Lecture 7: Robust Inference I

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

February 24, 2021

Robust inference unit:

- Today: conceptual issues and analytical approaches.
- Next time: bootstrap and permutation.

- Robust inferential procedure: hypothesis tests and intervals reject or cover, respectively, at stated rates (e.g., 95%) under a wide range of data distributions.
- Wider range of data distributions & closer to stated rates $\Rightarrow$ more "robust."

- Robust inferential procedure: hypothesis tests and intervals reject or cover, respectively, at stated rates (e.g., 95%) under a wide range of data distributions.
- Wider range of data distributions & closer to stated rates $\Rightarrow$ more "robust."
- E.g., suppose
  - coverage rate $1-\alpha$,
  - target parameter θ
  - sample of size N, $S_N \sim P_0$, where P_0 is the probability law for the data.

- Robust inferential procedure: hypothesis tests and intervals reject or cover, respectively, at stated rates (e.g., 95%) under a wide range of data distributions.
- Wider range of data distributions & closer to stated rates $\Rightarrow$ more "robust."
- E.g., suppose
  - coverage rate $1-\alpha$,
  - target parameter θ
  - sample of size N, $S_N \sim P_0$, where P_0 is the probability law for the data.
  - Then, suppose an estimator $\hat{\theta}_N(S_N)$, and
  - a mapping $C_N(\cdot)$ that returns a confidence interval such that

$$|\Pr[\theta \in C_N(\hat{\theta}_N(S_N))] - (1-\alpha)| < \varepsilon,$$

  for all $P_0 \in \mathscr{P}_0$.

- Robust inferential procedure: hypothesis tests and intervals reject or cover, respectively, at stated rates (e.g., 95%) under a wide range of data distributions.
- Wider range of data distributions & closer to stated rates $\Rightarrow$ more "robust."
- E.g., suppose
  - coverage rate $1-\alpha$,
  - target parameter θ
  - sample of size N, $S_N \sim P_0$, where P_0 is the probability law for the data.
  - Then, suppose an estimator $\hat{\theta}_N(S_N)$, and
  - a mapping $C_N(\cdot)$ that returns a confidence interval such that

    $$|\Pr[\theta \in C_N(\hat{\theta}_N(S_N))] - (1-\alpha)| < \varepsilon,$$

    for all $P_0 \in \mathscr{P}_0$.
  - For small ε, the larger the family $\mathscr{P}_0$, the more robust is $C_N(\hat{\theta}_N(S_N))$.

Inference challenges

1. Under our frequentist framework, we rely on *asymptotic* reference distributions for testing and intervals:

Inference challenges

1. Under our frequentist framework, we rely on *asymptotic* reference distributions for testing and intervals:
   - t or Normal distribution for individual coefficients, F or χ^2 distribution for multiple coefficients, etc.
   - We like this approach because interval coverage and decision error rates are assured to approach the nominal rate (e.g., 95% coverage, 5% error rate) as N grows.

   We never reach "asymptopia" however, and so we need to attend to finite sample problems.

Inference challenges

1. Under our frequentist framework, we rely on *asymptotic* reference distributions for testing and intervals:
   - t or Normal distribution for individual coefficients, F or χ^2 distribution for multiple coefficients, etc.
   - We like this approach because interval coverage and decision error rates are assured to approach the nominal rate (e.g., 95% coverage, 5% error rate) as N grows.

   We never reach "asymptopia" however, and so we need to attend to finite sample problems.
2. Sometimes data are not independent—that is, there is clustering in the way treatments are assigned or outcomes observed. How can we account for this in a way that is robust?

Inference challenges

1. Under our frequentist framework, we rely on *asymptotic* reference distributions for testing and intervals:
   - t or Normal distribution for individual coefficients, F or χ^2 distribution for multiple coefficients, etc.
   - We like this approach because interval coverage and decision error rates are assured to approach the nominal rate (e.g., 95% coverage, 5% error rate) as N grows.

   We never reach "asymptopia" however, and so we need to attend to finite sample problems.
2. Sometimes data are not independent—that is, there is clustering in the way treatments are assigned or outcomes observed. How can we account for this in a way that is robust?
3. With finite samples, clustering, or other difficulties, exact expressions for variance can be intractable, and asymptotic approximations may fail. What alternative, robust procedures are available?

Finite Samples

- Refs: Lin (2013), Imbens & Kolesar (2016), Puteovsky & Tipton (2018), and Samii & Aronow (2012).

- Refs: Lin (2013), Imbens & Kolesar (2016), Puteovsky & Tipton (2018), and Samii & Aronow (2012).
- Suppose the goal is robust inference for $\hat{\rho} = \bar{Y}_1 - \bar{Y}_0$. (results generalize.)

- Refs: Lin (2013), Imbens & Kolesar (2016), Puteovsky & Tipton (2018), and Samii & Aronow (2012).
- Suppose the goal is robust inference for $\hat{\rho} = \bar{Y}_1 - \bar{Y}_0$. (results generalize.)
- Suppose a small sample size—e.g., less than 40 units.

- Refs: Lin (2013), Imbens & Kolesar (2016), Puteovsky & Tipton (2018), and Samii & Aronow (2012).
- Suppose the goal is robust inference for $\hat{\rho} = \bar{Y}_1 - \bar{Y}_0$. (results generalize.)
- Suppose a small sample size—e.g., less than 40 units.
- Even for seemingly simple problems, inference can be analytically complicated:
  - e.g., suppose you want to test for a mean difference between two normally distributed variables with different variances.
  - No finite sample pivotal statistic exists ("Behrens-Fisher problem") and so *inferential approximations* are necessary.

