Econometrics II

Lecture 3: Inference Principles

Konrad Burchardi

Stockholm University

9th of April 2024

Plan for Today

Inference Principles: Introduction

Goal: "How certain is my estimate?"

Focus: What is standard deviation of estimator $\hat{\beta}$, $\sqrt{\mathbb{V}(\hat{\beta})}$?

$\rightarrow$ Estimator thereof is the "standard error of $\hat{\beta}$": $\sqrt{\hat{\mathbb{V}}(\hat{\beta})}$.[1]

Today: Some answers, and many questions.

Very active research area!
Basic insights I thought were true turn out to be misleading.

[1] Confusing: In statistics the standard deviation of an estimator is often called "standard error".

Plan for Today

1. Inference Principles: Introduction
2. Classic Approach: Analytic Standard Errors
3. Sampling- and Design-Based Uncertainty
4. Bootstrap
5. Randomisation Inference

Setup[2]

Suppose we have a sample of N individuals and estimate by OLS:

$$Y_i = \beta' X_i + \epsilon_i$$

where β and X_i are $k \times 1$ vectors.
We have $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. Assume $\mathbb{E}[\epsilon|X] = 0$ and denote $\Omega := \mathbb{E}[\epsilon\epsilon'|X]$.

Then:

$$\mathbb{V}(\hat{\beta}|X) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}.$$

Denote $\Omega_{ij} := Cov(\epsilon_i, \epsilon_j|X)$.

---

[2] Reading suggestions: Angrist and Pischke, Chapter 8; Hansen, Chapter 4.

Case 1: Homoskedastic Errors

Assume homoskedasticity: $\Omega_{ij} = 0, \forall i \neq j$ and $\Omega_{ii} = \sigma^2, \forall i$.
Then

$$
\begin{aligned}
\mathbb{V}_{Homosc.}(\hat{\beta}|X) &= (X'X)^{-1}(X'\Omega X)(X'X)^{-1} \\
&= (X'X)^{-1}(X'\sigma^2 I X)(X'X)^{-1} \\
&= (X'X)^{-1}(\sigma^2 X'X)(X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1}
\end{aligned}
$$

Two consistent estimators:

$$\hat{\mathbb{V}}_{HM0}(\hat{\beta}|X) = \frac{\sum_{i=1}^{N} \hat{\epsilon_i}^2}{N}(X'X)^{-1} \quad \text{or} \quad \hat{\mathbb{V}}_{HM1}(\hat{\beta}|X) = \frac{\sum_{i=1}^{N} \hat{\epsilon_i}^2}{N-k}(X'X)^{-1}.$$

Case 1: Homoskedastic Errors (Bias Correction)

Two consistent estimators:

$$\hat{\mathbb{V}}_{HM0}(\hat{\beta}|X) = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^{\,2}}{N}(X'X)^{-1} \quad \text{or} \quad \hat{\mathbb{V}}_{HM1}(\hat{\beta}|X) = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^{\,2}}{N-k}(X'X)^{-1}.$$

It can be shown that $\hat{\mathbb{V}}_{HM0}(\hat{\beta}|X)$ is biased.
Intuition: OLS overfits and hence $\hat{\epsilon}_i$ underestimates ϵ_i.

In contrast, $\hat{\mathbb{V}}_{HM1}(\hat{\beta}|X)$ is unbiased.[3] (It is the default in STATA.)
Intuition: More k, more overfitting. Turns out $N-k$ exactly right correction.

[3] For proofs check Hansen (2022), Chapters 4.11 and 4.13.

Case 2: Heteroskedastic Errors

More reasonable assumption: $\Omega_{ii} \neq \Omega_{jj}$ for at least some i, j.

$$\mathbb{V}_{Heterosc.}(\hat{\beta}|X) = (X'X)^{-1} \left(\sum_{i=1}^{N} \Omega_{ii} X_i X_i' \right) (X'X)^{-1}.$$

In case of heteroskedasticity, $\hat{\mathbb{V}}_{HM1}(\hat{\beta}|X)$ is inconsistent for $\mathbb{V}_{Heterosc.}(\hat{\beta}|X)$.[4]

Eicker-Huber-White (EHW) estimator consistent for $\mathbb{V}_{Heterosc.}(\hat{\beta}|X)$:

$$\hat{\mathbb{V}}_{EHW}(\hat{\beta}|X) = a \cdot (X'X)^{-1} \left(\sum_{i=1}^{N} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1},$$

where a is a bias correction factor.

