LECTURE 4: CLUSTERING IN SAMPLING AND IN ASSIGNMENT

Guido Imbens – Stanford University

Economics 272, GSB 507, Spring 2025

OUTLINE

ROBUST AND CLUSTER-ROBUST STANDARD ERRORS

- Linaer Model (here with binary covariate)

$$Y_i = \beta_0 + \beta_1 W_i + \varepsilon_i = X_i^\top \beta + \varepsilon_i$$

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{N} X_i Y_i \right), \qquad \hat{\varepsilon} = Y_i - X_i^\top \hat{\beta}$$

- Neyman / Eicker-Huber-White / robust Standard errors

$$\hat{\mathbb{V}}^{\text{EHW}} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{N} X_i X_i^\top \hat{\varepsilon}_i^2 \right) \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1}$$

- Liang-Zeger/ cluster-robust Standard errors

$$\hat{\mathbb{V}}^{\text{LZ}} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \sum_{g=1}^{G} \left\{ \left(\sum_{i:G_i=g} X_i \hat{\varepsilon}_i \right) \left(\sum_{i:G_i=g} X_i \hat{\varepsilon}_i \right)^\top \right\} \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1}$$

ROBUST AND CLUSTER-ROBUST STANDARD ERRORS FOR THE BINARY TREATMENT CASE

- With $W_i \in \{0,1\}$ the two variance estimators for the treatment effect estimator $\hat{\tau}$ simplify to

$$\hat{\mathbb{V}}^{\mathrm{EHW}}(\hat{\tau}) = \frac{1}{\overline{W}^2(1-\overline{W})^2}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i^2(W_i-\overline{W})^2\right\}$$

$$\hat{\mathbb{V}}^{\mathrm{LZ}}(\hat{\tau}) = \frac{1}{\overline{W}^2(1-\overline{W})^2}\left\{\frac{1}{N}\sum_{g=1}^{g_k}\left(\sum_{i|G_i=g}\hat{\varepsilon}_i(W_i-\overline{W})\right)^2\right\}$$

where $\overline{W} = \sum_{i=1}^{N} W_i/N$ is the average value for the treatment.

ASYMPTOTICS TO INCLUDE BOTH CLUSTERED SAMPLING AND CLUSTERED ASSIGNMENT, FORMAL SET UP (AS LAST TIME)

- Sequence of populations, $k = 1, 2, 3, \ldots,$
- In population k
  - there are g_k clusters,
  - $n_{k,g}$ units in cluster g, $n_k = \sum_{g=1}^{g_k} n_{k,g}$ units total in population.
- Sampling:
  - Cluster g is sampled with probability q_k.
  - Units from the sampled clusters are sampled with probability p_k.
- Assignment:
  - For each cluster a (random) probability $A_{k,g}$ is drawn randomly from a distribution with mean μ_k and variance σ_k^2.
  - Units in cluster g are assigned to the treatment with probability $A_{k,g}$.
- Observed is treatment assignment $W_{k,i} \in \{0, 1\}$ and outcome $Y_{k,i} = y_{k,i}(W_{k,i})$
- Interest in population average effect $\tau = \sum_{i=1}^{n_k}(y_i(\mathrm{T}) - y_{k,i}(\mathrm{C}))/n_k$

NOTATION

- Population quantities lower case, n_k, g_k.
- Sample quantities upper case, N_k, G_k

SPECIAL CASES

- Asymptotics generally use $n_k \to \infty$, sometimes $g_k \to \infty$, sometimes $g_k = g, \forall k$.
- Random sampling of clusters from a large population of clusters, q_k is small (clustered sampling, not that common in economics, more common in survey sampling literature)
- Random sampling from a large population, $q_k = 1,\ p_k$ is small.
- Completely random assignment, $A_{k,g} = A_k \forall g$ (first two classes)
  - robust standard errors
- Clustered random assignment, $A_{k,g} \in \{0, 1\}, \forall k, g$ (last two classes)
  - cluster-robust standard errors
- How do we decide what case we are interested in given a sample (W_i, Y_i, G_i)?
  - This is not purely a statistical question, cannot be answered on the basis of the data. We cannot see whether $q_k = 1$ or $q_k < 1$ (clustered sampling). We can test whether $\sigma_k^2 = 0$ or $\sigma_k^2 > 0$ (clustered assignment).
- Challenge: what to do if $A_{k,g}$ has a distribution with positive variance that has support different from $\{0, 1\}$ (today)

EXAMPLE

- A sample from the 2000 US decennial census containing information for 2,632,838 individuals on
  - the logarithm of earnings
  - an indicator for attending college (years of education exceeding twelve years).
- 52 clusters: 50 states plus Puerto Rico and DC.
- Two regressions, in both cases with a single binary regressor (once varing at unit level, once varying at cluster level).
  - In the first regression the sole regressor is a binary variable indicating whether the fraction of individuals in the state has at least some college exceeds 0.55
  - In the second regression the sole regressor is the individual-level indicator for attending college.