- Refs: Lin (2013), Imbens & Kolesar (2016), Puteovsky & Tipton (2018), and Samii & Aronow (2012).
- Suppose the goal is robust inference for $\hat{\rho} = \bar{Y}_1 - \bar{Y}_0$. (results generalize.)
- Suppose a small sample size—e.g., less than 40 units.
- Even for seemingly simple problems, inference can be analytically complicated:
  - e.g., suppose you want to test for a mean difference between two normally distributed variables with different variances.
  - No finite sample pivotal statistic exists ("Behrens-Fisher problem") and so *inferential approximations* are necessary.
- Robust inference usually involves approximations based on:

- Refs: Lin (2013), Imbens & Kolesar (2016), Puteovsky & Tipton (2018), and Samii & Aronow (2012).
- Suppose the goal is robust inference for $\hat{\rho} = \bar{Y}_1 - \bar{Y}_0$. (results generalize.)
- Suppose a small sample size—e.g., less than 40 units.
- Even for seemingly simple problems, inference can be analytically complicated:
  - e.g., suppose you want to test for a mean difference between two normally distributed variables with different variances.
  - No finite sample pivotal statistic exists ("Behrens-Fisher problem") and so *inferential approximations* are necessary.
- Robust inference usually involves approximations based on:
  1. A robust standard error measure for $\hat{\rho}$, $\widehat{s.e.}(\hat{\rho})$.
  2. Relating $t = \hat{\rho}/\widehat{s.e.}(\hat{\rho})$ to an approximate reference distribution that (over-)compensates for finite sample departures from the asymptotic distribution.

Getting a decent standard error is not so bad:

Getting a decent standard error is not so bad:

- Suppose a randomized experiment. (With an observational study, we proceed "as-if" it is an experiment.)

Getting a decent standard error is not so bad:

- Suppose a randomized experiment. (With an observational study, we proceed "as-if" it is an experiment.)
- *Exact* sampling+randomization variance for $\hat{\rho}$ is

$$\mathrm{Var}[\hat{\rho}] = \frac{\sigma^2_{Y_1}}{n_1} + \frac{\sigma^2_{Y_0}}{n_0}$$

Getting a decent standard error is not so bad:

- Suppose a randomized experiment. (With an observational study, we proceed "as-if" it is an experiment.)
- *Exact* sampling+randomization variance for $\hat{\rho}$ is

$$\text{Var}\left[\hat{\rho}\right] = \frac{\sigma^2_{Y_1}}{n_1} + \frac{\sigma^2_{Y_0}}{n_0}$$

- A direct analogue estimator for this is,

$$\hat{V}_{ehw} = \frac{\frac{1}{n_1}\sum_{i:D_i=1}(Y_i - \bar{Y}_1)^2}{n_1} + \frac{\frac{1}{n_0}\sum_{i:D_i=0}(Y_i - \bar{Y}_0)^2}{n_0}$$

With OLS, this is the Eicker-Huber-White estimator.

Getting a decent standard error is not so bad:

- Suppose a randomized experiment. (With an observational study, we proceed "as-if" it is an experiment.)
- *Exact* sampling+randomization variance for $\hat{\rho}$ is
$$\mathrm{Var}\left[\hat{\rho}\right] = \frac{\sigma^2_{Y_1}}{n_1} + \frac{\sigma^2_{Y_0}}{n_0}$$
- A direct analogue estimator for this is,
$$\hat{V}_{ehw} = \frac{\frac{1}{n_1}\sum_{i:D_i=1}(Y_i - \bar{Y}_1)^2}{n_1} + \frac{\frac{1}{n_0}\sum_{i:D_i=0}(Y_i - \bar{Y}_0)^2}{n_0}$$
With OLS, this is the Eicker-Huber-White estimator.
- By sampling theory, this estimator is biased. Unbiasedness requires a modest finite sample/degrees of freedom correction:
$$\hat{V}_{HC2} = \frac{\frac{1}{n_1-1}\sum_{i:D_i=1}(Y_i - \bar{Y}_1)^2}{n_1} + \frac{\frac{1}{n_0-1}\sum_{i:D_i=0}(Y_i - \bar{Y}_0)^2}{n_0}$$
With OLS, this is called the "HC2" estimator.

Getting a decent standard error is not so bad:

- Suppose a randomized experiment. (With an observational study, we proceed "as-if" it is an experiment.)
- *Exact* sampling+randomization variance for $\hat{\rho}$ is

$$\mathrm{Var}\left[\hat{\rho}\right] = \frac{\sigma^2_{Y_1}}{n_1} + \frac{\sigma^2_{Y_0}}{n_0}$$

- A direct analogue estimator for this is,

$$\hat{V}_{ehw} = \frac{\frac{1}{n_1}\sum_{i:D_i=1}(Y_i - \bar{Y}_1)^2}{n_1} + \frac{\frac{1}{n_0}\sum_{i:D_i=0}(Y_i - \bar{Y}_0)^2}{n_0}$$

  With OLS, this is the Eicker-Huber-White estimator.
- By sampling theory, this estimator is biased. Unbiasedness requires a modest finite sample/degrees of freedom correction:

$$\hat{V}_{HC2} = \frac{\frac{1}{n_1-1}\sum_{i:D_i=1}(Y_i - \bar{Y}_1)^2}{n_1} + \frac{\frac{1}{n_0-1}\sum_{i:D_i=0}(Y_i - \bar{Y}_0)^2}{n_0}$$

  With OLS, this is called the "HC2" estimator.
- Degrees of freedom adjustment increases in number of regressors.

