RECENT DEVELOPMENTS IN FINITE POPULATION AND CLUSTERED STANDARD ERRORS

IAAE 2017

Sapporo, Japan

June 25-29, 2017

Jeff Wooldridge

(Work with Alberto Abadie, Susan Athey, Guido Imbens)

1. Introduction and Motivation
2. Sampling-Based versus Design-Based Uncertainty
3. Cluster Sampling

1. Introduction and Motivation

- Typical microeconometrics approach to inference:

Assume independent, identically distributed draws from the population.

- Finite population, fixed sample size: Obtaining independent draws requires sampling with replacement.
- Sampling w/o replacement leads to correlation.
  - ▸ Finite population corrections available: $(1 - N/M)$.

Finite Population Problem

Question: What happens if we observe the entire population?

- We can collect information on state-level employment rates and right-to-work laws for all 50 U.S. states.
- Should we report a zero standard error? $(1 - M/M) = 0$.
- Abadie, Athey, Imbens, Wooldridge (2017): Introduce design uncertainty, possibly in addition to sampling uncertainty.

Clustering Problem

- You draw a large random sample, say $N = 10,000$, from a large population (working age adults in the United States).
- Workers are randomly assigned to a job training program (coin flip).
- You estimate a proportionate effect on earnings (simple regression):

$$\hat{\tau} = 0.058, \text{se}_{ehw} = 0.012$$

• Your RA points out that you know the state of residence of each person and suggests clustering the standard errors at the state level (50 states).

Question: Should you cluster?

(1) No.

(2) Yes.

(3) Does not matter.

• If you answer (2) or (3), why not cluster using the nine census regions?

- Suppose you cluster and obtain

$$\hat{\tau} = 0.058, \text{se}_{cluster,state} = 0.031$$

$$\hat{\tau} = 0.058, \text{se}_{cluster,region} = 0.049$$

versus

$$\hat{\tau} = 0.058, \text{se}_{ehw} = 0.012$$

- What should you do?
- Would your answer change if the RA suggested clustering by occupation. Or race?

2. Sampling versus Design Uncertainty

A. Difference in Means with Random Assignment

- Population of size M.
- X_i is the random binary treatment indicator.
  - ▸ For example, it indicates whether state i has a right-to-work law.
- We view X_i as a *potential cause*.
- Two potential outcomes for each unit i, $y_i(0)$ and $y_i(1)$.
- Essentially the Rubin Causal Model.

- The population average causal (treatment) effect:

$$\theta_M = \frac{1}{M}\sum_{i=1}^{M}[y_i(1) - y_i(0)] = \bar{y}_M(1) - \bar{y}_M(0)$$

$y_i(1) - y_i(0)$ is causal effect for population unit i

- The observed outcome is random:

$$Y_i = Y_i(X_i) = (1 - X_i)y_i(0) + X_i y_i(1)$$

- Population variances of the potential outcomes:

$$\sigma_M^2(x) = \frac{1}{M-1}\sum_{i=1}^{M}[y_i(x) - \bar{y}_M(x)]^2 \text{ for } x = 0, 1$$

- Population variance of the unit-level causal effects $y_i(1) - y_i(0)$:

$$\sigma_M^2(0,1) = \frac{1}{M-1}\sum_{i=1}^{M}\{y_i(1) - y_i(0) - [\bar{y}_M(1) - \bar{y}_M(0)]\}^2.$$

- Constant treatment effect: $\sigma_M^2(0,1) = 0$.

Assumptions:

1. Random assignment of the X_i.

2. Random Sampling without Replacement.

- R_i is a binary sampling indicator.
- $\rho \in (0, 1]$ is the sampling probability:

$$\mathrm{P}(\mathbf{R}_M = \mathbf{r}|\mathbf{X}_M) = \rho^{\left(\sum_{i=1}^{M} r_i\right)} \cdot (1-\rho)^{\left(M-\sum_{i=1}^{M} r_i\right)},$$

for all M-vectors $\mathbf{r}$ with i-th element $r_i \in \{0, 1\}$.

