Econometrics II

Lecture 4: Experiments

Konrad Burchardi

Stockholm University

11th of April 2024

Literature

1. **”Mostly Harmless Econometrics”**, Angrist and Pischke
   Chapter 2 [introduction]; Chapter 3.2.3 [bad controls]
2. **”Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction”**, Imbens and Rubin
   Chapter 4 [introduction]; Chapter 5 [Fisher inference]; Chapter 7.5 [controls]

All mistakes are mine.

Plan for Today

Unbiased Estimation

Why randomize?

$\Rightarrow$ Balanced distribution of potential outcomes.

Randomized experiment guarantees, by design, *ex ante*:

$$\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] = 0$$

Therefore:

$$\begin{aligned}
&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \equiv ATT
\end{aligned}$$

and in fact:

$$= \mathbb{E}[Y_i(1) - Y_i(0)] \equiv ATE(\text{'average treatment effect'})$$

How to Randomize?

Why should we think about how to randomise?

Purely random treatment assignment:

1. Suboptimal fractions of T/C.
    1. Ex-ante random assignment: unnecessarily low **power**.
2. Imbalanced distribution of potential outcomes across T/C.
    1. Ex-post random assignment: problematic **causal inference**.
    2. Ex-ante random assignment: unnecessarily low **power**.

How should we randomize optimally then?

Set Fraction of T and C

Assignment Mechanism: Randomisation conditional on N_1 and N_0.

Under this assignment mechanism: $V(\hat{\beta}|N_0, N_1) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N-N_1}$.

How to minimise $V(\hat{\beta})$?

With $\sigma_0^2 = \sigma_1^2$:	$N_1^* = N/2$.
With $\sigma_0^2 \neq \sigma_1^2$:	More observations for noisy outcome.
Budget & Costs:	With costs c_0 and c_1, and a fixed budget B, then $\min_{N,N_1} V(\hat{\beta})$ s.s. $(N - N_1)c_0 + N_1 c_1 \leq B$. More observations for cheap outcome.

How to Randomize?

Why should we think about how to randomise?

Purely random treatment assignment:

1. Suboptimal fractions of T/C.
    1. Ex-ante random assignment: unnecessarily low power.
2. Imbalanced distribution of potential outcomes across T/C.
    1. Ex-post random assignment: problematic causal inference.
    2. Ex-ante random assignment: unnecessarily low power.

How should we randomize optimally then?

Plan for Today

Balance Tests: How to judge imbalances ex-post?

In practice, treatment arms are unlikely "balanced".

Imbalances in potential outcomes show up in covariates!

How to detect imbalances?

- Often t-test for each covariate shown.
  - Conceptually problematic.
  - Statistical significance is not what matters.[1]
- What is useful then?
  - Focus on **size of differences** for **covariates that impact outcome**![2]
  - Estimate size of difference using same specification as for differences in outcomes.
  - To check randomisation was implemented correctly, use omnibus test.

[1] Altman (1985) notes that such tests amount to assessing the probability of something having occurred by chance when you know that it did occur by chance. "Such a procedure is clearly absurd".

[2] Imbens and Rubin, 2015

Forcing Balance

Without further information:

Randomisation conditional on N_1 is best we can do to achieve $(Y(0), Y(1)) \perp D$ in sample.

But generally have more information:

$$\mathbb{E}[Y_i(D_i)] = f(X_i^+, X_i^-, D_i) \tag{1}$$

where X_i^+ are observable and X_i^- non-observable covariates.

Cannot force balance of potential outcomes, but...
...can try to force balance of X_i^+. Reduce $V(\hat{\beta})$.

Plan for Today

1. Unbiased Estimation
2. Balance
3. Stratification/Paired Experiments
4. Power
5. Control/Bad Control
6. Attrition
7. Canonical Experimental Designs

Stratification: Simple Example

Idea: Do not leave imbalance of important covariates to chance.

