: CPS itself; the merged outgoing rotation groups (MORG) in CPS; the Australia Labour Force Survey (LFS) and the Brazil Continuous National Household Sample Survey (Continuous PNAD)

# Outline

- 1 Motivation
- 2 A Pseudo Panel Model
- 3 Sampling-Based Finite-Population Correction
- 4 Partially-Matched-Sample Correction

## Panel at the individual level

- Population model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + d_t\boldsymbol{\eta}' + f_i + u_{it}, \quad t = 1, \dots, T \quad (1)$$

or

$$y_{it} = \underline{\mathbf{x}}_{it}\boldsymbol{\theta} + e_i + u_{it}, \quad t = 1, \dots, T \quad (2)$$

where  $\underline{\mathbf{x}}_{it} \equiv (\mathbf{x}_{it}, d_t, c_i)$  for  $d_t$  the time dummy vector and  $c_i$  the group dummy vector.  $u_{it}$  is the idiosyncratic error.  $f_i$  is the individual-level fixed effects.

- $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'$  is the structural parameter of interest.
- $G$  predetermined time-invariant groups. Let  $g_i$  be the group membership for draw  $i$ . For  $g_i = g$ ,  $E(f_i|g) = \boldsymbol{\alpha}_g$ , the group effect.
- $e_i \equiv f_i - \boldsymbol{\alpha}_{g_i}$  is the individual fixed effect net of the group effect. The composite error is

$$\boldsymbol{\varepsilon}_{it} = e_i + u_{it} = (f_i - \boldsymbol{\alpha}_{g_i}) + u_{it} = y_{it} - \underline{\mathbf{x}}_{it}\boldsymbol{\theta}. \quad (3)$$

## Panel at the group level

- Imbens and Wooldridge (2007): minimum distance (MD) framework is a natural fit.
- Key identifying assumption: the exogeneity of  $g_i$  with respect to  $u_{it}$ , i.e.,

$$E(u_{it}|g) = 0, \quad g = 1, 2, \dots, G. \quad (4)$$

- $g_i$  is essentially a valid instrumental variable (IV), but used differently to get to the group-level model:

$$\mu_{gt}^y = \mu_{gt}^x \theta, \quad g = 1, \dots, G, \quad t = 1, \dots, T, \quad (5)$$

where  $\mu_{gt}^y \equiv E(y_{it}|g)$  and  $\mu_{gt}^x \equiv E(x_{it}|g)$  are the group means of the covariates in the population.

- $GT$  group-time cells. A  $G \times T$  pseudo panel.
- Inoue (2008) uses GMM. Conditional moments evaluated at all discrete values of  $g_i$  and get exactly the same conditions.



## The fixed-effects estimator for pseudo panel models

- Equation (5) suggests a naive approach to estimate  $\theta$

$$\check{\theta} = \left( \sum_{g,t} \hat{\mu}_{gt}^x \hat{\mu}_{gt}^x \right)^{-1} \sum_{g,t} \hat{\mu}_{gt}^x \hat{\mu}_{gt}^y. \quad (6)$$

where  $\hat{\mu}_{gt}^y$  and  $\hat{\mu}_{gt}^x$  are the sample group means of  $y_{it}$  and  $x_{it}$ .

- FE s.e.'s are invalid as they do not account for the estimation errors in  $\hat{\mu}_{gt}^y$  and  $\hat{\mu}_{gt}^x$

## MD for pseudo panel models

- Reduced-form parameter vector  $\pi \equiv (\mu_{11}^y, \mu_{11}^x, \mu_{12}^y, \mu_{12}^x, \dots, \mu_{GT}^y, \mu_{GT}^x)'$

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} N(0, \Omega) \quad (7)$$

- Rewrite the group-level equations in the “composite residual” form

$$h_{gt}(\pi, \theta) \equiv -\mu_{gt}^y + \mu_{gt}^x \theta = 0, \quad g = 1, \dots, G, \quad t = 1, \dots, T,$$

and stack all the residuals  $h_{gt}$  in one column vector  $h(\pi, \theta)$  as

$$h(\pi, \theta) \equiv -\mu^y + \mu^x \theta. \quad (8)$$

- Essentially by the delta method,

$$\sqrt{nh}(\hat{\pi}, \theta) \xrightarrow{d} N(0, M(\theta)). \quad (9)$$

$$M(\theta) = B(\theta)\Omega B(\theta). \quad (10)$$

where  $M(\theta)$  is the inverse of the optimal weighting matrix;

$B(\theta) = \nabla_{\pi} h(\pi, \theta) = \text{diag}\{(-1, \beta'_{gt})\}$  for  $\beta_{gt} \equiv \beta_1 + \alpha_g + \eta_t$ .

## MD for pseudo panel models, continued

- the optimal MD estimator for  $\theta$ :

$$\hat{\theta}^{opt} = \underset{\theta}{\operatorname{argmin}} h(\hat{\pi}, \theta)' \hat{M}^{-1} h(\hat{\pi}, \theta). \quad (11)$$

where  $\hat{M}$  is an estimator for  $M(\theta)$  using some initial estimator for  $\theta$ .