• Estimator: Difference in means between the treated and control subsamples:

$$\hat{\theta}_M = \bar{Y}_1 - \bar{Y}_0,$$

$$\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N} R_i X_i Y_i, \quad \bar{Y}_0 = \frac{1}{N_0}\sum_{i=1}^{N} R_i (1 - X_i) Y_i$$

$$N_1 = \sum_{i=1}^{M} R_i X_i, \qquad N_0 = \sum_{i=1}^{M} R_i (1 - X_i)$$

Lemma: Under the random assignment and random sampling assumptions, conditional on $N_0, N_1 > 0$,

$$\mathbb{E}_{(\mathbf{R},\mathbf{X})}\left(\hat{\theta}_M\right) = \theta_M$$

$$\mathbb{V}_{(\mathbf{R},\mathbf{X})}\left(\hat{\theta}_M\right) = \frac{\sigma_M^2(0)}{N_0} + \frac{\sigma_M^2(1)}{N_1} - \frac{\sigma_M^2(0,1)}{M}$$

- The usual variance formula is conservative: $\sigma_M^2(0,1) \geq 0$.
- When M is large the difference can be small.
- $N_1 = M_1$ and $N_0 = M_0$ are allowed [Neyman (1923)].

B. Regression with Attributes and Causes

- Asymptotics for multiple regression analysis with $M \to \infty$.
- Covariates of the potential cause type, X_i, such as state regulation.
  - Can be discrete, continuous, mixed random vectors.
- Covariates of the attribute or characteristic type, z_i, such as state geographic descriptors.
  - The z_i are nonrandom.

- There exists a set of potential outcomes, $y_i(x)$.
  - The $y_i(x)$ are nonrandom.
- The realized outcome for unit i, $Y_i = y_i(X_i)$, is random.
- Sample units from this population.
- R_i is the random binary indicator for whether unit i is sampled.

- The potential cause X_i can have a distribution that depends on z_i.
  - z_i contains "confounders."
- ρ_M is the sampling rate.
- Allow $\rho_M = 1$ (sample = population) as well as $\rho_M \to 0$ (random sampling from a large population).

- For a population of size M and sample of size N, we obtain OLS estimates:

$$(\hat{\theta}_{ols}, \hat{\gamma}_{ols}) = \arg\min_{\theta,\gamma} \sum_{i=1}^{M} R_i \cdot (Y_i - X_i\theta - z_i\gamma)^2,$$

where the R_i select the sample.

- Randomness comes from X_i and R_i:
  - Design Uncertainty (X_i)
  - Sampling Uncertainty (R_i)

Assumptions

(a) The assignments $X_1, \ldots, X_M$ are independent, but not (necessarily) identically distributed.

(b) Finite fourth moments that are uniformly bounded in M.

(c) Convergence of moments as $M \to \infty$ (for convenience).

(d) Random sampling.

(e) $M\rho_M \to \infty$ (sample size grows), $\rho_M \to \rho \in [0, 1]$.

The General Case

- Let θ_M and γ_M denote the minimizers of the expected MSE.
- Define the residuals as

$$U_i = Y_i - X_i\theta_M - z_i\gamma_M$$

- Then the proper middle of the sandwich is:

$$\mathbf{B}_V = \lim_{M\to\infty}\left[\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\mathbb{V}_{\mathbf{X}}\begin{pmatrix} X_i'U_i \\ z_i'U_i \end{pmatrix}\right]$$

- Define

$$\mathbf{B}_{ehw} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{\mathbf{X}} \left[\begin{pmatrix} X_i' U_i \\ z_i' U_i \end{pmatrix} \begin{pmatrix} X_i' U_i \\ z_i' U_i \end{pmatrix}' \right]$$

$$\mathbf{B}_E = \mathbf{B}_{ehw} - \mathbf{B}_V$$

$$= \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \left[\mathbb{E}_{\mathbf{X}} \begin{pmatrix} X_i' U_i \\ z_i' U_i \end{pmatrix} \right] \left[\mathbb{E}_{\mathbf{X}} \begin{pmatrix} X_i U_i & z_i U_i \end{pmatrix} \right]$$