Formally: Restrict assignment mechanisms.

Example 1: Potential outcomes determines as $Y_i(D_i) = g_i + 1 \times D_i$, where $g_i \in \{1, 2\}$ is only covariate, $N = 4$, $N_1 = 2$ and $i = 1, 2$ only have $g_i = 1$. Note: true $\beta = 1$.

Assignment	d_1	d_2	d_3	d_4	$\hat{\beta}$
#1	0	1	0	1	1
#2	0	1	1	0	1
#3	1	0	0	1	1
#4	1	0	1	0	1
#5	1	1	0	0	0
#6	0	0	1	1	2

$\Rightarrow V(\hat{\beta})$ lower by excluding last two assignments! Stratification.

Stratification: Practicalities

Extends to several categorical variables. **How?**

$\Rightarrow$ In Problem Set 2 you are asked to simulate effect of stratification on the variance of the estimator.

Implementation in STATA:

```
set seed 20230323
gen random = runiform()
sort cat_var1 cat_var2 ...  random
gen treatment = mod(_n,2)
```

Stratification: Practicalities

Question 1: Which variables should we stratify on?

General recommendation: covariates strongly related to the outcome.[3] Why?

Question 2: And when covariates are continuous, or cells sparse?

[3] See Bruhn and McKenzie, 2009; Glennerster and Takavarasha, 2013.

Stratification: Implementations

What do people do? (Bruhn and McKenzie, 2009)

1. Pure Randomisation.
2. Re-Randomisation.
   - Subjectively decide whether to make another draw; or
   - Re-randomise until some statistic of balance is achieved: or
   - Choose assignment with best balance amongst N draws.
3. Matched-Pair ('blocking'): stratified randomisation with 2 units in each stratum.

   Recent insight:

   Matched-Pair designs optimal (under some conditions, in some sense).

Stratification: Matched-Pair Designs

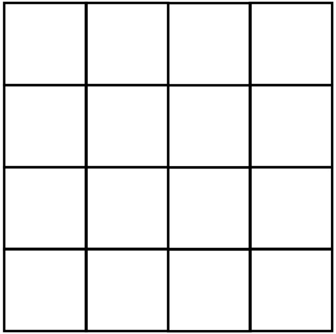
Matched-Pair Design is stratified randomisation with two units in each stratum.

1. Bai (AER, 2022):
   "Optimality of Matched-Pair Designs in Randomized Controlled Trials"
   - Shows optimality (in MSE sense) of a specific matched-pair design: Calculate $\mathbb{E}[Y_i(1)|X_i] + \mathbb{E}[Y_i(0)|X_i]$, match adjacent units on that "simple" scalar function.
   - Problem: we do not know that sum! Can be estimated using pilot data with large class of estimators, including machine learning techniques.
2. Barrios (2015) – special case:
   "Optimal Stratification in Randomized Experiments"
   - With homogeneous treatment effects, best way to choose matched-pairs is to match on $\mathbb{E}[Y_i(0)|X_i]$.
   - Baseline, but no pilot data needed.

Stratification: Taking it to the Extreme

Next level: No randomization.

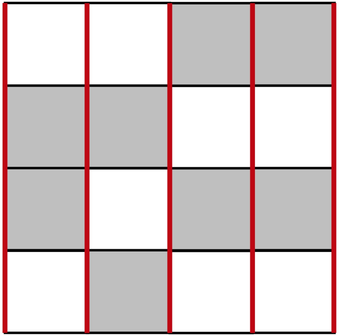
Consider this example, a 'field'. *What is optimal?*



Stratification: Taking it to the Extreme

Next level: No randomization.

Consider this example, a 'field'. *What is optimal?*

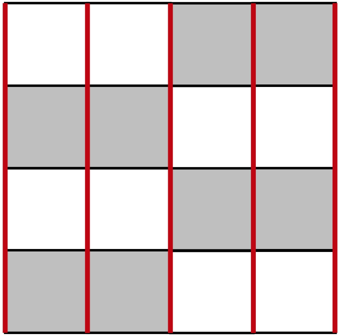


Stratification by x_1? Potentially unbalanced marginal distribution of x_2.