[4] See Hansen (2022), Chapter 4.13.

Case 2: Heteroskedastic Errors (Bias Correction)

$$\hat{\mathbb{V}}_{EHW}(\hat{\beta}|X) = a \cdot (X'X)^{-1} \left(\sum_{i=1}^{N} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}.$$

Again, bias correction, **different versions**:

HC0: $a = 1$, poor performance in small samples.

HC1: $a = N/(N-k)$, ad hoc correction. STATA: `, robust`.

HC2: $(X'X)^{-1} \left(\sum_{i=1}^{N} (1-h_{ii})^{-1} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}$. STATA: `, vce(hc2)`.

HC3: $(X'X)^{-1} \left(\sum_{i=1}^{N} (1-h_{ii})^{-2} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}$. STATA: `, vce(hc3)`.

In case of interest, check also Young (2019, QJE).

Non-diagonal Ω

So far we assumed Ω was diagonal. Why might it not be?

1. **Clusters** in the data, within which ϵs are correlated:
    - Students within schools,
    - Households within villages,
    - Firms within states.

   Errors may be correlated b/c of common shocks / unobserved characteristics.

2. **Serial correlation** in ϵs
    - Dataset consists of individuals / firms / ... observed on multiple occasions.

   Errors correlated with serially correlated shocks / persistent unobserved characteristics.

Non-diagonal Ω

So far we assumed Ω was diagonal. Why might it not be?

1. **Clusters** in the data, within which ϵs are correlated:
   - Students within schools,
   - Households within villages,
   - Firms within states.

   Errors may be correlated b/c of common shocks / unobserved characteristics.

2. **Serial correlation** in ϵs
   - Dataset consists of individuals / firms / ... observed on multiple occasions.

   Errors correlated with serially correlated shocks / persistent unobserved characteristics.

Non-diagonal Ω

So far we assumed Ω was diagonal. Why might it not be?

1. **Clusters** in the data, within which ϵs are correlated:
   - Students within schools,
   - Households within villages,
   - Firms within states.

   Errors may be correlated b/c of common shocks / unobserved characteristics.

2. **Serial correlation** in ϵs
   - Dataset consists of individuals / firms / ... observed on multiple occasions.

   Errors correlated with serially correlated shocks / persistent unobserved characteristics.

Case 3: Non-diagonal Ω (Kloek-Moulton)

- Each unit is observed once and belongs to one of C clusters of equal size M, denoted by $C_i \in \{1, \ldots, C\}$.[5]
- Error structure (note: homoskedastic-like):

$$\epsilon_{ic} = \alpha_c + \varepsilon_i$$

$$\Leftrightarrow \Omega_{ij} = \begin{cases} 0 & C_i \neq C_j \\ \rho_\epsilon \sigma^2 & C_i = C_j, i \neq j \\ \sigma^2 & i = j \end{cases}$$

We say that Ω is "block diagonal"

$$\Omega = \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_C \end{bmatrix} \quad \Omega_c = \begin{bmatrix} \sigma^2 & \rho_\epsilon \sigma^2 & \cdots & \rho_\epsilon \sigma^2 \\ \rho_\epsilon \sigma^2 & \sigma^2 & \cdots & \rho_\epsilon \sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_\epsilon \sigma^2 & \rho_\epsilon \sigma^2 & \cdots & \sigma^2 \end{bmatrix}$$

[5] See Angrist and Pischke for version with heterogeneous cluster sizes.

Case 3: Non-diagonal Ω (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M-1))}_{\text{“Moulton Factor”}},$$

where ρ_X is the intra-cluster correlation of X.

Insights:

bias: Expect $\rho_\epsilon > 0$, so if $\rho_X > 0$, $\mathbb{V}_{Homosc.} <$ true variance.

$\rho_\epsilon = 1$: If other covariates constant in cluster, adding new observations adds no new information.

