Lecture 2 – Introduction, continued

Economics 8379
George Washington University

Instructor: Prof. Ben Williams

- “The gold standard for drawing inferences about the effect of a policy is a randomized controlled experiment.” (Athey and Imbens, 2017, *JEP*)

- “The gold standard for drawing inferences about the effect of a policy is a randomized controlled experiment.” (Athey and Imbens, 2017, *JEP*)
- “...any special status for RCTs is unwarranted.” (Deaton and Cartwright, 2018)

- If D_i is randomly assigned then

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = E(Y_{1i} - Y_{0i})$$

- $E(Y_{1i} - Y_{0i})$ is the *average treatment effect* (ATE)

- In sample,

$$\bar{Y}_1 - \bar{Y}_0 = \bar{\beta}_1 + \bar{S}_1 - \bar{S}_0$$

  where
  - $\bar{\beta}_1$ is the sample average of $Y_{1i} - Y_{0i}$
  - $\bar{S}_1 - \bar{S}_0$ represents the selection bias in finite sample
- it is important to recognize that $\bar{S}_1 - \bar{S}_0$ is equal to zero only *in expectation*
- So it is super important to do *inference* well.

- Passing a balance test does not mean that $\bar{S}_1 - \bar{S}_0$ should be small.
  - it does detect problems with randomization mechanism
  - useful to describe balance, but not to *test*

- The imbalance is an issue of efficiency.
- This leads to the following issues:
  - stratification
  - regression adjustment
  - other experimental designs
  - getting the standard errors right

Stratification

Stratified random sampling

- Split the population into groups and randomly sample from each group.
- The estimate should use population proportions for the groups to aggregate back up.
- This is a useful way to increase precision of estimates.
- n.b: Clustered sampling is different – it generally *reduces* precision.
- See *Analysis of Household Surveys* by Angus Deaton for a good intro to these issues.

Stratification

Stratified random sampling

- Split the population into groups and randomly sample from each group.
- The estimate should use population proportions for the groups to aggregate back up.
- This is a useful way to increase precision of estimates.
- n.b: Clustered sampling is different – it generally *reduces* precision.
- See *Analysis of Household Surveys* by Angus Deaton for a good intro to these issues.

- This is *not* what Athey and Imbens are talking about.

Stratification

Stratified random assignment

- Split the sample into groups and randomly assign treatment within each group.
- Estimate treatment effects within each stratum and then aggregate using stratum shares.
- This is a useful way to increase precision of estimates.
- Athey and Imbens (2017) recommend stratifying as much as possible – on characteristics that are predictive of outcomes – with as few as 2 T and 2 C in each stratum.
- Paired random assignment is the extreme case of this with 1 T and 1 C in each stratum.

Stratification

"Stratified estimation" under completely random assignment

- Randomly assign treatment in the sample.
- Estimate treatment effects within "strata" and then aggregate using stratum shares.
- This increases precision of estimates but not as much as stratified random assignment does.

Stratification

Regression adjustment

- The approach on the previous slide is difficult if the covariates do not easily define "strata" and/or the sample size is small.
- Instead we often estimate the regression: $Y_i = \beta_0 + \beta_1 D_i + \beta_2' X_i + u_i$
- But the OLS estimator is biased in finite samples...random assignment of X_i does not ensure that $E(u_i \mid D_i, X_i) = 0$
- OLS estimator is consistent but Freedman (2008) shows that the finite sample bias can be considerable.
- Turn covariates into indicators and include interactions – benefits/costs?

Stratification

Other assignment mechanisms

- The key idea here is that randomization is not necessary and in fact sacrifices efficiency for ...?
- Section 2.4 in Deaton and Cartwright (2018)
- don't allow assignment based on unobservables!
- requires a priori knowledge about outcomes
- It's an old idea but papers by Banerjee, Chassang, Montero, and Snowberg; Banerjee, Chassang, and Snowberg; and Kasy are pushing the frontier on this.

Inference in RCTs

- Deaton and Cartwright (2018): inference is super important in RCTs as the selection bias is only zero in expectation.
- Some important issues:
  - the t statistic does not have Student's t distribution in small samples (Fisher-Behrens problem)
  - if the distribution of treatment effects is not symmetric, normality is problematic anyway
  - clustered standard errors are also problematic for similar reasons
  - regression adjustment can make the problem worse (Young, 2017)

Inference in RCTs

- Deaton and Cartwright (2018) simulation

Inference in RCTs

CHANNELLING FISHER: RANDOMIZATION TESTS AND THE STATISTICAL INSIGNIFICANCE OF SEEMINGLY SIGNIFICANT EXPERIMENTAL RESULTS*

ALWYN YOUNG

I follow R.A. Fisher's The Design of Experiments (1935), using randomization statistical inference to test the null hypothesis of no treatment effects in a comprehensive sample of 53 experimental papers drawn from the journals of the American Economic Association. In the average paper randomization tests of the significance of individual treatment effects find 13 to 22 percent fewer significant results than found using authors' methods. In joint tests of multiple treatment effects appearing together in tables, randomization tests yield 33 to 49 percent fewer statistically significant results than conventional tests. Bootstrap and jackknife methods support and confirm the randomization results. JEL Codes: C12, C90.