Stratification: Taking it to the Extreme

Next level: No randomization.

Consider this example, a ’field’. *What is optimal?*

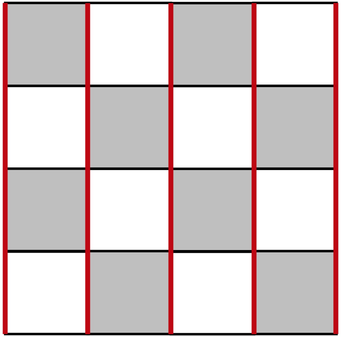


Something like this? Not maximizing power.

Stratification: Taking it to the Extreme

Next level: No randomization.

Consider this example, a ’field’. *What is optimal?*



Balance joint distribution of x_1, x_2.

But only two assignments achieve that! Randomization?

Stratification: Taking it to the Extreme

Given any prior on how potential outcomes are generated...

...argument is very general! (Kasy, 2016)

- For any assignment **D**, calculate loss function (MSE, V, ...).
- Expected loss is average across potential assignments.
- Find subset of assignments such that loss is minimised.
- Typically those are two.

Controlled Trial, not Randomized Controlled Trial

- Stratification special case where subset larger than 2.

Benefits

Notice: Stratification has benefits ex-post, and ex-ante!

After Kasy: *Why do we randomize then?*

- Optimal solution might be hard to find.
- Might set up the decision problem differently.
- Randomization Inference!

Plan for Today

1. Unbiased Estimation
2. Balance
3. Stratification/Paired Experiments
4. Power
5. Control/Bad Control
6. Attrition
7. Canonical Experimental Designs

Power: The Concept

Talked about experimental design choices that lead to lower variance of the estimator. Closely related concept: Power.

1. Eventually want to test hypothesis H_0 vs alternative H_1.
2. Two types of errors we can make:
   1. Type I error: reject H_0 when H_0 is true.
      Probability of type I error chosen by setting α.
   2. Type II error: fail to reject H_0 when H_0 is false.
      Probability of type II error depends on true effect size, experimental set-up and estimator/test statistics..

Power: How to increase it?

How to increase power?

1. Reduce the variance of the estimator:
   - Stratification (see before).
   - Baseline Controls (see later).
   - Sample Size.[4]
   - Measurement!
2. Choice of test statistic.

[4] When starting a project, make sure it is 'powered' to detect expected effects with reasonably high probability. Funders want to see this in grant applications.

Plan for Today

1. Unbiased Estimation
2. Balance
3. Stratification/Paired Experiments
4. Power
5. **Control/Bad Control**
6. Attrition
7. Canonical Experimental Designs

Analysis: Control Variables

- Control variables: pre-determined, observed variables.
- If simple random assignment: no need to include control to guarantee **unbiasedness**...
- ... but might make sense to increase **power**.
- If data is generated by *conditional random assignment*, need to hold variables you conditioned on in randomisation constant in the analysis. (Only *conditional* independence assumption will surely hold, will come back to this.)

Analysis: Bad Control

Control variables can not be outcomes themselves!

Mathematical reason:

$$\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$$

does not imply

$$\mathbb{E}[Y_i(0)|D_i = 1, X_i = x] = \mathbb{E}[Y_i(0)|D_i = 0, X_i = x]$$

where X_i refers to the (observed) outcomes for the covariate for unit i. Such an X_i would be a bad control.

Analysis: Bad Control

Example: Think of Y as achievement, D as class size (0 if large/1 if small), and X as parental help (1 if help/0 if not). D_i assigned by classical randomized experiment.
Plausibly parents' help responds to variation in D, so X is an outcome. Define X_i as potential outcome analogous to how we defined Y_i.