VARIANCE CALCULATION BASED ON ABADIE ET AL (2023)

- Two Estimators
  - least squares estimator
  - fixed effect estimator (fixed effects for each cluster)
- four standard errors,
  - robust (EHW, eicker-huber-white)
  - cluster-robust (LZ, liang-zeger)
  - new standard error (CCV, causal cluster variance),
  - bootstrap version of new standard error (TSCB, two-stage causal bootstrap).

ROBUST, CLUSTER-ROBUST AND CAUSAL CLUSTER VARIANCE (CCV) STANDARD ERRORS, FOR THE CENSUS SAMPLE

OLS Estimator

	Est	robust	CCV	cluster robust	tscb
Ave Coll > 0.55	0.102	(0.001)	(0.031)	(0.031)	(0.031)
Some Coll	0.466	(0.001)	(0.004)	(0.027)	(0.004)

Fixed Effect Estimator

	Est	robust	CCV	cluster robust	tscb
Ave Coll > 0.55	–	–	–	–	–
Some Coll	0.457	(0.001)	(0.001)	(0.028)	(0.001)

VARIANCE CALCULATION BASED ON ABADIE ET AL (2023)

- In population k we sample clusters with probability q_k, and units with probability p_k.
- assign cluster g to treatment probability A_{gk}, with mean μ_k and variance σ_k^2, independently across clusters.
- We estimate the average effect τ_{pop} as $\hat{\tau}_{\text{pop}} = \overline{Y}_{\text{T}} - \overline{Y}_{\text{C}}$.
- Assign unit i to treatment with probability $A_{kg_{ki}}$.
- Define the residuals

$$\varepsilon_{ki}(\text{C}) \equiv y_{ki}(\text{C}) - \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}(\text{C}) \qquad \varepsilon_{ki}(\text{T}) \equiv y_{ki}(\text{T}) - \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}(\text{T})$$

$$\varepsilon_{ki} \equiv \varepsilon_{ki}(W_{ki})$$

- $g_{k,i}$ indicates which cluster a unit is from.

VARIANCE CALCULATION BASED ON ABADIE ET AL (2023)

Result in Abadie et al:

$$\sqrt{n_k}(\hat{\tau}_k - \tau_k)/v_k^{1/2} \xrightarrow{d} N(0,1),$$

where for the (general case) we get a very messy expression:

$$v_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right) \tag{1}$$

$$- p_k \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0)\right)^2 - p_k\sigma_k^2 \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}(1)}{\mu_k} + \frac{\varepsilon_{k,i}(0)}{1-\mu_k}\right)^2 \tag{2}$$

$$+ p_k(1-q_k)\frac{1}{n_k}\sum_{g=1}^{g_k}\left(\sum_{i:G_{k,i}=g}\left(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0)\right)\right)^2 \tag{3}$$

$$+ p_k\sigma_k^2 \frac{1}{n_k}\sum_{g=1}^{g_k}\left(\sum_{i:G_{k,i}=g}\left(\frac{\varepsilon_{k,i}(1)}{\mu_k} + \frac{\varepsilon_{k,i}(0)}{1-\mu_k}\right)\right)^2. \tag{4}$$

How do we unpack / interpret this expression?

THE VARIANCE UNDER RANDOM SAMPLING, RANDOM ASSIGNMENT

- Consider the case with random sampling ($q_k = 1$), and random assignment ($\sigma_k^2 = 0$). Then the variance simplifies to

$$v_k(q_k = 1, \sigma_k^2 = 0) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right) - p_k\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0)\right)^2.$$

  - The first term in this variance is what is estimated by the robust variance estimator $\mathbb{V}^{\text{EHW}}$.
  - The second term is a finite sample correction in the Neyman variance that is familiar from the literature on randomized experiments
  - This finite sample correction vanishes if either there is no heterogeneity in the treatment effects, neithe within or between clusters $(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0))$, or if the sample is a small fraction of the population $(p_k \to 0)$.

THE ROBUST VARIANCE ESTIMATOR

- robust variance estimator

$$\hat{\mathbb{V}}_k^{\text{EHW}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2}\left\{\frac{1}{N_k}\sum_{i=1}^{N_k}\hat{\varepsilon}_{k,i}^2(W_{k,i}-\overline{W}_k)^2\right\},$$

- Define the estimand corresponding to the EHW variance estimator:

$$v_k^{\text{EHW}} = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k}+\frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right),$$

(same as first term in v_k)

- Then $\hat{\mathbb{V}}_k^{\text{EHW}}$ and v_k^{EHW} are close in the sense that

$$\frac{\hat{\mathbb{V}}_k^{\text{EHW}}-v_k^{\text{EHW}}}{v_k} = o_p(1).$$

In general the difference between the estimands $v_k^{\text{EHW}} - v_k$ can be positive or negative, so the robust variance estimator can be invalid even in large samples.