- Closed-form

$$\hat{\theta}^{opt} = \left[ \hat{\mu}^x' \hat{M}^{-1} \hat{\mu}^x \right]^{-1} \hat{\mu}^x' \hat{M}^{-1} \hat{\mu}^y. \quad (12)$$

- Variance

$$\operatorname{Avar} \left[ \sqrt{n}(\hat{\theta}^{opt} - \theta) \right] = \left[ \mu^x' M^{-1} \mu^x \right]^{-1} \quad (13)$$

# Structure of M

- Diagonal if samples are independent over time: recall  $\varepsilon_{it} \equiv e_i + u_{it}$

$$M = \text{diag} \{ (\rho_{gt} \kappa_t)^{-1} \text{Var}(\varepsilon_{it}|g) \} \quad (14)$$

where

- ▶  $\rho_{gt}$  is the fraction of population  $t$  in group  $g$ . Usually  $\rho_{gt} = \rho_g$  for a stable population.
- ▶  $\kappa_t$  the fraction of the whole population panel accounted for by population  $t$

# Outline

- 1 Motivation
- 2 A Pseudo Panel Model
- 3 Sampling-Based Finite-Population Correction**
- 4 Partially-Matched-Sample Correction

## A finite population: sample mean

- Notation:  $N$  for quantities in the population and  $S$  for quantities in the sample.
- Given a finite population  $\Pi_N = \{y_{N1}, y_{N2}, \dots, y_{NN}\}$ , suppose we are interested in the population mean of  $y$

$$\bar{y}_N = \frac{1}{N} \sum_{i=1}^N y_{Ni}$$

- A sample is a subset of  $\Pi_N$  represented by the vector of inclusion indicators  $(Z_1, Z_2, \dots, Z_N) \in \{0, 1\}^N$ . An intuitive estimator for  $\bar{y}_N$  is the sample average

$$\bar{y}_S = \frac{1}{n} \sum_{i:Z_i=1} y_{Ni} = \frac{1}{n} \sum_{i=1}^N Z_i y_{Ni}$$

where  $n = \sum_{i=1}^N Z_i$

- No design-based uncertainty in the sense of Abadie et al. (2020).

## A finite population: variance of the sample mean

- It can be shown that  $\bar{y}_S$  has mean  $\bar{y}_N$  and variance

$$\text{var}(\bar{y}_S) = \left( \frac{1}{n} - \frac{1}{N} \right) v_N,$$

where  $v_N = \frac{1}{N-1} \sum_{i=1}^N (y_{Ni} - \bar{y}_N)^2$  is the population variance of  $y$ . An unbiased estimator for  $v_N$  is the sample variance

$$\hat{v}_N = \frac{1}{n-1} \sum_{i:Z_i=1} (y_{Ni} - \bar{y}_S)^2 = \frac{1}{n-1} \sum_i Z_i (y_{Ni} - \bar{y}_S)^2.$$

- Therefore, an unbiased estimator for  $\text{var}(\bar{y}_S)$  is

$$\widehat{\text{var}(\bar{y}_S)} = \left( \frac{1}{n} - \frac{1}{N} \right) \hat{v}_N. \quad (15)$$

## A finite population CLT

- Adopt the classical finite population CLTs for simple random sampling by Hájek (1960), reviewed recently by Li and Ding (2017). Also used is Lehmann (1999).
- Theorem 1 in Li and Ding 2017 (originally in Hájek (1960)): as  $N \rightarrow \infty$ , we have

$$\frac{\bar{y}_S - \bar{y}_N}{\sqrt{\text{var}(\bar{y}_S)}} \xrightarrow{d} \text{Normal}(0, 1) \quad (16)$$

if  $\frac{1}{\min(n, N-n)} \cdot \frac{m_N}{v_N} \rightarrow \infty$ , where  $m_N = \max_{1 \leq i \leq N} (y_{Ni} - \bar{y}_N)^2$  is the maximum squared distance of the  $y_{Ni}$ 's from the population mean  $\bar{y}_N$ .

- Multiply both sides of Equation (16) by  $\sqrt{n} \sqrt{\text{var}(\bar{y}_S)}$  to get the root-n format

$$\sqrt{n}(\bar{y}_S - \bar{y}_N) \xrightarrow{d} \text{Normal}(0, (1 - \lambda) v_N)$$

where

$$\lambda \equiv n/N \quad (17)$$



## The finite population correction to M

- Recall  $M = \text{diag} \{ (\rho_{gt} \kappa_t)^{-1} \text{Var}(e_{it}|g) \}$
- With fpc:

$$M_\lambda = (1 - \lambda) \text{diag} \{ (\rho_g \kappa_t)^{-1} \text{Var}(e_{it}|g) \} \quad (18)$$

and

$$\text{Avar} \left[ \sqrt{n}(\hat{\theta}^{\text{opt}} - \theta) \right] = [\mu^x{}' M_\lambda^{-1} \mu^x]^{-1} \quad (19)$$

$$= (1 - \lambda) [\mu^x{}' M^{-1} \mu^x]^{-1} \quad (20)$$

- If  $\lambda_t$ ,  $M_\lambda = \text{diag} \{ (1 - \lambda_t) (\rho_g \kappa_t)^{-1} \text{Var}(e_{it}|g) \}$
- Estimation is straightforward.
- An example where this may matter: the Integrated Public Use Microdata Series (IPUMS) include a random sample of 10% of the census