Implementing this is also simple:

- You can ask for "HC2" in Stata (`vce(hc2)`) or R (using the `sandwich` package).
- Stata's `, robust` command uses $\hat{V}_{ehw}$ but then applies a different degrees of freedom adjustment.
- As the sample size gets larger, there should be no appreciable difference.

The second step is more complicated

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho}/\widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho} / \widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.
- This is motivated by two considerations:

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho}/\widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.
- This is motivated by two considerations:
    1. If the population residuals are in fact normal with equal variances, this is the *exact* finite sample distribution for t (and so under the null hypothesis, t is pivotal).

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho}/\widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.
- This is motivated by two considerations:
  1. If the population residuals are in fact normal with equal variances, this is the *exact* finite sample distribution for t (and so under the null hypothesis, t is pivotal).
  2. Otherwise, it is a blunt way to try to account for departures from normality:

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho} / \widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.
- This is motivated by two considerations:
  1. If the population residuals are in fact normal with equal variances, this is the *exact* finite sample distribution for t (and so under the null hypothesis, t is pivotal).
  2. Otherwise, it is a blunt way to try to account for departures from normality:
     - When the population residuals are not normal, the exact finite sample distribution is typically intractable (though for, e.g., binomial outcomes, one can derive it).

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho}/\widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.
- This is motivated by two considerations:
  1. If the population residuals are in fact normal with equal variances, this is the *exact* finite sample distribution for t (and so under the null hypothesis, t is pivotal).
  2. Otherwise, it is a blunt way to try to account for departures from normality:
     - When the population residuals are not normal, the exact finite sample distribution is typically intractable (though for, e.g., binomial outcomes, one can derive it).
     - All we know is that the asymptotic distribution is normal.

The second step is more complicated

- Standard practice is to relate $t = \hat{\rho}/\widehat{s.e.}(\hat{\rho})$ to the t_{n-k} distribution.
- This is motivated by two considerations:
  1. If the population residuals are in fact normal with equal variances, this is the *exact* finite sample distribution for t (and so under the null hypothesis, t is pivotal).
  2. Otherwise, it is a blunt way to try to account for departures from normality:
     - When the population residuals are not normal, the exact finite sample distribution is typically intractable (though for, e.g., binomial outcomes, one can derive it).
     - All we know is that the asymptotic distribution is normal.
     - Using t_{n-k} instead "fattens the tails" of our reference distribution to account for finite sample departures from normality.

In certain cases, this may not be robust enough:

In certain cases, this may not be robust enough:

- Suppose n_1 is very large but n_0 is very small.

In certain cases, this may not be robust enough:

- Suppose n_1 is very large but n_0 is very small.
- Then, $\bar{Y}_1$ will be very precisely estimated, but $\bar{Y}_0$ will be imprecisely estimated.

In certain cases, this may not be robust enough:

- Suppose n_1 is very large but n_0 is very small.
- Then, $\bar{Y}_1$ will be very precisely estimated, but $\bar{Y}_0$ will be imprecisely estimated.
- That being the case, using $n-k=(n_1+n_0)-k$ as the degrees of freedom adjustment would overstate the stability of the $\hat{\rho}$ sampling+randomization distribution.

In certain cases, this may not be robust enough:

- Suppose n_1 is very large but n_0 is very small.
- Then, $\bar{Y}_1$ will be very precisely estimated, but $\bar{Y}_0$ will be imprecisely estimated.
- That being the case, using $n-k=(n_1+n_0)-k$ as the degrees of freedom adjustment would overstate the stability of the $\hat{\rho}$ sampling+randomization distribution.
- The correct degrees of freedom adjustment ought to be closer to n_0-k.

In certain cases, this may not be robust enough:

- Suppose n_1 is very large but n_0 is very small.
- Then, $\bar{Y}_1$ will be very precisely estimated, but $\bar{Y}_0$ will be imprecisely estimated.
- That being the case, using $n - k = (n_1 + n_0) - k$ as the degrees of freedom adjustment would overstate the stability of the $\hat{\rho}$ sampling+randomization distribution.
- The correct degrees of freedom adjustment ought to be closer to $n_0 - k$.
- A way to account for the consequences of such skew is the Welch-Satterthwaite degrees of freedom approximation, which derives a degrees of freedom adjustment for normal data but $n_0 \neq n_1$.
- Lin (2013), Imbens & Kolesar (2016), and Pusteovsky & Tipton (2018) generalize this to regression and find that it works quite well even for non-normal data.

Clustering

Suppose an experiment:

- Some candidates from party A are randomly assigned to issue “pork barrel” appeals to their constituents, while others are randomly assigned to issue “national welfare” appeals.
- We measure effects in terms of voters’ tendency to vote for the party A candidate in their constituency.

(photo from http://www.tzaffairs.org/2009/01/by-election-shock-for-ccm/)

What determines precision of our effect estimates for this study?