THE CLUSTER ROBUST VARIANCE

$$\hat{\mathbb{V}}_k^{\text{cluster}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2} \left\{ \frac{1}{N_k} \sum_{g=1}^{g_k} \left(\sum_{i:G_{ki}=g} \hat{\varepsilon}_{ki}(W_{k,i} - \overline{W}_k) \right)^2 \right\}.$$

Define the estimand corresponding to the LZ variance estimator:

$$\begin{aligned}
v_k^{\text{cluster}} = & \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k} \right) \\
& - p_k \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0) \right)^2 - p_k \sigma_k^2 \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{\varepsilon_{k,i}(1)}{\mu_k} + \frac{\varepsilon_{k,i}(0)}{1-\mu_k} \right)^2 \\
& + p_k \frac{1}{n_k} \sum_{g=1}^{g_k} \left(\sum_{i:G_{ki}=g} \left(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0) \right) \right)^2 \\
& + p_k \sigma_k^2 \frac{1}{n_k} \sum_{g=1}^{g_k} \left(\sum_{i:G_{ki}=g} \left(\frac{\varepsilon_{k,i}(1)}{\mu_k} + \frac{\varepsilon_{k,i}(0)}{1-\mu_k} \right) \right)^2 .
\end{aligned}$$

THE CLUSTER ROBUST VARIANCE

- Then $\hat{\mathbb{V}}_k^{\text{cluster}}$ gets close to v_k^{cluster}:

$$\frac{\hat{\mathbb{V}}_k^{\text{cluster}} - v_k^{\text{cluster}}}{v_k} = \mathcal{O}_p(1).$$

- The difference $v_k^{\text{cluster}} - v_k$ is always nonnegative so that the cluster-robust variance can be conservative, but cannot underestimate the variance.
- The difference is

$$v_k^{\text{cluster}} - v_k = p_k q_k \frac{1}{n_k} \sum_{g=1}^{g_k} \left(\sum_{i:G_{ki}=g} \left(\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0) \right) \right)^2$$

coming from variation in the treatment effects $\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0)$

CHALLENGE

- If $0 < \sigma_k^2 < \mu_k(1-\mu_k)$ (so $A_{k,g}$ takes on values outside of $\{0,1\}$ and is not constant), and $pq_k = 1$, what to do?
  - Neither $\hat{\mathbb{V}}^{\text{EHW}}$ nor $\hat{\mathbb{V}}_k^{\text{cluster}}$ is appropriate
  - If σ_k^2 close to zero, variance estimator should be close to $\hat{\mathbb{V}}^{\text{EHW}}$
  - If σ_k^2 is close to $\mu_k(1-\mu_k)$, variance estimator should be close to $\hat{\mathbb{V}}_k^{\text{cluster}}$

NEW CAUSAL CLUSTER VARIANCE

- Focus on case with $q_k = p_k = 1$, all unit sampled.
- The first step is to approximate the normalized error of the least squares estimator $\widehat{\tau}_k$ by a normalized sample average over clusters:

$$\sqrt{n_k}\frac{\widehat{\tau}_k - \tau_k}{\sqrt{v_k}} = \frac{1}{\sqrt{n_k p_k}\mu_k(1-\mu_k)\sqrt{v_k}} \sum_{g=1}^{g_k} C_{k,g} + \mathcal{O}_p(1),$$

- the g_k cluster terms $C_{k,g}$, independent across clusters, are defined as

$$C_{k,g} = \sum_{i=1}^{n_k} 1\{G_{ki} = g\}\varepsilon_{k,i}(W_{k,i} - \mu_k),$$

- The cluster-robust variance estimator is approximately equal to the sum of squares of these terms:

$$\hat{\mathbb{V}}^{\text{cluster}} = \frac{1}{n_k v_k \mu_k^2 (1-\mu_k)^2} \sum_{g=1}^{g_k} C_{k,g}^2 + \mathcal{O}_p(1).$$

THE NEW CAUSAL CLUSTER VARIANCE

- Note: the sum of the expectations of $C_{k,g}$ is equal to zero, $\sum_{m=1}^{m_k} \mathbb{E}[C_{k,g}] = 0$,
- But for each cluster separately the expectation of the cluster terms is not equal to zero:

$$\mathbb{E}[C_{k,g}] = n_{k,m}\, p_k \mu_k (1 - \mu_k)(\tau_{k,g} - \tau_k) \neq 0.$$