## Simulation: setup

- Individual-level panel DGP

$$y_{it} = \beta_1 + \beta_2 x_{it} + \eta_t + f_i + u_{it}, \quad i = 1, \dots, N_t, \quad t = 1, \dots, T. \quad (21)$$

- ▶  $\beta = (\beta_1, \beta_2) = (1, 1)$ ,  $\eta_t = t - 1$ ,  $\alpha_g = g - 1$ ,
  - ▶  $f_i \sim N(\alpha_g, 10)$ .
  - ▶  $G = 6$ ,  $T = 4$ .
  - ▶ “education”  $x_{it} \sim N(gt/6, 1)$ ,
- Key: In the first step,  $T$  finite populations of fixed sizes are simulated using (21) and then kept fixed. In the second step, a random sample is independently drawn from each of the  $T$  populations simulated in the first step with the sampling rate  $\lambda$ .

Figure: Inference comparison for the optimal pseudo-panel MD estimators with and without the finite-population correction ( $\tilde{\beta}_2$  and  $\hat{\beta}_2$ , respectively).

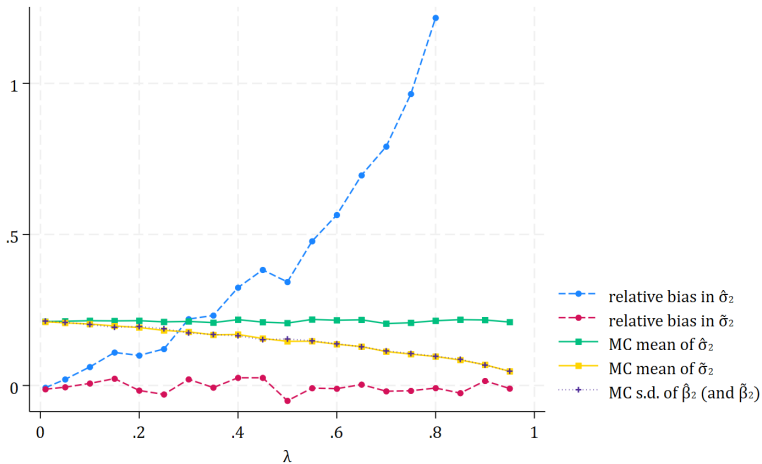


Table 1: Inference comparison for the optimal pseudo-panel MD estimators with and without the finite-population correction.

$\lambda$	$\hat{\beta}_2$ ( $\hat{\beta}_2$ ) s.d.	$\tilde{\sigma}_2$ mean	$\hat{\sigma}_2$ mean	$\tilde{\sigma}_2$ % Bias	$\hat{\sigma}_2$ % Bias	$\hat{\sigma}_2$ % Bias Asy.
.01	0.213	0.211	0.212	-1.2	-0.7	0.5
.05	0.209	0.207	0.213	-0.6	2.0	2.6
.10	0.202	0.203	0.214	0.7	6.1	5.4
.15	0.193	0.197	0.214	2.3	10.9	8.5
.20	0.195	0.192	0.215	-1.7	10.0	11.8
.25	0.188	0.182	0.211	-2.9	12.1	15.5
.30	0.174	0.177	0.212	2.1	22.0	19.5
.35	0.169	0.168	0.208	-0.7	23.2	24.0
.40	0.165	0.169	0.218	2.6	32.4	29.1
.45	0.152	0.156	0.210	2.5	38.3	34.8
.50	0.154	0.146	0.207	-5.1	34.3	41.4
.55	0.148	0.147	0.219	-0.9	47.8	49.1
.60	0.138	0.137	0.216	-1.0	56.5	58.1
.65	0.128	0.129	0.217	0.3	69.6	69.0
.70	0.114	0.112	0.205	-1.9	79.1	82.6
.75	0.106	0.104	0.208	-1.7	96.5	100.0
.80	0.097	0.096	0.214	-0.8	121.7	123.6
.85	0.087	0.084	0.218	-2.5	151.7	158.2
.90	0.068	0.069	0.217	1.5	221.0	216.2
.95	0.047	0.047	0.210	-1.0	342.8	347.2

Mean and standard deviation (s.d.) are across Monte Carlo Simulations.

Percentage biases are calculated relative to the s.d..

The last column is the asymptotic value of the percentage bias in  $\hat{\sigma}_2$ .

## fpc in general

- Can be trivially extended to any regression analysis where the sample is a nontrivial proportion of the finite population.