$\rho_X = 0$: Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.

$\rho_X = 1$: Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...

clusters: More severe with fewer clusters (big M given N).

Case 3: Non-diagonal Ω (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M-1))}_{\text{“Moulton Factor”}},$$

where ρ_X is the intra-cluster correlation of X.

Insights:

- bias: Expect $\rho_\epsilon > 0$, so if $\rho_X > 0$, $\mathbb{V}_{Homosc.} <$ true variance.
- $\rho_\epsilon = 1$: If other covariates constant in cluster, adding new observations adds no new information.
- $\rho_X = 0$: Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.
- $\rho_X = 1$: Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...
- \# clusters: More severe with fewer clusters (big M given N).

Case 3: Non-diagonal Ω (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M-1))}_{\text{“Moulton Factor”}},$$

where ρ_X is the intra-cluster correlation of X.

Insights:

bias: Expect $\rho_\epsilon > 0$, so if $\rho_X > 0$, $\mathbb{V}_{Homosc.} <$ true variance.

$\rho_\epsilon = 1$: If other covariates constant in cluster, adding new observations adds no new information.

$\rho_X = 0$: Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.

$\rho_X = 1$: Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...

clusters: More severe with fewer clusters (big M given N).

Case 3: Non-diagonal Ω (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M-1))}_{\text{“Moulton Factor”}},$$

where ρ_X is the intra-cluster correlation of X.

Insights:

- bias: Expect $\rho_\epsilon > 0$, so if $\rho_X > 0$, $\mathbb{V}_{Homosc.} <$ true variance.
- $\rho_\epsilon = 1$: If other covariates constant in cluster, adding new observations adds no new information.
- $\rho_X = 0$: Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.
- $\rho_X = 1$: Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...
- # clusters: More severe with fewer clusters (big M given N).

Case 4: Non-diagonal Ω (Liang-Zeger)

Clustered errors typically estimated assuming more general error structure:

- Let X_c correspond to the submatrix of X with $C_i = c$.
- Allow for unrestricted Ω_{ij} **within clusters**.
- Impose $\Omega_{ij} = 0$ for $C_i \neq C_j$.

Then:

$$\mathbb{V}_{ClustersGeneral}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{c=1}^{C} X_c' \Omega_c X_c \right) (X'X)^{-1}$$

$$\hat{\mathbb{V}}_{LZ}(\hat{\beta}) = a \cdot (X'X)^{-1} \left(\sum_{c=1}^{C} X_c' \hat{\epsilon}_c \hat{\epsilon}_c' X_c \right) (X'X)^{-1}$$

$\hat{\mathbb{V}}_{LZ}(\hat{\beta})$ is consistent for $\mathbb{V}_{ClustersGeneral}(\hat{\beta})$ (as $C \rightarrow \infty$), and a is bias correction.

STATA: `cluster(cluster_id)` or `vce(cluster cluster_id)`, with $a = \frac{N-1}{N-k}\frac{C}{C-1}$.

"Classic" Advice: When to Cluster?

"Classic" recommendations:

- Cluster if there could be intra-cluster correlation in the error term.
- Compare robust and clustered standard errors, and pick the bigger ones: If clustering increases the standard errors then it is conservative to do it, if not then no harm done.
- Cluster at the highest level, subject to having "sufficiently many" clusters.

I am afraid those recommendations might not age well, see later.

Case 5: Non-diagonal Ω (Serial Correlation)

Often units are observed on multiple occasions over time.

- Typical case: panel data,
  - e.g. individuals in different states in annual tax data,
  - e.g. schools pre/post education reform,
  - e.g. an individual's sequence of decisions in a lab experiment.
- Serially correlated shocks or unobservables: correlation between the residuals.
- Conceptually very similar to correlation between disturbances within clusters.
- There exist variance estimators designed for serial correlation (Newey-West).
- Common to just cluster at the unit level or higher (e.g. person, state, school) which allows for more general variance-covariance structure.

Bertrand, Duflo, Mullainathan (QJE, 2004)

"How Much Should We Trust Difference-In-Difference Estimates?"

Bertrand et al. (2004) focus on the case of D-in-D estimation, with a treatment that affects some units (e.g. states) at some point in time.