Theorem: Under assumptions (a)-(e),

$$\sqrt{N}\begin{pmatrix} \hat{\theta}_{ols} - \theta_M \\ \hat{\gamma}_{ols} - \gamma_M \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{A}^{-1}(\mathbf{B}_{ehw} - \rho \cdot \mathbf{B}_E)\mathbf{A}^{-1}\right)$$

where

$$\mathbf{A} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\left[\begin{pmatrix} X_i' \\ z_i' \end{pmatrix}\begin{pmatrix} X_i & z_i \end{pmatrix}\right]$$

- Not degenerate when $\rho = 1$.
- The standard EHW case corresponds to $\rho = 0$:

$$\mathbf{A}^{-1}\mathbf{B}_{ehw}\mathbf{A}^{-1}.$$

- Difference between the EHW variance and the correct variance is positive semi-definite:

$$\rho \cdot \mathbf{A}^{-1}\mathbf{B}_{E}\mathbf{A}^{-1},\ \rho = \text{plim}(N/M)$$

- Not important if N/M is small or $\mathbb{E}_{\mathbf{X}}(W_i' U_i) \approx 0$ [$W_i = (X_i, z_i)$].

Regression Function is Correctly Specified

- The usual EHW variance formula is generally conservative with $\rho > 0$.
- Two additional assumptions ensure:

1. We consistently estimate the causal effect.
2. The asymptotic variance formula is not conservative (for any ρ).

  - ▸ Extends constant treatment effect assumption.

I. **Linearity of Potential Outcomes**: The potential outcomes satisfy

$$y_i(x) = y_i(0) + x\theta. \ \square$$

II. **Linearity of Potential Causes**: For some $K \times J$ matrix Λ_M, and for $i = 1, \ldots, M$,

$$\mathbb{E}_{\mathbf{X}}[X_i] = \Lambda_M z_i. \ \square$$

- Partial the z_i out of X_i: $\dot{X}_i = X_i - \Lambda_M z_i$.

$$\mathbf{A}_{\dot{X}} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}(\dot{X}_i' \dot{X}_i),$$

$$\mathbf{B}_{ehw,\dot{X}} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}(U_i^2 \dot{X}_i' \dot{X}_i).$$

Theorem: Under the previous assumptions, including linearity of potential outcomes and potential causes,

$$\sqrt{N}\left(\hat{\theta}_{ols} - \theta\right) \xrightarrow{d} Normal(0, \mathbf{A}_{\dot{X}}^{-1}\mathbf{B}_{ehw,\dot{X}}\mathbf{A}_{\dot{X}}^{-1}). \ \square$$

- A robust Frisch-Waugh formula.
- The asymptotic variance of $\sqrt{N}\left(\hat{\theta}_{ols} - \theta\right)$ does not depend on ρ, the ratio of the sample to the population size.
- The usual EHW variance matrix is correct for $\hat{\theta}_{ols}$.

Estimating the Asymptotic Variance

- The EWH estimator is asymptotically conservative for all parameters.
- Can legimately shrink the variance estimates.
- Attributes $\{z_i\}$ need to help explain variation in $E(\dot{X}_i' U_i)$.
- Simple approach: Use the variance estimator for stratified sampling, where functions of z_i are used in the stratification.
- See AAIW (2017) for more.

3. Cluster Sampling

- Sampling Uncertainty Approach:
  - ▸ Assumes a large population of clusters.
  - ▸ We randomly sample clusters from the population.
  - ▸ If sampling without replacement, finite population corrections are available.

- What if we observe the entire population?
- Clustering becomes a design-based problem.
- Even if we randomly sample from a population, design-based considerations are important for determining when, and at what level, to cluster.
  - Policy assignment at the cluster level.