# Outline

- 1 Motivation
- 2 A Pseudo Panel Model
- 3 Sampling-Based Finite-Population Correction
- 4 Partially-Matched-Sample Correction**

## Recall: Sampling may not be independent over time

- A standard pseudo panel data set is composed of repeated cross sections that are independent over time.
- Depending on the sampling design, samples from different time period may not be independent. Example: CPS itself; the merged outgoing rotation groups (MORG) in CPS; the Australia Labour Force Survey (LFS) and the Brazil Continuous National Household Sample Survey (Continuous PNAD)

# Sampling Design

## Definition

### Simplified Partially Matched Sampling Design

- 1 At time  $t = 1$ : (i) randomly sample  $n_1$  observations to form sample 1 and (ii) randomly select a subsample of  $n_{1,stay}$  observations from sample 1 to keep to period 2.
- 2 At time  $t$  for  $t = 2, \dots, T$ : (i) randomly sample  $(n_t - n_{t-1,stay})$  observations from the population and combine with the subsample left from  $t - 1$  to get sample  $t$  and (ii) if  $t < T$ , randomly select a subsample of  $n_{t,stay}$  observations in the newly drawn subsample of  $(n_t - n_{t,stay})$  observations to keep to period  $t + 1$ ; if  $t = T$ , stop. Repeat step 2 until stop is reached.



# Temporal correlation

- Recall  $M(\theta) = \text{diag} \{ (\rho_g \kappa_t)^{-1} \text{Var}(\varepsilon_{it}|g) \}$  in the standard case. It's diagonal because  $\sqrt{n}\hat{\mu}_{gt}^\varepsilon$  and  $\sqrt{n}\hat{\mu}_{gs}^\varepsilon$   $t \neq s$  are asymptotically independent.
- Now need to derive the joint asymptotic distribution of  $\sqrt{n}\hat{\mu}_{gt}^\varepsilon$  and  $\sqrt{n}\hat{\mu}_{gs}^\varepsilon$  for  $t \neq s$
- Notice that  $\mu_{gt}^\varepsilon = \mu_{gt}^y - \mu_{gt}^x \theta = -h_{gt}(\pi, \theta)$ . So  $\sqrt{n}\hat{\mu}_{gt}^\varepsilon = -\sqrt{nh_{gt}}(\pi, \theta)$

Decompose  $\hat{\mu}_{gt}^\varepsilon$  into two parts: the matched and the unmatched

- Decompose  $\hat{\mu}_{gt}^\varepsilon$  as

$$\hat{\mu}_{gt}^\varepsilon = n_{gt}^{-1} \sum_{i \in \mathcal{I}_{ts}} r_{it,g} \varepsilon_{it} + n_{gt}^{-1} \sum_{i \in \mathcal{I}_t \setminus \mathcal{I}_{ts}} r_{it,g} \varepsilon_{it}. \quad (22)$$

where  $r_{itg} = 1_{\{g_i=g, i \in \mathcal{I}_t\}}$ ,  $\mathcal{I}_t$  the index set for sample  $t$ ,  $\mathcal{I}_{ts} = \mathcal{I}_t \cap \mathcal{I}_s$  for matched, and  $\mathcal{I}_t \setminus \mathcal{I}_{ts}$  for unmatched in sample  $t$ .

- Matching rate: Let  $\psi_{ts,t}$  be the fraction of sample  $t$  that is in the matched subsample between sample  $t$  and sample  $s$  and  $\psi_{ts,s}$  be defined similarly.
- For ease of illustration, assume  $\psi_{ts,s} = \psi_{ts,t} = \psi$  and  $\rho_{gt} = \rho_g$ .

Decompose  $\hat{\mu}_{gt}^\varepsilon$  into two parts: the matched and the unmatched

- CLT on the 1st part (the unmatched), for  $t$  and  $s$  jointly

$$\begin{pmatrix} \sqrt{n_t} \hat{\mu}_{gt1}^\varepsilon \\ \sqrt{n_s} \hat{\mu}_{gs1}^\varepsilon \end{pmatrix} \xrightarrow{d} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (1-\psi)\rho_g E[\varepsilon_{it}^2|g] & 0 \\ 0 & (1-\psi)\rho_g E[\varepsilon_{is}^2|g] \end{pmatrix} \right]$$

- CLT on the 2nd part (the matched), for  $t$  and  $s$  jointly,

$$\begin{pmatrix} \sqrt{n_t} \hat{\mu}_{gt2}^\varepsilon \\ \sqrt{n_s} \hat{\mu}_{gs2}^\varepsilon \end{pmatrix} \xrightarrow{d} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi\rho_g E[\varepsilon_{it}^2|g] & \psi\rho_g \text{Var}(f_i|g) \\ \psi\rho_g \text{Var}(f_i|g) & \psi\rho_g E[\varepsilon_{is}^2|g] \end{pmatrix} \right],$$

## Merge the two asymptotic distributions

- Combine the two parts, and notice  $n_{gt}/n_t \rightarrow \rho_g, n_t/n \rightarrow \kappa_t, n_s/n \rightarrow \kappa_s$ , we have

$$\sqrt{n} \begin{pmatrix} \hat{\mu}_{gt}^\varepsilon \\ \hat{\mu}_{gs}^\varepsilon \end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\rho_g \kappa_t)^{-1} E[\varepsilon_{it}^2 | g] & \psi \rho_g^{-1} (\kappa_t \kappa_s)^{-1/2} \text{Var}(f_i | g) \\ \psi \rho_g^{-1} (\kappa_t \kappa_s)^{-1/2} \text{Var}(f_i | g) & (\rho_g \kappa_t)^{-1} E[\varepsilon_{is}^2 | g] \end{pmatrix} \right)$$