Inference in RCTs

- Solutions:
  - be aware of asymmetry in data
  - randomization inference
  - Imbens and Kolesar (2016) (later)

Randomization inference

- Fisher's exact p-value.
- The idea is to reassign treatment artificially in your data to get D_i^* and re-estimate your effects.
- Doing this a bunch of times you get a distribution of estimates and you locate the real-data estimate in this distribution.
- Allows you to mimic the real process of random assignment.
- The key drawback is that this is a test of the null of no treatment effects, *not* of $ATE = 0$.

One last thing on inference

- Athey and Imbens (2017) suggest that we should think of our sample as fixed, not drawn from an infinite superpopulation.
- From this viewpoint, they argue that our usual standard errors are actually *conservative*.
- Similar idea in Abadie, Athey, Imbens, and Wooldridge (2017) regarding clustering.

Bias in RCTs

- There are well-known issues to be aware of that can cause *bias* in ATE estimates:
  - post-randomization differences
  - non-compliance
  - violations of SUTVA

External validity

- the claim that RCTs are often limited in their external validity is uncontroversial
- Deaton and Cartwright (2018) and Athey and Imbens (2017) differ in their view of the importance of internal vs external validity
- Heckman (2008) is relevant here too but I will focus on new ideas from the Deaton and Cartwright (2018) reading

External validity

- Some issues
  - ATE versus other policy countefactuals
  - Bertrand's chicken and why replication in other settings is not enough
  - interactions; "support factors"
  - general equilibrium effects and other scale problems
  - what do average effects tell us about individual effects?

External validity

- Some fixes
  - there is a lot of recent work on using reweighting/stratifying to extrapolate
    - in addition to references in Deaton and Cartwright (2018), section 3.5, see Rachel Meager's work
  - using RCTs to build and test theory
    - in addition to references in Deaton and Cartwright (2018), section 3.6, see Karen Ye's work

External validity

- Some fixes
  - there is a lot of recent work on using reweighting/stratifying to extrapolate
    - in addition to references in Deaton and Cartwright (2018), section 3.5, see Rachel Meager's work
  - using RCTs to build and test theory
    - in addition to references in Deaton and Cartwright (2018), section 3.6, see Karen Ye's work
  - this work is the future (imho)

Overview

- Deaton and Cartwright (2018) stress the importance of valid inference in RCTs.
- But issues with the appropriate standard errors are relevant not for RCTs only

Overview

- Deaton and Cartwright (2018) stress the importance of valid inference in RCTs.
- But issues with the appropriate standard errors are relevant not for RCTs only
- we will talk about issues with std errors in OLS, with RCT as a special case
- we will talk more next class about OLS

OLS estimator

- Let X be an $n \times K$ matrix of covariates and y a $n \times 1$ vector of outcomes.
- Denote the rows of X as X_i' and elements of y as y_i.
- The OLS estimator can be written as

$$(X'X)^{-1}X'y = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i y_i$$

- Let $\beta = E(X_i X_i')^{-1} E(X_i y_i)$ and define $e = y - X\beta$, with elements e_i.

Variance of OLS

- Treat X as non-stochastic and define $\Psi = E(ee')$.
- Then $\Omega = Var(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Psi X(X'X)^{-1}$

Variance of OLS

- Treat X as non-stochastic and define $\Psi = E(ee')$.
- Then $\Omega = Var(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Psi X(X'X)^{-1}$
- HAC estimators - heteroskedasticity and autocorrelation consistent
  - different formulas based on different restrictions on Ψ

"Conventional"

- if $\Psi = \sigma^2 I_n$ then $X'\Psi X = \sigma^2 X'X$ so

$$\hat{\Omega}_c = n(X'X)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\right)$$

where $\hat{e}_i = Y_i - X_i'\hat{\beta}$

Heteroskedasticity robust

- if heteroskedasticity but no autocorrelation, $\Psi = diag(\psi_i)$
- then $X'\Psi X = \sum_{i=1}^{n} X_i X_i' \psi_i$
- the Eicker-Huber-White standard errors -

$$\hat{\Omega}_r = n(X'X)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \hat{e}_i^2 \right) (X'X)^{-1}$$