Ex Post Clustering of a Random Sample

- Important to know when *not* to cluster.
- Job training program example: Random sampling, random assignment.
- **Do not cluster**!

$$\hat{\tau} = 0.058, \text{se}_{ehw} = 0.012$$

$$\hat{\tau} = 0.058, \text{se}_{cluster,state} = 0.031$$

$$\hat{\tau} = 0.058, \text{se}_{cluster,region} = 0.049$$

- Consider estimating the mean, μ, from a population.
- Randomly draw N observations, $\{Y_i : i = 1, \ldots, N\}$.
- Within-group means are heterogeneous.
- **The usual standard error is valid**.
- Let $\hat{U}_i = Y_i - \bar{Y}$. If we cluster,

$$\hat{\mathbb{V}}_{cluster} = N^{-2}\left(\sum_{i=1}^{N} \hat{U}_i^2 + \sum_{i=1}^{N}\sum_{h\neq i}^{N}\sum_{g=1}^{G} S_{ig}S_{hg}\hat{U}_i\hat{U}_h\right)$$

- S_{ig} are the group indicators, collect along with Y_i.

- In expected value, the expression is, roughly,

$$\mathbb{E}\left(\hat{\mathbb{V}}_{cluster}\right) \approx \frac{\sigma_Y^2}{N} + \gamma^2$$

$$\gamma^2 \equiv \sum_{g=1}^{G} \rho_g^2 \tau_g^2 \geq 0$$

$$\rho_g = P(S_{ig} = 1)$$

$$\tau_g = E(Y_i | S_{ig} = 1) - \mu$$

- Equal cluster shares, $\rho_g = 1/G$:

$$\mathbb{E}\left(\hat{\mathbb{V}}_{cluster}\right) \approx \frac{\sigma_Y^2}{N} + \frac{\eta^2}{G}$$

$$\frac{\mathbb{E}\left(\hat{\mathbb{V}}_{cluster}\right)}{\mathbb{E}\left(\hat{\mathbb{V}}_{usual}\right)} \approx 1 + \frac{\eta^2}{\sigma_Y^2} \cdot \frac{N}{G}$$

$$\eta^2 = G^{-1} \sum_{g=1}^{G} \tau_g^2 \leq \sigma_Y^2.$$

- N/G is the average number of observations per cluster.

- Formula works well in simulations, for both large G and smallish G.
- Following are random sampling ($N = 10,000$).
- $\mu = 2$, but heterogeneous means across g ($\tau_g \neq 0$).
- $G = 50$, $H = 10$.
- 1,000 replications.

Sampling SD	.1605
Usual SE	.1589
Clustered SE (*g*)	.4316
Clustered SE (*h*)	.9489
Cluster Correlation Coefficient (*g*)	.3368

Important Lesson:

- The data cannot tell us whether to cluster.
- In the previous example, the within-cluster correlation is fairly large:

$$Corr \approx 0.337$$

- But the right thing to do is not cluster!

Regression Analysis

- Similar result can be shown for regression when the uncertainty is due to sampling only.
- Nature of how covariates are distributed in population is irrelevant.
  - Can have arbitrary heterogeneity across g.
  - More heterogeneity in $\mathbb{D}(X)$ is worse for clustering.
  - Can be constant within cluster.

Simulations

- Finite population ($M = 100,000$).
- $G = 50$, $H = 10$
- Heterogeneity across clusters.
- X_i is binary.
- 10 percent random sampling: $N = 10,000$.
- 1,000 replications.
- Regress Y_i on $1, X_i$, $i = 1, \ldots, N$.

Constant Slope

	Design 1		Design 2	
	Homogeneous $\mathbb{D}(X)$		Heterogeneous $\mathbb{D}(X)$	
	OLS	FE	OLS	FE
Sampling SD	.0792	.0746	.0912	.0835
EHW SE	.0854	.0806	.0952	.0880
Clustered SE (g)	.0839	.0793	.3010	.0875
Clustered SE (h)	.0815	.0785	.6580	.0852
Cluster Corr (g)	.1169		.0128	

Heterogeneous Slope, Homogeneous $\mathbb{D}(X)$

	OLS	FE
Sampling SD	.0460	.0246
EHW SE	.0498	.0248
Clustered SE (g)	.2110	.2101
Clustered SE (h)	.4782	.4851
Cluster Corr (g)	.7572	

- Clustering is wrong and much too conservative, for both OLS and FE.