# The Main Theorem

## Theorem

Define  $\Psi_{gts}^\varepsilon = (\Psi_{ts,t} \Psi_{ts,s})^{\frac{1}{2} \mathbb{I}_{ts}} (\rho_g \mathbf{K}_t \rho_g \mathbf{K}_s)^{-\frac{1}{2}} \sigma_{\varepsilon, gts}$  where  $\mathbb{I}_{ts} \equiv 1_{\{t \neq s\}}$ ,  $\sigma_{\varepsilon, g}^2 \equiv \text{Var}(\varepsilon_{it} | g)$  and  $\sigma_{\varepsilon, gts} \equiv \text{Cov}(\varepsilon_{it}, \varepsilon_{is} | g)$  which degenerates to  $\sigma_{\varepsilon, g}^2$  if  $t = s$ . Then, under the given DGP (1) and the sampling design in Assumption 1, the joint asymptotic distribution of  $\sqrt{n} \hat{\mu}_{gt}^\varepsilon$  and  $\sqrt{n} \hat{\mu}_{gs}^\varepsilon$  is

$$\sqrt{n} \begin{pmatrix} \hat{\mu}_{gt}^\varepsilon \\ \hat{\mu}_{gs}^\varepsilon \end{pmatrix} \xrightarrow{d} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Psi_{ggt}^\varepsilon & \Psi_{gst}^\varepsilon \\ \Psi_{gts}^\varepsilon & \Psi_{gss}^\varepsilon \end{pmatrix} \right] \quad (23)$$

In addition, the joint asymptotic distribution of  $\sqrt{n} \hat{\mu}_{gt}^\varepsilon$  for  $g = 1, \dots, G$  and  $t = 1, \dots, T$  can be written as

$$\sqrt{n} \hat{\mu}^\varepsilon \xrightarrow{d} N(0, \Psi^\varepsilon) \quad (24)$$

where  $\Psi^\varepsilon = \text{diag}(\Psi_g^\varepsilon)_G$  and

$$\Psi_g^\varepsilon = \{\Psi_{gts}^\varepsilon\}_{TT}. \quad (25)$$

## The Optimal MD estimator for partially matched samples

- Let  $\hat{\theta}_\psi$  be the optimal MD estimator using the inverse of  $\hat{\Psi}^\varepsilon$  as the weighting matrix,

$$\hat{\theta}_\psi = \left[ \hat{\mu}^x{}' \left( \hat{\Psi}^\varepsilon \right)^{-1} \hat{\mu}^x \right]^{-1} \hat{\mu}^x{}' \left( \hat{\Psi}^\varepsilon \right)^{-1} \hat{\mu}^y. \quad (26)$$

The asymptotic variance of  $\hat{\theta}_\psi$  can be obtained by replacing the M matrix in (13) with  $\Psi^\varepsilon$ , i.e.,

$$Avar(\hat{\theta}_\psi) = n^{-1} \left[ \mu^x{}' (\Psi^\varepsilon)^{-1} \mu^x \right]^{-1} \quad (27)$$

## Discussion

- $\Psi_{gts}^{\varepsilon} = (\psi_{ts,t} \psi_{ts,s})^{\frac{1}{2}} \mathbb{I}_{ts} (\rho_g \mathbf{K}_t \rho_g \mathbf{K}_s)^{-\frac{1}{2}} \sigma_{\varepsilon, gts}$  reveals why the off-diagonal elements in their asymptotic variance is nonzero: for  $\varepsilon_{it}$  and  $\varepsilon_{is}$  with  $i$  referring to the same individual,

$$\sigma_{\varepsilon, gts} = \sigma_{f, g}^2 + \sigma_{u, gts}$$

where  $\sigma_{f, g}^2 = \text{Var}(f_i | g)$  and  $\sigma_{u, gts} = \text{Cov}(u_{it}, u_{is} | g)$ .

- Therefore, the partially matched subsample causes correlation between the composite error  $\varepsilon_{it}$  and  $\varepsilon_{is}$  via the fixed effect  $f_i$  as well as the serial correlation in  $u_{it}$  if there is any.
- Numerically, a relatively large  $\psi$  is also important.

## $\Psi^\varepsilon$ for MORG, $T = 3$

- For MORG, that  $\psi_1 = 1/2$  implies only data from adjacent years have matched subsamples,

$$\Psi_g^\varepsilon = (1 - \lambda) \times \begin{bmatrix} (\rho_g \kappa_1)^{-1} \sigma_{g1}^2 & \psi_1 (\rho_g^2 \kappa_1 \kappa_2)^{-\frac{1}{2}} \sigma_{g12} & 0 \\ \psi_1 (\rho_g^2 \kappa_1 \kappa_2)^{-\frac{1}{2}} \sigma_{g12} & (\rho_g \kappa_2)^{-1} \sigma_{g2}^2 & \psi_1 (\rho_g^2 \kappa_2 \kappa_3)^{-\frac{1}{2}} \sigma_{g23} \\ 0 & \psi_1 (\rho_g^2 \kappa_2 \kappa_3)^{-\frac{1}{2}} \sigma_{g23} & (\rho_g \kappa_3)^{-1} \sigma_{g3}^2 \end{bmatrix}. \quad (28)$$