Suppose 'control for X' in difference in outcomes:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1, X_i = 1] - \mathbb{E}[Y_i|D_i = 0, X_i = 1] \\
&= \mathbb{E}[Y_i(1)|D_i = 1, X_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0, X_i = 1] \\
&= \mathbb{E}[Y_i(1)|X_i(1) = 1] - \mathbb{E}[Y_i(0)|X_i(0) = 1] \quad | \quad \text{by Random Assignment} \\
&= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i(1) = 1]}_{\text{'some' causal effect}} + \underbrace{\mathbb{E}[Y_i(0)|X_i(1) = 1] - \mathbb{E}[Y_i(0)|X_i(0) = 1]}_{\text{selection bias}}
\end{aligned}
$$

Analysis: Bad Control

$$\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i(1) = 1]}_{\text{'some' causal effect}} + \underbrace{\mathbb{E}[Y_i(0)|X_i(1) = 1] - \mathbb{E}[Y_i(0)|X_i(0) = 1]}_{\text{selection bias}}$$

Causal effect: for population that helps when class size is small.

Compositional bias: control sample *shouls be* those that help when class size is small, but *it is* those that help when class size is large - who are probably more than those who help when size is small. Probably special group!

Same argument applies *generally* when restricting to subgroup with some outcome. Example: effect of training on wages – which are only observed for employed.

Extensive margin: easy; intensive margin: infeasible.

Analysis: Estimate only reduced form of effects

Example: Educational production function $Y = f(D, X, \theta)$. Suppose that X responds to D.

Comparing outcomes across treatment arms, we estimate:

$$\Delta Y = \frac{\partial f}{\partial D}\Delta D + \frac{\partial f}{\partial X}\frac{\partial X}{\partial D}\Delta D \tag{2}$$

- Sometimes that is what you are interested in.
- But for other questions you might need $\frac{\partial f}{\partial D}$ or $\frac{\partial f}{\partial X}$.

P. Fredriksson: “One reason for the slight dismay in certain quarters over the experimental approach.”

Plan for Today

1. Unbiased Estimation
2. Balance
3. Stratification/Paired Experiments
4. Power
5. Control/Bad Control
6. **Attrition**
7. Canonical Experimental Designs

Attrition: Practical Advise

Attrition is common in experiments:

cannot obtain follow-up data for some treated and some un-treated observations.

- Generally good to avoid!
- Unproblematic when attrition is unrelated to potential outcomes.

  Unfortunately impossible to know!
- Commonly accepted test: attrition rates unrelated to treatment.

  Might still be that attrition is selected differently in T and C.
- When attrition rates differ across T and C:

  Some bounding exercise, commonly Lee (2009) bounds.

Attrition: Lee (2009) Bounds

Suppose the attrition rate is higher in C than in T.

Worry: Those with 'best' $Y_i(0)$ cannot be observed in Control. Comparison of T with C exaggerates ATE.

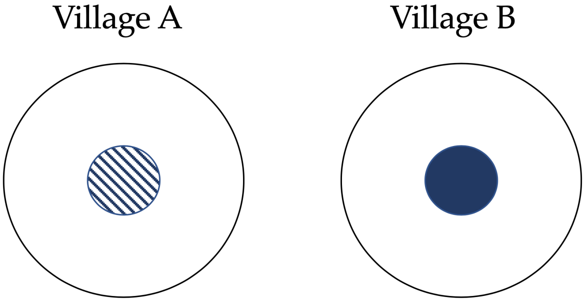
Lee Bounds are an extreme bounding exercise:

- Drop the 'best' from T, s.t. T and C have same attr. rate, and rerun analysis.
- Drop the 'worst' from T, s.t. T and C have same attr. rate, and rerun analysis.
- Report results from both exercises as bounds.