Clustering with Design Uncertainty

- What if we have obtained a random sample, but a policy is applied at a group level?
  - ▸ We might want to allow for design uncertainty.
- If we observe the entire population there is no sampling uncertainty.
- Abadie, Athey, Imbens, Wooldridge (in process, "Clustering as a Design Problem").

- In the binary treatment case:

$$\theta_{pate} = \frac{1}{M}\sum_{i=1}^{M}[y_i(1) - y_i(0)]$$

$$\theta_{sate} = \frac{1}{N}\sum_{i=1}^{M}R_i[y_i(1) - y_i(0)]$$

$$\theta_{date} = \frac{1}{M_1}\sum_{i=1}^{M}X_i y_i(1) - \frac{1}{M_0}\sum_{i=1}^{M}(1 - X_i)y_i(0)$$

- We observe θ_{date} if we sample entire population.

- Note that

$$\mathbb{E}_{\mathbf{R}}(\theta_{sate}) = \mathbb{E}_{\mathbf{X}}(\theta_{date}) = \theta_{pate}$$

$$\begin{aligned}\hat{\theta} &= \frac{1}{N_1}\sum_{i=1}^{N} R_i X_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N} R_i(1-X_i)Y_i \\ &= \bar{Y}_1 - \bar{Y}_0\end{aligned}$$

- Conditional on $\mathbf{R}$, $\hat{\theta}$ is unbiased for θ_{sate}.
- Conditional on $\mathbf{X}$, $\hat{\theta}$ is unbiased for θ_{date}.
- $\hat{\theta}$ is unconditionally unbiased for θ_{pate}.

Scenarios

1. Random Sampling, Random Assignment

- $\hat{\mathbb{V}}_{robust}$ is valid.
- $\hat{\mathbb{V}}_{cluster}$ (for any set of clusters) is not generally valid (and too conservative).

2. Random Sampling, Clustered Assignment

- X_g is assigned by cluster: all units in the cluster are treated or not.
- $\hat{\mathbb{V}}_{robust}$ is valid for θ_{date}, but need $\hat{\mathbb{V}}_{cluster}$ for θ_{pate} or θ_{sate}.
- Unclustered SE is valid for studying the effect of policy as it was implemented.
- Clustering accounts for uncertainty due to other possible assignments.

- As we draw random samples, is X_g reassigned, or fixed?
- If fixed, do not cluster. (Estimating DATE.)
- If reassigned, need to cluster. (Estimating PATE or SATE.)
- The reassigning case is technically similar to creating a group-level variable ex post, such as a peer effects variable.
  - ▸ Clustering is generally needed for the group-level variable.

	No Reassign X_g	Reassign X_g
Sampling SD	.0830	.5991
EHW SE	.0874	.0709
Clustered SE (g)	.4375	.6175
Clustered SE (h)	.3089	.5611
Cluster Corr (g)	.1171	.3768

- In very small G cases (difference-in-differences), cannot cluster.
- Donald and Lang (2007) is one solution.

3. Clustered Sampling, Random Assignment

- No clustering if interested in SATE.
- Have to cluster for PATE or DATE.

4. Random Sampling, Heterogeneous Assignment with Cluster

- Neither $\hat{\mathbb{V}}_{robust}$ nor $\hat{\mathbb{V}}_{cluster}$ is generally valid.
- "Fuzzy Clustering" in AAIW (in progress).

Summary

- The data cannot tell you whether, or at what level, to cluster.
  - Estimated within-cluster correlations are generally uninformative.
- Need to take a stand on:
  - How the data were sampled.
  - How the covariates were assigned.
  - Is an SE for a descriptive treatment effect enough?