Influential: by far Esther Duflo's most cited paper!

- Outcomes within a state correlated over time, so over-time observations are not independent measures of state.
- Show that failing to correct for serial correlation leads to over-rejection of the null of no effect.
- Clustering performs well with "sufficiently many" clusters.

Popularised clustering.

Plan for Today

Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:

Where is uncertainty about estimate coming from?

Think about some scenarios:

1. Estimate is average age in this room...

   ...and you have data on age of all of us.

2. Estimate is average age in this room...

   ...and you have data on age of randomly selected 5 of us.

3. Estimate is effect of treatment D for those in this room...

   ...and you have data on D and Y for all of us.

Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:
Where is uncertainty about estimate coming from?

Think about some scenarios:

1. Estimate is average age in this room...

   ...and you have data on age of all of us.

2. Estimate is average age in this room...

   ...and you have data on age of randomly selected 5 of us.

3. Estimate is effect of treatment D for those in this room...

   ...and you have data on D and Y for all of us.

Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:
Where is uncertainty about estimate coming from?

Think about some scenarios:

1. Estimate is average age in this room...

   ...and you have data on age of all of us.

2. Estimate is average age in this room...

   ...and you have data on age of randomly selected 5 of us.

3. Estimate is effect of treatment D for those in this room...

   ...and you have data on D and Y for all of us.

Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:

Where is uncertainty about estimate coming from?

Think about some scenarios:

1. Estimate is average age in this room...

   ...and you have data on age of all of us.

2. Estimate is average age in this room...

   ...and you have data on age of randomly selected 5 of us.

3. Estimate is effect of treatment D for those in this room...

   ...and you have data on D and Y for all of us.

Sources of Uncertainty

Something is confusing!

- What type of uncertainty do we express with standard standard errors?
- When having a sample of size $N = 100$, our standard errors do not take into account whether it is drawn from a population of 1.000, or 1.000.000.
- What on earth are the errors?
- What does it mean that "the Xs are fixed"?
- ...

Guido Imbens talks about the status quo when presenting his current work like Steve Jobs about the blackberry when presenting the iPhone: "Bäääää!"

Abadie, Athey, Imbens and Wooldridge (2020)

Abadie et al. (2020) distinguish between sampling-based uncertainty and design-based uncertainty.

They propose that it is useful to think (again) about:

1. the estimand of interest,
2. the population of interest,
3. the sampling process, and
4. the assignment process.

Set-Up:

- Finite population consisting of n units.
- Each unit characterized by (Y_i, X_i).
- Whether unit i is in the sample is indicated by $R_i \in \{0, 1\}$.

Sampling-Based Uncertainty

Consider:

- estimand which is a function of the full set $\{(Y_i, X_i)\}_{i=1}^n$, and
- estimator which is a function of the observed data $\{(R_i, R_i Y_i, R_i X_i)\}_{i=1}^n$.

$\rightarrow$ Uncertainty about estimand arises when we observe the values (Y_i, X_i) only for sample, i.e. subset of population!

$\rightarrow$ Sampling-based inference uses information about the sampling process that determines $\{R_i\}_{i=1}^n$ to assess variability of estimators across different samples.

Sampling-Based Uncertainty

Table 1: Sampling-based Uncertainty (✓ is observed, ? is missing)

Unit	Actual Sample Y_i	Actual Sample Z_i	Actual Sample R_i	Alternative Sample I Y_i	Alternative Sample I Z_i	Alternative Sample I R_i	Alternative Sample II Y_i	Alternative Sample II Z_i	Alternative Sample II R_i	...
1	✓	✓	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	✓	✓	1	✓	✓	1	...
4	?	?	0	✓	✓	1	?	?	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	✓	1	?	?	0	?	?	0	...

From Abadie et al. (2020).

Different Scenario:

- Observe for each unit in the population the value of one of two potential outcomes, $Y_i^*(1)$ or $Y_i^*(0)$, but not both.
- Which potential outcome is observed is indicated by $X_i \in \{0, 1\}$.[6]
- Denote the observed outcome as $Y_i = Y_i^*(X_i)$.

[6] Otherwise we will use D_i as treatment indicator in this course. But here the main point is that we are talking about an explanatory variable, and those we called X_i today.