What determines precision of our effect estimates for this study?

- The number of voters?

What determines precision of our effect estimates for this study?

- The number of voters?
- The number of party A candidates?

What determines precision of our effect estimates for this study?

- The number of voters?
- The number of party A candidates?
- Both?

(photo from http://www.tzaffairs.org/2009/01/by-election-shock-for-ccm/)

For causal effect estimation, here are some rules of thumb:

For causal effect estimation, here are some rules of thumb:

- The distribution of the treatment is the first thing to consider.

For causal effect estimation, here are some rules of thumb:

- The distribution of the treatment is the first thing to consider.
- If treatment is assigned to clusters of units, then you need to accounting for clustering.

For causal effect estimation, here are some rules of thumb:

- The distribution of the treatment is the first thing to consider.
- If treatment is assigned to clusters of units, then you need to accounting for clustering.
- Each cluster contributes independent information: assignment is uncorrelated from cluster to cluster, by definition.

For causal effect estimation, here are some rules of thumb:

- The distribution of the treatment is the first thing to consider.
- If treatment is assigned to clusters of units, then you need to accounting for clustering.
- Each cluster contributes independent information: assignment is uncorrelated from cluster to cluster, by definition.
- Cluster members each contribute less information: correlation of treatment assignment among cluster co-members, combined with correlation of outcomes among cluster co-members, makes cluster co-members *redundant* to some degree.

For causal effect estimation, here are some rules of thumb:

- The distribution of the treatment is the first thing to consider.
- If treatment is assigned to clusters of units, then you need to accounting for clustering.
- Each cluster contributes independent information: assignment is uncorrelated from cluster to cluster, by definition.
- Cluster members each contribute less information: correlation of treatment assignment among cluster co-members, combined with correlation of outcomes among cluster co-members, makes cluster co-members *redundant* to some degree.
- Thus, it is the number of clusters much more than the size of the clusters, that drives inflation of the variance and standard errors.

If you have heard about clustering before, this is probably a different way to think about it.

If you have heard about clustering before, this is probably a different way to think about it.

- In classical regression analysis, random variation is due to error term.

If you have heard about clustering before, this is probably a different way to think about it.

- In classical regression analysis, random variation is due to error term.
- Focuses on clustering in terms of “correlated errors” in outcome distribution.

If you have heard about clustering before, this is probably a different way to think about it.

- In classical regression analysis, random variation is due to error term.
- Focuses on clustering in terms of "correlated errors" in outcome distribution.
- For causal inference, correlated outcomes only matter when there is correlated treatment assignment.

If you have heard about clustering before, this is probably a different way to think about it.

- In classical regression analysis, random variation is due to error term.
- Focuses on clustering in terms of “correlated errors” in outcome distribution.
- For causal inference, correlated outcomes only matter when there is correlated treatment assignment.
- Correlations in treatment assignment are, in principle, knowable, whereas correlations in “errors” are not.

If you have heard about clustering before, this is probably a different way to think about it.

- In classical regression analysis, random variation is due to error term.
- Focuses on clustering in terms of “correlated errors” in outcome distribution.
- For causal inference, correlated outcomes only matter when there is correlated treatment assignment.
- Correlations in treatment assignment are, in principle, knowable, whereas correlations in “errors” are not.
- Therefore, practical consideration of what is “knowable” also favors emphasis on correlation in treatment assignment.

What are the clusters in the experiment?

Clustering and the Distribution of $\hat{\rho}$

- Suppose the population of interest is partitioned into a large number of clusters indexed by $h = 1, 2, ...$, with cluster h having N_h members.

Clustering and the Distribution of $\hat{\rho}$

- Suppose the population of interest is partitioned into a large number of clusters indexed by $h = 1, 2, ...,$ with cluster h having N_h members.
- Treatment is assigned in a way that is independent of potential outcomes, but for i, j in the same cluster, $\mathrm{Cor}[D_i, D_j] \neq 0$.

Clustering and the Distribution of $\hat{\rho}$

- Suppose the population of interest is partitioned into a large number of clusters indexed by $h = 1, 2, ...$, with cluster h having N_h members.
- Treatment is assigned in a way that is independent of potential outcomes, but for i, j in the same cluster, $\text{Cor}[D_i, D_j] \neq 0$.
- We randomly sample H clusters from the population.

Clustering and the Distribution of $\hat{\rho}$

- Suppose the population of interest is partitioned into a large number of clusters indexed by $h = 1, 2, ...$, with cluster h having N_h members.
- Treatment is assigned in a way that is independent of potential outcomes, but for i, j in the same cluster, $\text{Cor}[D_i, D_j] \neq 0$.
- We randomly sample H clusters from the population.
- We estimate the ATE with,

$$\hat{\rho} = \overline{Y_1} - \overline{Y_0}$$

Clustering and the Distribution of $\hat{\rho}$

- Suppose the population of interest is partitioned into a large number of clusters indexed by $h = 1, 2, ...$, with cluster h having N_h members.
- Treatment is assigned in a way that is independent of potential outcomes, but for i, j in the same cluster, $\text{Cor}[D_i, D_j] \neq 0$.
- We randomly sample H clusters from the population.
- We estimate the ATE with,

$$\hat{\rho} = \overline{Y_1} - \overline{Y_0}$$

- What are the consequences of clustering for the bias or consistency of $\hat{\rho}$?
- What about for the variance of the sampling/randomization distribution of $\hat{\rho}$, and therefore the standard error?