Plan for Today

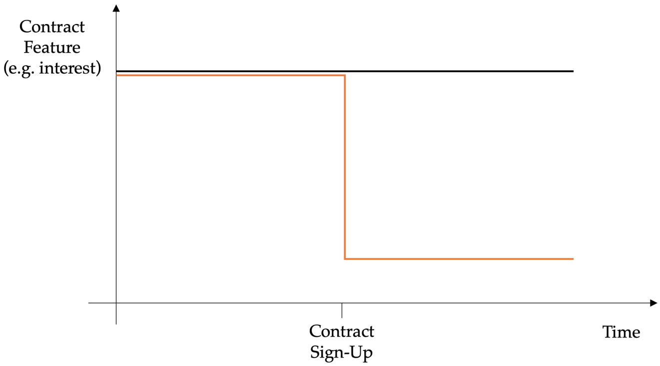
Spill-Overs: *Miguel and Kremer* (2004)

Suppose you believe your treatment might affect other units, i.e. has **spill-over effects.** *How can you estimate those?*



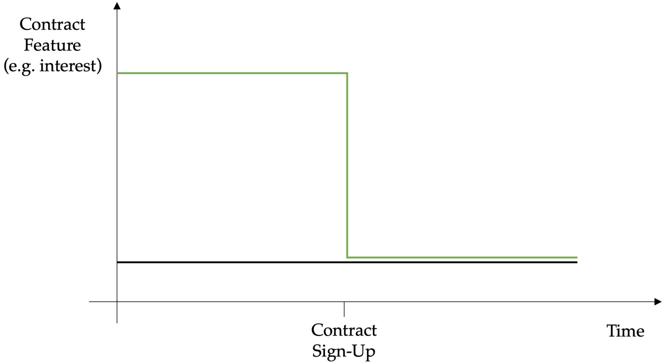
Moral Hazard and Adverse Selection: *Karlan and Zinman* (2009)

Distinguishing Moral Hazard and Adverse Selection *empirically* is generally hard.
Karlan and Zinman figured out one way:



Moral Hazard and Adverse Selection: *Karlan and Zinman* (2009)

Distinguishing Moral Hazard and Adverse Selection *empirically* is generally hard.
Karlan and Zinman figured out one way:



Randomized Experiments

- Experiments important to understand on their own right.

  Today just a small introduction.
- Modern econometric approaches approximate that ideal.
- The experimental ideal will often be very useful to think in *observational studies* whether you estimate causal effects.
- Issues we discussed will come up, more or less explicitly: assignment mechanism, conditional independence, balance, biases, reduced form estimation, average treatment effects for subpopulations...

Questions?

References

1. Angrist and Pischke (2008): Chapter 2 [introduction]; Chapter 3.2.3 [bad controls]
2. Imbens and Rubin (2015): Chapter 4 [introduction]; Chapter 5 [Fisher inference], Chapter 7.5 [controls]
3. Kasy, Maximilian (2016) “Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead” Political Analysis: 1-15. [stratification]
4. Lee, D. S. (2009) “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects”. Review of Economic Studies 76: 1071–1102. [attrition]
5. Bai, Yuehao (2022) “Optimality of Matched-Pair Designs in Randomized Controlled Trials”. Working Paper. [matched-pair designs]
6. Barrios, Thomas (2014) ”Optimal Stratification in Randomized Experiments”. Working Paper. [matched-pair design]
7. Bruhn, M. and D. McKenzie (2009) “In Pursuit of Balance: Randomization in Practice in Development Field Experiments”. American Economic Journal: Applied Economics 1:4, 200-232.
8. Karlan, D. and J. Zinman (2009) “Observing Unobservables: Indentifying Information Asymmetries With a Consumer Credit Field Experiment”. Econometrica 77:6, 1993-2008.
9. Miguel and Kremer (2004) “Worms: Indentifying Impacts on Education and Health in the Presence of Treatment Externalities”. 72:1, 159-217.