## $\Psi^\varepsilon$ for LFS , $T = 3$

- For the Australia Labour Force Survey,  $\psi_1 = 7/8$ ,  $\psi_2 = 6/8$ , .... When  $T = 3$ ,

$$\Psi_g^\varepsilon = (1 - \lambda) \times \begin{bmatrix} (\rho_g \kappa_1)^{-1} \sigma_{g1}^2 & \psi_1 (\rho_g^2 \kappa_1 \kappa_2)^{-\frac{1}{2}} \sigma_{g12} & \psi_2 (\rho_g^2 \kappa_1 \kappa_3)^{-\frac{1}{2}} \sigma_{g13} \\ \psi_1 (\rho_g^2 \kappa_1 \kappa_2)^{-\frac{1}{2}} \sigma_{g12} & (\rho_g \kappa_2)^{-1} \sigma_{g2}^2 & \psi_1 (\rho_g^2 \kappa_2 \kappa_3)^{-\frac{1}{2}} \sigma_{g23} \\ \psi_2 (\rho_g^2 \kappa_1 \kappa_3)^{-\frac{1}{2}} \sigma_{g13} & \psi_1 (\rho_g^2 \kappa_2 \kappa_3)^{-\frac{1}{2}} \sigma_{g23} & (\rho_g \kappa_3)^{-1} \sigma_{g3}^2 \end{bmatrix} \quad (29)$$

## Estimation

- For  $t = s$ ,  $\sigma_{\varepsilon, gts} = \sigma_{\varepsilon, gt}^2 = \text{Var}(\varepsilon_{it}|g)$ ,

$$\hat{\sigma}_{\varepsilon, gt}^2 = n_{gt}^{-1} \sum_{i \in \mathcal{I}_t} r_{itg} (\check{u}_{it} - \tilde{u}_{gt})^2 \quad (30)$$

where  $\check{u}_{it}$  is the residual obtained using the FE estimator  $\check{\theta}$ , and  $\tilde{u}_{gt} = n_{gt}^{-1} \sum_{i=1}^{n_t} r_{itg} \check{u}_{it}$ .

- For  $t \neq s$ , use the matched subsample between sample  $t$  and sample  $s$ : notice that  $r_{itg} = r_{isg}$  for  $i \in \mathcal{I}_{ts}$ ,

$$\hat{\sigma}_{\varepsilon, gts} = m_{gts}^{-1} \sum_{i \in \mathcal{I}_{ts}} r_{itg} (\check{u}_{it} - \tilde{u}_{gts})(\check{u}_{is} - \tilde{u}_{gts}), \text{ for } t \neq s, \quad (31)$$

where  $m_{gts} = \sum_{i \in \mathcal{I}_{ts}} r_{itg}$ .

## Simulation

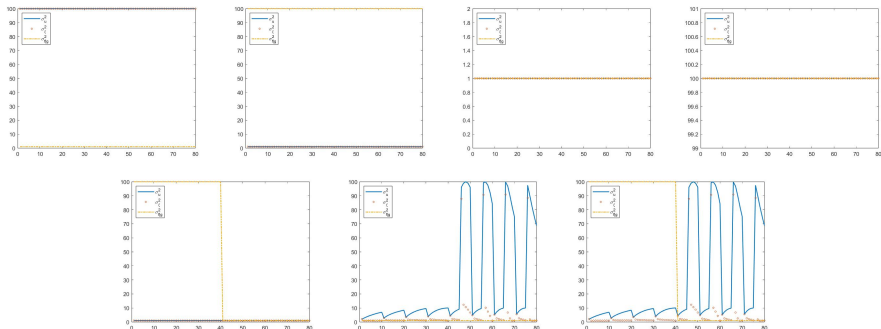
- $x_{it} \sim N(gt/6, 1)$ ,  $\beta_1 = \beta_2 = 1$ ,  $\eta_t = t - 1$  and  $\alpha_g = g - 1$ .  $x_{it}$  is independent of  $g_i$ ,  $e_i$  and  $u_{it}$ .

$$y_{it} = \beta_1 + \beta_2 x_{it} + \eta_t + \underbrace{(\alpha_{g_i} + e_i)}_{f_i} + u_{it}, \quad i \in \mathcal{I}_t, \quad t = 1, \dots, T, \quad (32)$$

- The simulation cases to be considered differ in their variance specifications on  $f_i|g$  (equivalently,  $e_i|g$ ) and variance and serial covariance specifications on  $u_{it}|g$ . Four benchmark cases (Cases 1.1 to 1.4) and three extended cases (Cases 2 to 4) are designed
- The yearly MORG sampling design is adopted with constant sample sizes ( $n_t = n_0$  and thus  $\kappa_t = 1/T$ ) and fixed group proportions ( $\rho_{gt} = \rho_g = 1/G$ ). This implies a constant matching rate  $\psi_{ts,t} = \psi_{ts,s} = \psi = 50\%$
- $G = 8$  and  $T = 10$ .