Clustering and the Distribution of $\hat{\rho}$

When the number of clusters, H, is small, $\hat{\rho}$ can be substantially biased. This is because of $\hat{\rho}$ may have a varying denominator:

$$\hat{\rho} = \overline{Y_1} - \overline{Y_0} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{N_h} D_{hi} Y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{N_h} D_{hi}} - \frac{\sum_{h=1}^{H} \sum_{i=1}^{N_h} (1 - D_{hi}) Y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{N_h} (1 - D_{hi})}$$

A toy example to illustrate this kind of bias: Suppose three clusters with outcomes, $\{1,1\}, \{3,4\}, \{10,20,30\}$. Sample 2, take mean.

Overall mean:	$[(1+1)+(3+4)+(10+20+30)]/7 = 9.86$
Sample 1 mean estimate:	$[(1+1)+(3+4)]/4 = 2.25$
Sample 2 mean estimate :	$[(1+1)+(10+20+30)]/5 = 12.4$
Sample 3 mean estimate :	$[(3+4)+(10+20+30)]/5 = 13.4$
	Expected value of estimator: 9.35

Clustering and the Distribution of $\hat{\rho}$

This kind of bias goes away as the number of clusters gets large. Thus, $\hat{\rho}$ is consistent for ρ in H.

Clustering and the Distribution of $\hat{\rho}$

This kind of bias goes away as the number of clusters gets large. Thus, $\hat{\rho}$ is consistent for ρ in H. To see this, let $H \to \infty$,

$$\begin{aligned}
\hat{\rho} = \overline{Y_1} - \overline{Y_0} &= \frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}Y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}} - \frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})Y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_hi)} \\
&= \frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}Y_{1hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}} - \frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})Y_{0hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})} \\
&= \frac{\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}Y_{1hi}}{\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}} - \frac{\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})Y_{0hi}}{\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})} \\
&\xrightarrow{p} \frac{\mathrm{E}\left[\sum_{i=1}^{N_h} D_{hi}Y_{1hi}\right]}{\mathrm{E}\left[\sum_{i=1}^{N_h} D_{hi}\right]} - \frac{\mathrm{E}\left[\sum_{i=1}^{N_h}(1-D_{hi})Y_{0hi}\right]}{\mathrm{E}\left[\sum_{i=1}^{N_h}(1-D_{hi})\right]} \\
&= \frac{\mathrm{E}\left[\sum_{i=1}^{N_h} D_{hi}\right]\mathrm{E}\left[Y_{1hi}\right]}{\mathrm{E}\left[\sum_{i=1}^{N_h} D_{hi}\right]} - \frac{\mathrm{E}\left[\sum_{i=1}^{N_h}(1-D_{hi})\right]\mathrm{E}\left[Y_{0hi}\right]}{\mathrm{E}\left[\sum_{i=1}^{N_h}(1-D_{hi})\right]} = \rho
\end{aligned}$$

Clustering and the Distribution of $\hat{\rho}$

So the usual estimator will be accurate, on average, so long as the number of clusters is large.

Clustering and the Distribution of $\hat{\rho}$

So the usual estimator will be accurate, on average, so long as the number of clusters is large.

Clustering and the Distribution of $\hat{\rho}$

So the usual estimator will be accurate, on average, so long as the number of clusters is large.

The variance of the sampling/randomization distribution of $\hat{\rho}$ is affected more strongly, however.

Clustering and the Distribution of $\hat{\rho}$

So the usual estimator will be accurate, on average, so long as the number of clusters is large.

The variance of the sampling/randomization distribution of $\hat{\rho}$ is affected more strongly, however.

- Given large H, the population level variance for $\hat{\rho}$ is given by,

$$\begin{aligned}\text{Var}\left[\hat{\rho}\right] &= \text{Var}\left[\overline{Y_1}\right] + \text{Var}\left[\overline{Y_0}\right] \\ &= \text{Var}\left[\frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}Y_{1hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}}\right] + \text{Var}\left[\frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})Y_{0hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})}\right].\end{aligned}$$

Clustering and the Distribution of $\hat{\rho}$

So the usual estimator will be accurate, on average, so long as the number of clusters is large.

The variance of the sampling/randomization distribution of $\hat{\rho}$ is affected more strongly, however.

- Given large H, the population level variance for $\hat{\rho}$ is given by,

$$\begin{aligned}\mathrm{Var}\left[\hat{\rho}\right] &= \mathrm{Var}\left[\overline{Y_1}\right] + \mathrm{Var}\left[\overline{Y_0}\right] \\ &= \mathrm{Var}\left[\frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}Y_{1hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h} D_{hi}}\right] + \mathrm{Var}\left[\frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})Y_{0hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(1-D_{hi})}\right].\end{aligned}$$

- Both the numerator and denominator are random, so evaluating is a little complicated (start with a Taylor expansion).