Design-Based Uncertainty

Consider:

- estimand which is a function of the full set $\{(Y_i^*(1), Y_i^*(0), X_i)\}_{i=1}^n$, and
- estimator which is a function of the observed data $\{(Y_i, X_i)\}_{i=1}^n$.

$\rightarrow$ Uncertainty about estimand arises because different observations are assigned to treatment across different realisations of the assignment.

$\rightarrow$ Design-based inference uses information about the assignment process that determines $\{X_i\}_{i=1}^n$ to assess the variability of the estimator.

Design-Based Uncertainty

Table 2: DESIGN-BASED UNCERTAINTY (✓ IS OBSERVED, ? IS MISSING)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	...
1	✓	?	1	✓	?	1	?	✓	0	...
2	?	✓	0	?	✓	0	?	✓	0	...
3	?	✓	0	✓	?	1	✓	?	1	...
4	?	✓	0	?	✓	0	✓	?	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	?	1	?	✓	0	?	✓	0	...

From Abadie et al. (2020).

Abadie, Athey, Imbens and Wooldridge (2020)

Estimands

$\mathbf{Y}$, $\mathbf{Y}^*(1)$, $\mathbf{Y}^*(0)$, $\mathbf{R}$, $\mathbf{X}$ stacked vectors of corresponding unit-level variables.

Classification of Estimands:

Descriptive Estimand: An estimand which can be written as a function of $(\mathbf{Y}, \mathbf{X})$, free of dependence on $\mathbf{R}$ and on potential outcomes beyond the realized outcome.

Causal Estimand: An estimand that depends on potential outcomes $\mathbf{Y}^*(1)$, $\mathbf{Y}^*(0)$.

Abadie, Athey, Imbens and Wooldridge (2020)

Consider three estimands:

$$\theta^{\text{sampling}}(\mathbf{Y}, \mathbf{X}) = \frac{1}{n_1} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - X_i) Y_i$$

$$\theta^{\text{design}}(\mathbf{Y}^*(1), \mathbf{Y}^*(0), \mathbf{R}) = \frac{1}{N} \sum_{i=1}^{n} R_i (Y_i^*(1) - Y_i^*(0))$$

$$\theta^{\text{causal}}(\mathbf{Y}^*(1), \mathbf{Y}^*(0)) = \frac{1}{n} \sum_{i=1}^{n} (Y_i^*(1) - Y_i^*(0)),$$

where n_0 and n_1 refer to the number of units in the population who are untreated and treated, respectively, and N_0 and N_1 refer to the sample similarly.

Consider the difference-in-sample-means estimator (OLS of Y_i on X_i and constant):

$$\hat{\theta} = \frac{1}{N_1} \sum_{i=1}^{n} R_i X_i Y_i - \frac{1}{N_0} \sum_{i=1}^{n} R_i (1 - X_i) Y_i.$$

Abadie, Athey, Imbens and Wooldridge (2020)

Estimator

Assume random sampling and random assignment.

With appropriate conditioning, the $\hat{\theta}$ estimator is unbiased for each estimand:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{R}}[\hat{\theta}|\mathbf{X}, N_1, N_0] &= \theta^{\text{sampling}} \\
\mathbb{E}_{\mathbf{X}}[\hat{\theta}|\mathbf{R}, N_1, N_0] &= \theta^{\text{design}} \\
\mathbb{E}_{\mathbf{X},\mathbf{R}}[\hat{\theta}|N_1, N_0] &= \theta^{\text{total}}
\end{aligned}
$$

Interpretation of conditioning:

- Considering randomness of **R** only gives sampling-based uncertainty.
- Considering randomness of **X** only gives design-based uncertainty.
- Not conditioning accounts for both types of uncertainty.

Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

Finally, we can write out the variances of our estimator for each estimand:

$$
\begin{aligned}
V^{\text{sampling}} &= \mathbb{E}_{\mathbf{X}}[\text{Var}_{\mathbf{R}}(\hat{\theta}|\mathbf{X}, N_1, N_0)|N_1, N_0] &&= \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right) \\
V^{\text{design}} &= \mathbb{E}_{\mathbf{R}}[\text{Var}_{\mathbf{X}}(\hat{\theta}|\mathbf{R}, N_1, N_0)|N_1, N_0] &&= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1} \\
V^{\text{total}} &= \text{Var}_{\mathbf{X},\mathbf{R}}(\hat{\theta}|\mathbf{X}, N_1, N_0) &&= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1},
\end{aligned}
$$

where denote with S_0^2, S_1^2, and S_θ^2 the population variance of $Y_i^*(0)$, $Y_i^*(1)$ and the treatment effect $Y_i^*(1) - Y_i^*(0)$.[7] [8]

[7] To arrive at the former two expressions we take expectations over the conditional variances.

[8] For proofs check the supplementary material to the paper, and also Imbens and Rubin (2015), Chapter 6.

Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$V^{\text{sampling}} = \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right)$$

$$V^{\text{design}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}$$

$$V^{\text{total}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}$$

1. For fixed N_0 and N_1, if $n_0, n_1 \to \infty$, the total and sampling variance are equal.

   $\to$ All uncertainty comes from randomness in sampling.

Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$V^{\text{sampling}} = \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right)$$

$$V^{\text{design}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}$$

$$V^{\text{total}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}$$

2. Both when the estimand is $\theta^{\text{descriptive}}$ or θ^{causal}, ignoring finite population leads to overstatement of variance, but not for $\theta^{\text{causal, sample}}$.

Intuition?

Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$
\begin{aligned}
V^{\text{sampling}} &= \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right) \\
V^{\text{design}} &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1} \\
V^{\text{total}} &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}
\end{aligned}
$$

3. The expectation of the Eicker-Huber-White estimator is $\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$.

$\rightarrow$ Generally over-estimates variance for well-defined estimand.

$\rightarrow$ Eicker-Huber-White estimator is assuming infinite super-population!

Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$
\begin{aligned}
V^{\text{sampling}} &= \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right) \\
V^{\text{design}} &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1} \\
V^{\text{total}} &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}
\end{aligned}
$$

4. Problem: Unclear how to estimate S_θ^2!

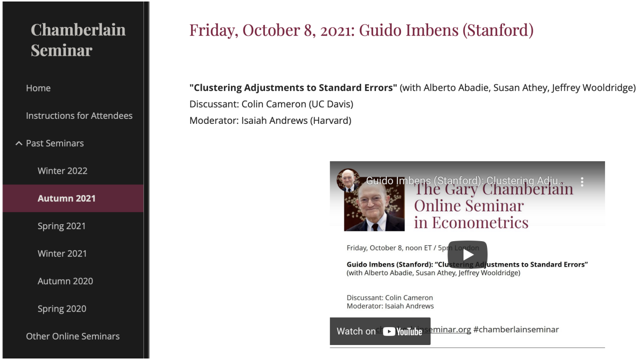
$\rightarrow$ Eicker-Huber-White estimator implicitly sets it to 0.

$\rightarrow$ Check paper for approaches.

Abadie, Athey, Imbens and Wooldridge

"When Should You Adjust Standard Errors for Clustering?"

What does all of this imply for clustering?



https://www.chamberlainseminar.org/past-seminars/autumn-2021

https://academic.oup.com/qje/article/138/1/1/6750017

Plan for Today

Bootstrap

Conventional econometrics: *Infer* the distribution of a statistic, f (e.g., t-statistic)

- calculated from a sample with empirical distribution F_1
- drawn from a infinite population with distribution F_0.

Call this the distribution of $f(F_1|F_0)$.

The Bootstrap: *Estimates* the distribution of $f(F_1|F_0)$

- by drawing random samples F_2 (**with replacement**) from F_1,
- and calculate the statistic f each time.
- If f is a smooth function of the data, then $f(F_2|F_1) \rightarrow_d f(F_1|F_0)$.

Intuition: treat sample distribution F_1 *as though it were the population distribution.*

Bootstrap

Conventional econometrics: *Infer* the distribution of a statistic, f (e.g., t-statistic)

- calculated from a sample with empirical distribution F_1
- drawn from a infinite population with distribution F_0.

Call this the distribution of $f(F_1|F_0)$.