- `robust` option in Stata

Heteroskedasticity robust

- note:
  - this is only an asymptotic result
  - it works because $\frac{1}{n} X' diag(\hat{e}_i^2) X \rightarrow_p \frac{1}{n} X' diag(\psi_i) X$

Clustered formula

- suppose Ω is block diagonal with L blocks
- (more on when this is the appropriate model later)
- cluster-robust standard errors are based on the fact that

$$L^{-1} \sum_{l=1}^{L} X_l' \hat{e}_l \hat{e}_l' X_l \to_p L^{-1} X' \Psi X$$

  as $L \to \infty$
- This is the Liang-Zeger formula.
- `cluster` option in Stata; have to specify clustering variable

Other HAC estimators

- spatial autocorrelation
  - Conley (1999) and others
  - Stata .ado file on Conley's website
- time series with serial correlation
  - Newey West is most common (`newey` in Stata)
  - but many other approaches out there too (see, e.g., https://scholar.harvard.edu/files/stock/files/...)
- Multiway clustering – Cameron, Gelbach, Miller (2011) (`cgmreg`)

Bias in heteroskedasticity robust ses

- The mean of the robust standard error, $E(\hat{\Omega}_r)$, is

$$(X'X)^{-1}\left(\sum_{i=1}^{n} X_i X_i'(Var(e_i) - 2Var(e_i)h_{ii} + h_i'\Psi h_i)\right)(X'X)^{-1}$$

  where $H = X(X'X)^{-1}X$ and h_{ii} is called the leverage.
- we can improve it with bias corrections

Bias in heteroskedasticity robust ses

- Bias corrections involve replacing $\hat{e}_i^2$ with
  - $HC_1 = \frac{N}{N-K}\hat{e}_i^2$
  - $HC_2 = \frac{1}{1-h_{ii}}\hat{e}_i^2$
  - or $HC_3 = \frac{1}{(1-h_{ii})^2}\hat{e}_i^2$
- AP: generally, HC_3 is preferred to HC_2 is preferred to HC_1 when there is heteroskedasticity, but the reverse when errors are homoskedastic.
- this is a small sample problem – AP's monte carlo uses $n = 30$
- actually the asymmetry is crucial too – in AP's example, X_i is binary with $Pr(X_i = 1) = 0.1$!

- Imbens and Kolesar (2015) clarify a few points here.
  - HC_2 adjustment makes the variance estimator unbiased under homoskedasticity.
  - the adjustments are very intuitive in the case of a single binary regressor
    - robust variance estimator:

$$\frac{\tilde{\sigma}_0^2}{n_0} + \frac{\tilde{\sigma}_1^2}{n_1}$$

      where $\tilde{\sigma}_d^2 = n_d^{-1} \sum_{i \in d} (Y_i - \bar{Y}_d)^2$
    - HC_2 adjustment replaces $\tilde{\sigma}_d^2$ with $\hat{\sigma}_d^2 = (n_d - 1)^{-1} \sum_{i \in d} (Y_i - \bar{Y}_d)^2$

- Imbens and Kolesar (2015) clarify a few points here.
  - the adjustments are very intuitive in the case of a single binary regressor
    - A well-known problem: Behrens-Fisher
    - Welch solution: use HC_2 and use the Student's t distribution with K^*_{Welch} dof.
    - Note: K^*_{Welch} involves unknown variances...feasible version K_{Welch} uses sample variances

- Imbens and Kolesar (2015) for general regression case.
  - Bell and McCaffrey (2002) adjustment
  - use HC_2 and use the Student's t distribution with K^*_{BM} dof.
  - Note: K^*_{BM} does not involve estimating the variances...
- Imbens and Kolesar (2015) find that BM is better in their simulations.
- Formulas and R code are provided by Imbens and Kolesar (2015).

from Imbens and Kolesar, 2015

TABLE 2.—COVERAGE RATES AND NORMALIZED STANDARD ERRORS (IN PARENTHESES) FOR DIFFERENT CONFIDENCE INTERVALS IN THE BEHRENS-FISHER PROBLEM