Clustering and the Distribution of $\hat{\rho}$

To get at the essential issues, let's simplify the situation:

Clustering and the Distribution of $\hat{\rho}$

To get at the essential issues, let's simplify the situation:

- Suppose total number assigned to treatment and control is fixed to M_1 and M_0, respectively.
- Number of cluster is sufficiently large that we can ignore any negative correlation in assignment *between* clusters

Clustering and the Distribution of $\hat{\rho}$

To get at the essential issues, let's simplify the situation:

- Suppose total number assigned to treatment and control is fixed to M_1 and M_0, respectively.
- Number of cluster is sufficiently large that we can ignore any negative correlation in assignment *between* clusters
- Then,

$$
\begin{aligned}
\operatorname{Var}[\hat{\rho}] &\approx \frac{1}{M_1^2}\sum_{h=1}^{H}\operatorname{Var}\left[\sum_{i=1}^{N_h} D_{hi}Y_{1hi}\right] + \frac{1}{M_0^2}\sum_{h=1}^{H}\operatorname{Var}\left[\sum_{i=1}^{N_h}(1-D_{hi})Y_{0hi}\right] \\
&= \frac{1}{M_1^2}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\left(\underbrace{\operatorname{Var}[D_{hi}Y_{1hi}]}_{A} + 2\sum_{j\neq i}\underbrace{\operatorname{Cov}[D_{hi}Y_{1hi}, D_{hj}Y_{1hj}]}_{B}\right) \\
&\quad + \frac{1}{M_0^2}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\left(\underbrace{\operatorname{Var}[(1-D_{hi})Y_{0hi}]}_{A} + 2\sum_{j\neq i}\underbrace{\operatorname{Cov}[(1-D_{hi})Y_{0hi}, (1-D_{hj})Y_{0hj}]}_{B}\right),
\end{aligned}
$$

where A terms are usual unit-level contributions to variance, and B terms characterize variance inflation due to clustering.

Clustering and the Distribution of $\hat{\rho}$

- Closer look at cluster variance inflation term:

$$\begin{aligned} \mathrm{Cov}\left[D_{hi}Y_{1hi}, D_{hj}Y_{1hj}\right] &= \mathrm{E}\left[D_{hi}Y_{1hi}D_{hj}Y_{1hj}\right] - \mathrm{E}\left[D_{hi}Y_{1hi}\right]\mathrm{E}\left[D_{hj}Y_{1hj}\right] \\ &= \mathrm{E}\left[D_{hi}D_{hj}\right]\mathrm{E}\left[Y_{1hi}Y_{1hj}\right] - \mathrm{E}\left[D_{hi}\right]\mathrm{E}\left[D_{hj}\right]\mathrm{E}\left[Y_{1hi}\right]\mathrm{E}\left[Y_{1hj}\right]. \end{aligned}$$

- Variance inflation depends on treatment covariance multiplied by outcome covariance.

Clustering and the Distribution of $\hat{\rho}$

So, to recap, correlated assignment among co-members of a cluster results in the following:

- No problems in terms of consistency so long as the number of clusters is large. Usual estimators (e.g., $\hat{\rho}$) are accurate.
- Larger sampling/randomization variance when there is also outcome correlation among co-members of a cluster.
- This means that we need to adjust our standard error estimates accordingly.

Clustering in the Regression Context

- Regression provides clean results in derivation of “cluster robust” standard errors.

Clustering in the Regression Context

- Regression provides clean results in derivation of "cluster robust" standard errors.
- Consider a generic least squares regression of Y_i on some regressors, X_i.
- Recall distribution of OLS fit under unit-level sampling (even if misspecified):

$$\mathrm{E}\,[X_i X_i']^{-1}\mathrm{E}\,[X_i X_i' e_i^2]\mathrm{E}\,[X_i X_i']^{-1}.$$

- With clustering, things are not quite so simple.

Clustering in the Regression Context

- Recall H clusters were sampled and treatment assignment is correlated within clusters, with N_h units per cluster.

Clustering in the Regression Context

- Recall H clusters were sampled and treatment assignment is correlated within clusters, with N_h units per cluster.
- Let $N = \sum_{h=1}^{H} N_h$, the total number of units.
- We use the index hi to denote unit i in cluster h.

Clustering in the Regression Context

- Recall H clusters were sampled and treatment assignment is correlated within clusters, with N_h units per cluster.
- Let $N = \sum_{h=1}^{H} N_h$, the total number of units.
- We use the index hi to denote unit i in cluster h.
- For asymptotics in H, we evaluate

$$\sqrt{H}(\hat{\beta} - \beta) = \left[\frac{1}{N} \sum_h \sum_i X_{hi} X'_{hi} \right]^{-1} \frac{\sqrt{H}}{N} \sum_h \sum_i X_{hi} e_{hi}$$

.

- Under standard regularity conditions, the first term converges in H to $\mathrm{E}\left[X_{hi} X'_{hi}\right]^{-1}$.
- By Slutsky, this leaves $\frac{\sqrt{H}}{N} \sum_h \sum_i X_{hi} e_{hi}$ for us to evaluate in the limit.
- A before, this asymptotic distribution has mean zero.

Clustering in the Regression Context

- The variance follows

$$\begin{aligned}
\mathrm{Var}\left[\sum_{h=1}^{H}\sum_{i=1}^{N_h} X_{hi}e_{hi}\right] &= \sum_{h=1}^{H}\mathrm{Var}\left[\sum_{i=1}^{N_h} X_{hi}e_{hi}\right] = \sum_{h=1}^{H}\mathrm{Var}\left[\mathbf{X}_h' e_h\right] \\
&= \sum_{h=1}^{H}\mathrm{E}\left[(\mathbf{X}_h' e_h - E[\mathbf{X}_h' e_h])(e_h'\mathbf{X}_h - E[e_h'\mathbf{X}_h])\right] \\
&= \sum_{h=1}^{H}\mathrm{E}\left[\mathbf{X}_h' e_h e_h' \mathbf{X}_h\right].
\end{aligned}$$

- This bears a very strong resemblance to what we saw before with the difference in means estimator. Taking it a step further reveals some more insights...