The Bootstrap: *Estimates* the distribution of $f(F_1|F_0)$

- by drawing random samples F_2 (**with replacement**) from F_1,
- and calculate the statistic f each time.
- If f is a smooth function of the data, then $f(F_2|F_1) \rightarrow_d f(F_1|F_0)$.

Intuition: treat sample distribution F_1 *as though it were the population distribution.*

Bootstrap

Some Remarks:

- Sometimes analytic errors are not available, or hard to compute.
  (For example when your regression includes "generated regressors".)
- "Asymptotic refinement": can sometimes get closer to the true finite-sample distribution than asymptotic approximations.
  $\rightarrow$ Requires the bootstrapped statistics to be asymptotically pivotal.
- Bootstrap "feels" like it is addressing sampling uncertainty. But Abadie et al. (2020) clarify in their setting the expectation of the bootstrapped variance equals the Eicker-Huber-White estimator.

Bootstrap

Some Remarks:

- Sometimes analytic errors are not available, or hard to compute.
  (For example when your regression includes "generated regressors".)
- "Asymptotic refinement": can sometimes get closer to the true finite-sample distribution than asymptotic approximations.
  $\rightarrow$ Requires the bootstrapped statistics to be asymptotically pivotal.
- Bootstrap "feels" like it is addressing sampling uncertainty. But Abadie et al. (2020) clarify in their setting the expectation of the bootstrapped variance equals the Eicker-Huber-White estimator.

Bootstrap

Some Remarks:

- Sometimes analytic errors are not available, or hard to compute.
  (For example when your regression includes "generated regressors".)
- "Asymptotic refinement": can sometimes get closer to the true finite-sample distribution than asymptotic approximations.
  $\rightarrow$ Requires the bootstrapped statistics to be asymptotically pivotal.
- Bootstrap "feels" like it is addressing sampling uncertainty. But Abadie et al. (2020) clarify in their setting the expectation of the bootstrapped variance equals the Eicker-Huber-White estimator.

Bootstrap

Different approaches:

1. "Pairs bootstrap" or "nonparametric bootstrap":

   Repeatedly sample (with replacement) N observations from data.

2. "Parametric bootstrap":

   Keep the Xs fixed, but generate a new dependent variable by resampling from the distribution of residuals $\hat{e}$. (Bad if there is heteroscedasticity).

3. "Wild bootstrap":

   Hold Xs fixed, generate new depend. variable $y_i = X_i'\hat{\beta} \pm \hat{\epsilon}_i$ with probability 1/2.

4. "Block bootstrap":

   If there are clusters in the data, you need to resample *whole clusters* (with replacement), to preserve the correlation structure. E.g. for wild bootstrap, all observations within a cluster get $+\hat{\epsilon}$ or $-\hat{\epsilon}$.

x

Plan for Today

Randomisation Inference[9]

Long-known approach to design-based uncertainty:

- Under **sharp null hypothesis** (e.g. $\theta = 0\ \forall i$), we know $Y_i^*(1)$ *and* $Y_i^*(0)$ for all i.
- Can create many / all alternative assignments, given assignment mechanism, and recalculate $\hat{\beta}$ or test statistic each time.
- Gives exact, finite sample distribution of $\hat{\beta}$ or test statistic!
- No assumptions on disturbances!
- Downside: Allows to test sharp hypotheses only.

[9] Check Imbens and Rubin (2015), Chapter 5.

Questions?

References

1. Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2020. “Sampling-Based versus Design-Based Uncertainty in Regression Analysis”. *Econometrica* 88:1, 265–296.
2. Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2023. “When Should You Adjust Standard Errors for Clustering?”. *Quarterly Journal of Economics* 138:1, 1–35.
3. Angrist, Joshua and Jörn-Steffen Pischke. 2009. “Mostly Harmless Econometrics”. Princeton University Press.
4. Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. “How Much Should We Trust Differences-In-Differences Estimates?”. *The Quarterly Journal of Economics* 119:1, 249–275.
5. Hansen, Bruce E. 2022. “Econometrics”. Princeton University Press. Downloaded at `https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf`.
6. Young, Alwyn. 2019. “Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.” *The Quarterly Journal of Economics* 134, 557–598.