Because $\sum_{m=1}^{m_k} \mathbb{E}[C_{k,g}] = 0$, we can replace the $C_{k,g}$ by deviations from mean $\dot{C}_{k,g} \equiv C_{k,g} - E[C_{k,g}]$, where

$$\dot{C}_{k,g} = \sum_{i=1}^{N_k} 1\{G_{ki} = g\} \left\{ \varepsilon_{ki}(W_{k,i} - \mu_k) - (\tau_{k,g} - \tau_k)\mu_k(1 - \mu_k) \right\},$$

so that

$$\sqrt{N_k} \frac{\hat{\tau}_k - \tau_k}{\sqrt{v_k}} = \frac{1}{\sqrt{n_k\, p_k} \mu_k (1 - \mu_k) \sqrt{v_k}} \sum_{g=1}^{g_k} \dot{C}_{k,g} + o_p(1).$$

THE NEW CAUSAL CLUSTER VARIANCE ESTIMATOR

- A naive way to estimate the variance of the second sum is to put in estimated counterparts:

$$\frac{1}{N_k(\hat{\mu}_k(1-\hat{\mu}_k))^2}\sum_{g=1}^{g_k}\left(\sum_{i|G_{ki}=g}\left\{\hat{\varepsilon}_{k,i}(W_{k,i}-\hat{\mu}_k)-(\widehat{\tau}_{k,g}-\widehat{\tau}_k)\hat{\mu}_k(1-\hat{\mu}_k)\right\}\right)^2$$

- The problem is that the estimation error in $\widehat{\tau}_{k,g}$ is positively correlated with the estimation error in $\sum_{i|G_{ki}=g}\hat{\varepsilon}_{k,i}(W_{k,i}-\hat{\mu}_k)$, leading to a downward bias in this variance estimator.
- To get a variance estimator with better properties we use sample splitting.

THE NEW CAUSAL CLUSTER VARIANCE

- Let $Z_{k,i} \in \{0,1\}$ be a binomial random variable with probability 1/2.
- Estimate the normalized variance for the case with $q_k = 1$ as $\hat{\mathbb{V}}_k^{\mathrm{CCV}}$, equal to

$$\frac{4}{N_k(\hat{\mu}_k(1-\hat{\mu}_k))^2}\sum_{g=1}^{g_k}\left\{\left(\sum_{i:G_{ki}=g,Z_{ki}=0}\left\{\hat{\varepsilon}_{k,i}(W_{k,i}-\hat{\mu}_k)-(\widehat{\tau}_{k,g}^1-\widehat{\tau}_k^1)\hat{\mu}_k(1-\hat{\mu}_k)\right\}\right)^2\right.$$

$$-2\sum_{i:G_{ki}=g}(1-Z_{k,i})\left\{\hat{\varepsilon}_{k,i}(W_{k,i}-\hat{\mu}_k)-(\widehat{\tau}_{k,g}^1-\widehat{\tau}_k^1)\hat{\mu}_k(1-\hat{\mu}_k)\right\}^2$$

$$\left.+(1-p_k)\sum_{g=1}^{g_k}\frac{N_{k,m}}{N_k}(\widehat{\tau}_{k,g}^1-\widehat{\tau}_k^1)^2\right\}.$$

- The factor 4 in the first term of the variance expression accounts for the fact that we only use half the observations to estimate the sum.
- The second term accounts for the double counting of the square terms.

A RESAMPLING-BASED VARIANCE ESTIMATOR

- Two-stage-cluster-bootstrap (tscb), consists of two resampling stages and a couple of additional steps:
  - Calculate for each cluster $g = 1, \ldots, g_k$, the fraction treated units, $\overline{W}_{k,g} = N_{k,g,1}/N_{k,g}$.
  - Next draw for cluster g a fraction treated $\overline{W}^b_{k,g}$ from the empirical distribution of these g_k fractions $\{\overline{W}_{k,g'}\}_{g'=1}^{g_k}$, with replacement.
  - Given the sample cluster size for this cluster, $N_{k,g}$, draw $N_{k,g} \times \overline{W}^b_{k,g}$ units with replacement from the set of $N_{k,g,1}$ treated units in this cluster.
  - Similarly draw $N_{k,g} \times (1 - \overline{W}^b_{k,g})$ units with replacement from the set of $N_{k,g,0}$ control units in this cluster. Add these $N_{k_{g1}}$ treated and $N_{k_{g0}}$ control units to the bootstrap sample.
  - We do this for all g_k clusters to create the bootstrap sample.
  - Then for each bootstrap sample we calculate the least squares and fixed effect estimators, and we use these bootstrap estimates to calculate bootstrap standard errors.

REFERENCES

- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. "When should you adjust standard errors for clustering?." *The Quarterly Journal of Economics* 138, no. 1 (2023): 1-35.