Angrist-Pischke Unbalanced Design, $N_0 = 27, N_1 = 3$, Log-Normal Errors

		I		II		III		IV		V	
	$\sigma(0)$	0.5		0.85		1		1.18		2	
A. Coverage Rates and Median Standard Errors											
Variance Estimator	Dist/dof	Cov. Rate	Med. SE	Cov. Rate	Med. SE	Cov. Rate	Med. SE	Cov. Rate	Med. SE	Cov. Rate	Med. SE
$\hat{\mathbb{V}}_{\text{homo}}$	∞	75.9	(0.26)	91.8	(0.41)	93.3	(0.47)	94.4	(0.55)	97.0	(0.91)
	$N-2$	78.2	(0.27)	92.6	(0.43)	93.9	(0.49)	94.9	(0.58)	97.3	(0.96)
$\hat{\mathbb{V}}_{\text{EHW}}$	∞	66.8	(0.22)	73.4	(0.26)	76.7	(0.27)	80.6	(0.30)	91.1	(0.41)
	$N-2$	68.2	(0.23)	75.0	(0.27)	78.4	(0.29)	82.4	(0.31)	92.6	(0.42)
	wild	76.1	(0.36)	78.8	(0.36)	81.1	(0.36)	84.2	(0.38)	94.3	(0.46)
	wild_0	95.2	(0.44)	99.0	(0.63)	99.1	(0.72)	99.2	(0.84)	99.5	(1.37)
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	71.3	(0.26)	77.2	(0.29)	80.2	(0.31)	83.7	(0.33)	93.3	(0.44)
	$N-2$	72.5	(0.27)	78.6	(0.31)	81.7	(0.33)	85.3	(0.35)	94.5	(0.46)
	wild	77.2	(0.38)	79.7	(0.38)	81.8	(0.38)	84.7	(0.39)	94.4	(0.48)
	wild_0	95.2	(0.44)	99.0	(0.63)	99.1	(0.72)	99.2	(0.84)	99.5	(1.37)
	K_{Welch}	79.9	(0.47)	82.2	(0.44)	84.3	(0.44)	87.1	(0.44)	95.7	(0.52)
	K^*_{Welch}	90.1	(0.54)	95.8	(0.57)	97.2	(0.57)	98.3	(0.58)	98.9	(0.62)
	K_{BM}	87.2	(0.48)	94.9	(0.54)	97.2	(0.57)	98.8	(0.61)	99.7	(0.81)
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	75.4	(0.31)	80.6	(0.34)	83.2	(0.36)	86.4	(0.38)	94.9	(0.49)
	$N-2$	76.5	(0.33)	81.9	(0.36)	84.6	(0.37)	87.8	(0.40)	95.9	(0.51)
$\max_{\text{EHW}}$	∞	85.7	(0.30)	97.8	(0.44)	98.4	(0.50)	98.6	(0.58)	98.8	(0.93)
$\max_{\text{HC2}}$	∞	86.9	(0.33)	98.5	(0.46)	99.0	(0.52)	99.2	(0.60)	99.3	(0.94)
B. Mean Effective dof											
	K^*_{Welch}	2.1		2.3		2.5		2.7		4.1	
	K_{Welch}	4.9		7.5		8.5		9.7		14.0	
	K_{BM}	2.5		2.5		2.5		2.5		2.5	

"Cov. Rate" refers to coverage of nominal 95% confidence intervals (in percentages), and "Med. SE" refers to standard errors normalized by $t^K_{0.975}/t^\infty_{0.975}$. Variance estimators and degrees-of-freedom (dof) adjustments are described in the text, and wild bootstrap confidence intervals ("wild" and "wild_0") are described in section 2 in the appedix; $\max_{\text{EHW}} = \max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{EHW}})$, and $\max_{\text{HC2}} = \max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}})$. Results are based on 1 million replications, except for wild bootstrap-based confidence intervals, which use 100,000 replications and 1,000 bootstrap draws in each replication.

Clustered standard errors

When should we cluster?

Clustered standard errors

- Moulton factor:
  - Suppose $e_i = \nu_{l_i} + \eta_i$.
  - Intraclass correlation: $\rho_e = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$
  - The ratio of the correct standard error formula to the conventional one is:

$$1 + \rho_x \rho_e \left(\frac{Var(n_l)}{\bar{n}} + \bar{n} - 1 \right)$$

- Generally, clustering matters when (1) group effects explain a lot of variation in outcome *and* (2) the regressors are correlated within groups

Clustered standard errors

Abadie, Athey, Imbens, Wooldridge (2017)

1. Actually it's the intraclass correlation in *residual* $\times$ *covariate* that matters.
   - Suppose, e.g., $Y_i = \tau_{l_i} W_i + \nu_i$ where ν_i is iid.
   - there is no intraclass correlation in ν_i or in W_i in their example
   - but there is intralcass correlation in this product
2. If we observe all (or most) of the clusters then we actually shouldn't cluster.

Clustered standard errors

- justification of the clustered std error formula requires a large number of clusters
- Imbens and Kolesar (2015) also provide a HC_2 formula and dof adjustment for the clustering case from Bell and McCaffrey (2002).
- alternatively, we can use a block bootstrap
- we will talk more about this when we talk about panel data