Clustering in the Regression Context

$$\sum_{h=1}^{H} \mathrm{E}\left[\mathbf{X}_h' e_h e_h' \mathbf{X}_h\right] = \sum_{h=1}^{H} \mathrm{E}\left\{\mathbf{X}_h' \mathrm{E}\left[e_h e_h' | \mathbf{X}_h\right] \mathbf{X}_h\right\}$$

$$= \sum_{h=1}^{H} \mathrm{E}\left\{\mathbf{X}_h' \mathrm{E}\left[\left.\begin{pmatrix} e_{h1}^2 & e_{h1}e_{h2} & \cdots \\ e_{h1}e_{h2} & e_{h1}^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}\right| X\right] X_h\right\}$$

$$= \sum_{h=1}^{H} \mathrm{E}\left[\mathbf{X}_h' \begin{pmatrix} \mathrm{Var}\left[e_{h1}|\mathbf{X}\right] & \mathrm{Cov}\left[e_{h1}e_{h2}|\mathbf{X}\right] & \cdots \\ \mathrm{Cov}\left[e_{h1}e_{h2}|\mathbf{X}\right] & \mathrm{Var}\left[e_{h1}|\mathbf{X}\right] & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \mathbf{X}_h\right].$$

which, for X_{hi} of length K, yields a sum of $K \times K$ matrices with elements of the form,

$$\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{N_h} \mathrm{E}\{\underbrace{X_{hi,k}X_{hj,k}}_{A}\underbrace{\mathrm{Cov}\left[e_{hi}, e_{hj}|\mathbf{X}\right]}_{B}\},$$

combining regressor covariance (A) with residual covariance (B).

Asymptotically valid "cluster robust" standard errors are constructed by substituting in sample analogues for the expectations, variances, and covariances, yielding the estimator,

$$\hat{\mathbf{V}}_{CR,a} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{h=1}^{H} \mathbf{X}_h' \hat{e}_h \hat{e}_h' \mathbf{X}_h \right) (\mathbf{X}'\mathbf{X})^{-1}.$$

Asymptotically valid "cluster robust" standard errors are constructed by substituting in sample analogues for the expectations, variances, and covariances, yielding the estimator,

$$\hat{\mathbf{V}}_{CR,a} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{h=1}^{H}\mathbf{X}_h'\hat{e}_h\hat{e}_h'\mathbf{X}_h\right)(\mathbf{X}'\mathbf{X})^{-1}.$$

In some software packages (e.g Stata), a finite sample correction is applied to improve performance in moderately sized samples. The correction is derived from sample theoretic arguments and yields,

$$\hat{\mathbf{V}}_{CR,f} = \frac{H}{H-1}\frac{H\bar{N}-1}{H\bar{N}-K}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{h=1}^{H}\mathbf{X}_h'\hat{e}_h\hat{e}_h'\mathbf{X}_h\right)(\mathbf{X}'\mathbf{X})^{-1},$$

where $\bar{N}$ is the average cluster size. This is what you get with Stata's "`cluster`" option.

Asymptotically valid "cluster robust" standard errors are constructed by substituting in sample analogues for the expectations, variances, and covariances, yielding the estimator,

$$\hat{\mathbf{V}}_{CR,a} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{h=1}^{H} \mathbf{X}_h' \hat{e}_h \hat{e}_h' \mathbf{X}_h \right) (\mathbf{X}'\mathbf{X})^{-1}.$$

In some software packages (e.g Stata), a finite sample correction is applied to improve performance in moderately sized samples. The correction is derived from sample theoretic arguments and yields,

$$\hat{\mathbf{V}}_{CR,f} = \frac{H}{H-1} \frac{H\bar{N}-1}{H\bar{N}-K} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{h=1}^{H} \mathbf{X}_h' \hat{e}_h \hat{e}_h' \mathbf{X}_h \right) (\mathbf{X}'\mathbf{X})^{-1},$$

where $\bar{N}$ is the average cluster size. This is what you get with Stata's "`cluster`" option.
See Imbens and Kolesar (2016) and Pustejovsky and Tipton (2018) for further small sample refinements based on on Welch-Satterthwaite approximation.

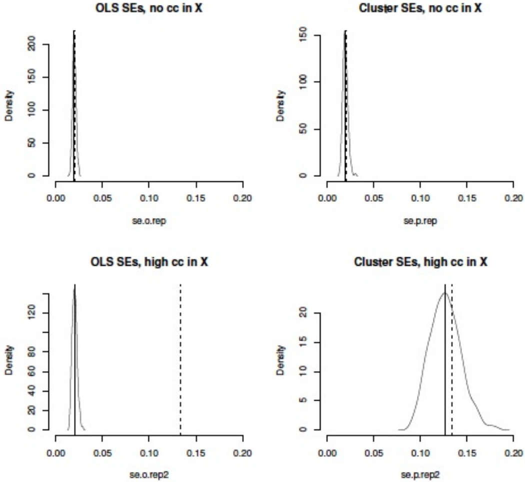

Figure 1: *Kernel density plots showing the distribution of standard error estimates for the coefficient on x_{it} from 500 simulation runs for OLS standard errors and cluster robust standard errors. The dashed line shows the actual standard deviation of the regression coefficient over the 500 runs, and the solid line shows the mean of the standard error estimates. For all four cases, there is substantial intra-cluster correlation in the errors, but only for the bottom two is there any intra-correlation in the x'_{it}s. "cc" in the plot titles refers to "clustered correlation."*

Clustering in the Regression Context

Another way to characterize these properties is in the manner of Moulton (1986) (cf. MHE, Ch. 8).