# $\sigma_{f,g}^2$ and $\sigma_{u,gt}^2$ in the seven designs

Figure: The plots of  $\sigma_{u,gt}^2$ ,  $\sigma_{\xi,gt}^2$  and  $\sigma_{f,g}^2$  against  $j(g,t) = (g-1)T + t$  for  $g = 1, \dots, G$  and  $t = 1, \dots, T$  under  $G = 8$  and  $T = 10$ .



# Simulation

- Benchmark cases:
  - ▶ Case 1.1,  $\sigma_{f,g}^2 = 1$ ,  $\sigma_{u,gt}^2 = 100$ ; ( $\sigma_{u,gt}^2$  dominates  $\sigma_{f,g}^2$ )
  - ▶ Case 1.2,  $\sigma_{f,g}^2 = 100$ ,  $\sigma_{u,gt}^2 = 1$ ; ( $\sigma_{f,g}^2$  dominates  $\sigma_{u,gt}^2$ )
  - ▶ Case 1.3,  $\sigma_{f,g}^2 = 1$ ,  $\sigma_{u,gt}^2 = 1$ ; ( $\sigma_{f,g}^2$ ,  $\sigma_{u,gt}^2$  same low level)
  - ▶ Case 1.4,  $\sigma_{f,g}^2 = 100$ ,  $\sigma_{u,gt}^2 = 100$ . ( $\sigma_{f,g}^2$ ,  $\sigma_{u,gt}^2$  same high level)
- Case 2: cohort-wise heteroskedastic  $f_i$  only

$$\sigma_{f,g}^2 = 100 \cdot 1_{\{g \leq G/2\}} + 1_{\{g > G/2\}}, \quad (33)$$

( $\sigma_{f,g}^2$  dominates  $\sigma_{u,gt}^2$  in half of the cells)

- Case 3: cell-wise heteroskedastic and serial correlated  $u_{it}$  only.

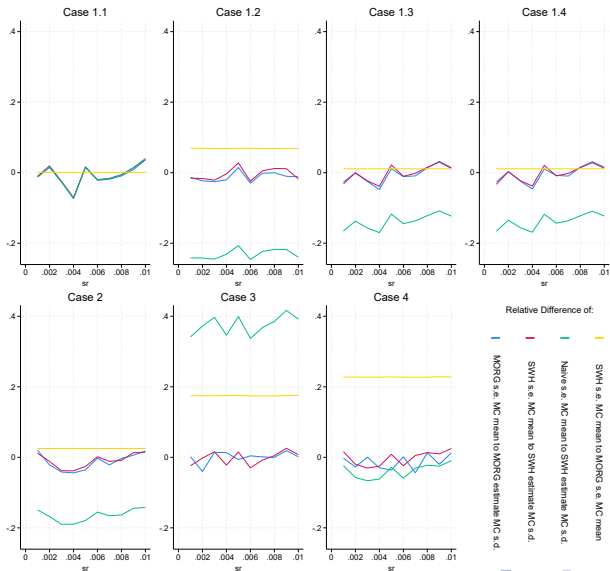
$$u_{it} = \gamma_0 u_{it-1} + \xi_{it} \quad (34)$$

$$\sigma_{\xi,gt}^2 = \text{Var}(\xi_{it}|g) = \max \left\{ 1, \left[ \sigma_{b,gt}^2 - \gamma_0^2 \sigma_{b,g(t-1)}^2 \right] \right\}. \quad (35)$$

where  $\sigma_{b,gt}^2 \equiv b_{gt} \left[ \sin \left( a \frac{gt}{GT} \right) \right]^p$  and  $b_{gt} \equiv 10 \cdot 1_A + 100 \cdot 1_{\bar{A}}$  with  $A = \{g \leq G/2 \text{ or } t \leq T/2\}$  and  $(\gamma_0, a, p) = (-0.95, 3.1415, 0.5)$ .

- Case 4, merges Case 2 and Case 3

Figure:  $G = 6$  and  $T = 10$ . 1000 replications. The sampling rate (sr) varies from 0.1% to 1% with step 0.1%. The sample cohort size in each period is roughly 240.



# The Sandwich-Form Inference

- SWH

$$\widehat{Avar}(\hat{\theta}_M) = n^{-1} \hat{\Xi} \hat{A} \hat{\Xi} \quad (36)$$

where  $\hat{\Xi} \equiv (\hat{\mu}^x' M^{-1} \hat{\mu}^x)^{-1}$  and  $\hat{A} \equiv \hat{\mu}^x' \hat{M}^{-1} \hat{\Psi}^\varepsilon \hat{M}^{-1} \hat{\mu}^x$ .

- Yellow line: MORG s.e. (using  $\hat{\Psi}^\varepsilon$ ) is more efficient than SWH s.e. when  $\hat{\Psi}^\varepsilon$  is nontrivial different from the diagonal  $\hat{M}$ .
- Blue and red line: Both MORG and SWH provide good finite-sample approximations to their respective asymptotic targets.
- Green line: Naive s.e. generally biased.

## An empirical illustration: estimating monetary returns to education

- For surveys of this literature, see Card 1999, Card 2001, Heckman, Lochner, and Todd 2006, McMahon 2009 and Oreopoulos and Petronijevic 2013 among others.
- The specification is similar to that in Angrist and Krueger (1991).
- The MORG files spanning from 2010 to 2019 ( $T = 10$ ).



Table 1: CPS, 2010-2019

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	PPOLS	POLS	MORG	MORG	SWH	SWH	Naive	Naive
Years of Edu.	0.085*** (0.000)	0.084*** (0.000)	0.249*** (0.053)	0.218*** (0.052)	0.294*** (0.063)	0.245*** (0.061)	0.294*** (0.068)	0.245*** (0.067)
Married	0.141*** (0.002)	0.127*** (0.002)	0.675*** (0.120)	0.297 (0.164)	0.688*** (0.131)	0.319 (0.183)	0.688*** (0.121)	0.319 (0.165)
Black	-0.184*** (0.003)	-0.187*** (0.003)	0.306 (0.801)	0.009 (0.681)	1.109 (0.922)	0.498 (0.781)	1.109 (0.897)	0.498 (0.768)
Age		0.083*** (0.001)		0.008 (0.099)		-0.035 (0.117)		-0.035 (0.121)
Age Squared		-0.001*** (0.000)		-0.000** (0.000)		-0.000* (0.000)		-0.000** (0.000)
Metropolitan	0.081*** (0.002)	0.081*** (0.002)	0.224 (0.544)	0.344 (0.468)	0.192 (0.703)	0.365 (0.586)	0.192 (0.803)	0.365 (0.671)
R-squared	0.269	0.276						
N	449568	449568						

Years 2010-2019

Results on region dummies, group dummies and time dummies are omitted.






\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

- Consistent with the theory and the simulation results, the SWH s.e. for education is greater than the MORG s.e. no matter the age function is included or not in the specification (Column 3 vs 5, 11%; Column 4 v.s. 6, 14%).
- noticeable discrepancy between the Naive s.e. and the SWH s.e. under both specifications, indicating potential bias in the Naive s.e. for this particular application.
- additional estimates separately obtained from 2010-2014 and 2015-2019 in online appendix.
- The results from pooled OLS (POLS, first two columns) are reported for comparison.







# Summary

- The simplest version of the finite population correction is a shrinkage factor that equals  $1 - \lambda$  to the usual optimal weighting matrix and asymptotic variance under infinite super population, where  $\lambda$  is the sampling rate.
- The partially matched sampling design causes temporal correlation between samples from different time periods, giving rise to a block-diagonal weighting matrix in the minimum distance (MD) estimation of pseudo panel models.
  - ▶ (Asymptotically) more efficient.
  - ▶ Conventional s.e. conservative.




## References I

-  Abadie, Alberto et al. (2020). “Sampling-based versus design-based uncertainty in regression analysis”. In: *Econometrica* 88.1, pp. 265–296.
-  Angrist, Joshua D and Alan B Krueger (1991). “Does Compulsory School Attendance Affect Schooling and Earnings?” In: *The Quarterly Journal of Economics* 106.4, pp. 979–1014. ISSN: 00335533, 15314650. URL: <http://www.jstor.org/stable/2937954>.
-  Barnard, George A (1973). “Discussion of a paper by VP Godambe and ME Thompson, Bayes, Fiducial and Frequentist Aspects of Regression Analysis in Survey-sampling”. In: *Journal of the Royal Statistical Society, B* 33, pp. 361–390.
-  Card, David (1999). “The causal effect of education on earnings”. In: *Handbook of labor economics* 3, pp. 1801–1863.
-  — (2001). “Estimating the return to schooling: Progress on some persistent econometric problems”. In: *Econometrica* 69.5, pp. 1127–1160.

## References II

-  Hájek, Jaroslav (1960). “Limiting distributions in simple random sampling from a finite population”. In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, pp. 361–374.
-  Hartley, H O (1975). “A "super-population viewpoint" for finite population sampling”. In: *Biometrics*, pp. 411–422.
-  Heckman, James J, Lance J Lochner, and Petra E Todd (2006). “Earnings functions, rates of return and treatment effects: The Mincer equation and beyond”. In: *Handbook of the Economics of Education* 1, pp. 307–458.
-  Imbens, Guido W. and Jeffrey M Wooldridge (2007). *What's new in econometrics?* NBER.
-  Inoue, Atsushi (2008). “Efficient estimation and inference in linear pseudo-panel data models”. In: *Journal of Econometrics* 142.1, pp. 449–466.
-  Lehmann, Erich Leo (1999). *Elements of large-sample theory*. New York: Springer, New York.

## References III

-  Li, Xinran and Peng Ding (2017). “General forms of finite population central limit theorems with applications to causal inference”. In: *Journal of the American Statistical Association* 112.520, pp. 1759–1769.
-  McMahon, Walter W (2009). *Higher learning, greater good: The private and social benefits of higher education*. JHU Press.
-  Oreopoulos, Philip and Uros Petronijevic (2013). “Making college worth it: A review of research on the returns to higher education”. In.