Clustering in the Regression Context

Another way to characterize these properties is in the manner of Moulton (1986) (cf. MHE, Ch. 8).

- Suppose a cluster randomized experiment, where outcomes can be modeled as,

$$Y_{ih} = \beta_0 + \beta_1 D_h + \nu_h + \eta_{hi},$$

  and ν_h is a zero-mean, cluster-specific "random effect" that is independent across groups and has variance σ_ν^2, while η_{hi} is a zero-mean, unit specific error term that is independent across individuals and has variance σ_η^2.

Clustering in the Regression Context

$$Y_{ih} = \beta_0 + \beta_1 D_h + \nu_h + \eta_{hi},$$

- The compound error term has total variance, $\sigma_\nu^2 + \sigma_\eta^2$.
- For units in the same group, compound error terms have covariance, $\mathrm{E}\left[\nu_h + \eta_{hi}\right]\left[\nu_h + \eta_{hj}\right] = \sigma_\nu^2$.
- Then, the correlation between outcomes for units i and j in cluster h is given by,

$$\rho_{ICC,\nu} = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2},$$

a quantity known as the "intra-class correlation" coefficient.

Clustering in the Regression Context

- Under this model, we can obtain a neat expression for the effects of clustering on $\mathrm{Var}[\hat{\beta}_1]$ from OLS.
- The expression, called the "Moulton factor," relates the true variance of $\hat{\beta}_1$ to the expected value of the homoskedasticity variance estimator:

$$\frac{V_{true}(\hat{\beta}_1)}{V_{homosk.}(\hat{\beta}_1)} \approx 1 + (\bar{N} - 1)\rho_{ICC,\nu},$$

where $\bar{N}$ is the average cluster size.

Clustering in the Regression Context

- If we have X_{hi} that varies within clusters, then the generalized Moulton factor is (cf. MHE, p. 311):

$$\frac{V_{true}(\hat{\beta}_1)}{V_{homosk.}(\hat{\beta}_1)} = 1 + \left(\frac{\mathrm{Var}\left[N_h\right]}{\bar{N}} + \bar{N} - 1\right) \rho_{ICC,x}\rho_{ICC,v}$$

where $\rho_{ICC,X}$ is the intra-class correlation of the X_{hi}'s.

Clustering in the Regression Context

- If we have X_{hi} that varies within clusters, then the generalized Moulton factor is (cf. MHE, p. 311):

$$\frac{V_{true}(\hat{\beta}_1)}{V_{homosk.}(\hat{\beta}_1)} = 1 + \left(\frac{\text{Var}[N_h]}{\bar{N}} + \bar{N} - 1\right) \rho_{ICC,x}\rho_{ICC,v}$$

where $\rho_{ICC,X}$ is the intra-class correlation of the X_{hi}'s.

- Reinforces what we have seen: consequences of clustering arise from treatment clustering $\times$ outcome clustering.

Clustering, Regression, and Causal Effect Estimation

- Putting it all together, OLS with unit-level data is consistent.

Clustering, Regression, and Causal Effect Estimation

- Putting it all together, OLS with unit-level data is consistent.
- When treatment assignment exhibits clustering – e.g., if it is a cluster- or group- randomized experiment or quasi-experiment– then cluster-robust standard errors will provide confidence intervals with proper coverage when number of clusters is large.

Remarks

- In cluster-randomized experiments and clustered natural experiments, all cluster co-members typically receive the *same* treatment, and so their correlation is 1. This maximizes the degree of potential variance inflation.

Remarks

- In cluster-randomized experiments and clustered natural experiments, all cluster co-members typically receive the *same* treatment, and so their correlation is 1. This maximizes the degree of potential variance inflation.
- Clustering may refer to spatial clustering or to any other relationships between units that makes units' exposure to treatment likely to be correlated.
- E.g., Suppose the treatment is a country's external trade policy and you want to know the effect on A's trade *partners*. Then, exposure to the treatment is clustered among the network of trade partners (cf. Aronow, Samii, and Assenova, 2015, and Tabord-Meehan, 2018, for "dyadic robust").

Remarks

- The approach covered here has remained true to our "agnostic" approach:
  - minimal assumptions on outcomes,
  - make use of known (or more "knowable") design—namely the sampling and treatment assignment process.

Remarks

- The approach covered here has remained true to our "agnostic" approach:
  - minimal assumptions on outcomes,
  - make use of known (or more "knowable") design—namely the sampling and treatment assignment process.
- Other approaches exist for handling the clustering problem, including as parametric random effects estimation, multi-level models, etc.
- They rely on more stringent assumptions, which, when valid, make the estimation more precise. See Gelman & Hill (2007), Green & Vavreck (2008).