Session II
Randomized experiments and conditional independence

Evaluating public policies

Arthur Heim (PSE & Cnaf)

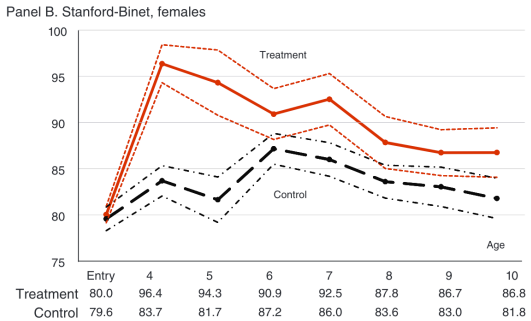


FIGURE 1. STANFORD-BINET IQ TEST SCORES BY GENDER AND TREATMENT STATUS

Notes: Bold lines display mean IQs. Fine lines represent standard errors for the corresponding means (one standard error above and below). For a detailed description of the cognitive measures and results for other IQ tests, see online Appendix B. Numbers below each chart are treatment [illegible] Source: Heckman, Pinto, and Savelyev (2013) effect of Perry Preschool Program on IQ for girls

Outline

introduction

What we have seen so far

- We introduced the notations and framework of the Rubin 1974 causal model
- The core : postulate the existence of **potential outcomes** corresponding to the possible states of the world according to the presence or absence of the policy to be evaluated.
- Some causal parameters of interest can be identified and estimated with appropriate hypotheses, design and measures.
- = **Identification strategy**

introduction

Today's agenda

- Randomisation: Where it comes from, how it's done.
- Regressions

Outline

Randomisation: An introduction

The "magic" of randomisation

What is and what's not randomisation (1/2)

- Randomisation is the process of making something random.
- A random process is a sequence of random variables describing a process whose outcomes **do not follow a deterministic pattern**, but one that can be described by **probability distributions**.
- For example, a **random sample** of individuals from a population refers to a sample where every individual has a known probability of being sampled. We use this sampling probability to infer plausible values of parameters of the full population with the statistics from the sample
- Large random sample ensure representativeness of the population and more precise estimates i.e. less **sampling variability**
- The sampling variability of an estimate is a measure of how much the estimate may vary from sample to sample

Randomisation: An introduction

The "magic" of randomisation

What is and what's not randomisation (2/2)

- Randomized experiments are **lotteries** that randomly assign subjects from a sample to research groups, each of which is offered a different treatment
- Random assignment ensure balance between groups $\Rightarrow$ no ex-ante differences ; post-treatment average differences estimate the average treatment effect in this sample.
- The sample from a randomized experiment **may or may not be a random sample**. These are two distinct issues.
- Randomisation ensures **internal validity** i.e. unbiased estimator of treatment effects in this sample.
- Large randomized experiments give more precise estimators, but inference to a larger population depend on the sampling process, participants' awareness, the specific time and location, etc.
- Issues of **external validity**

Randomisation: An introduction

Experiments in social sciences

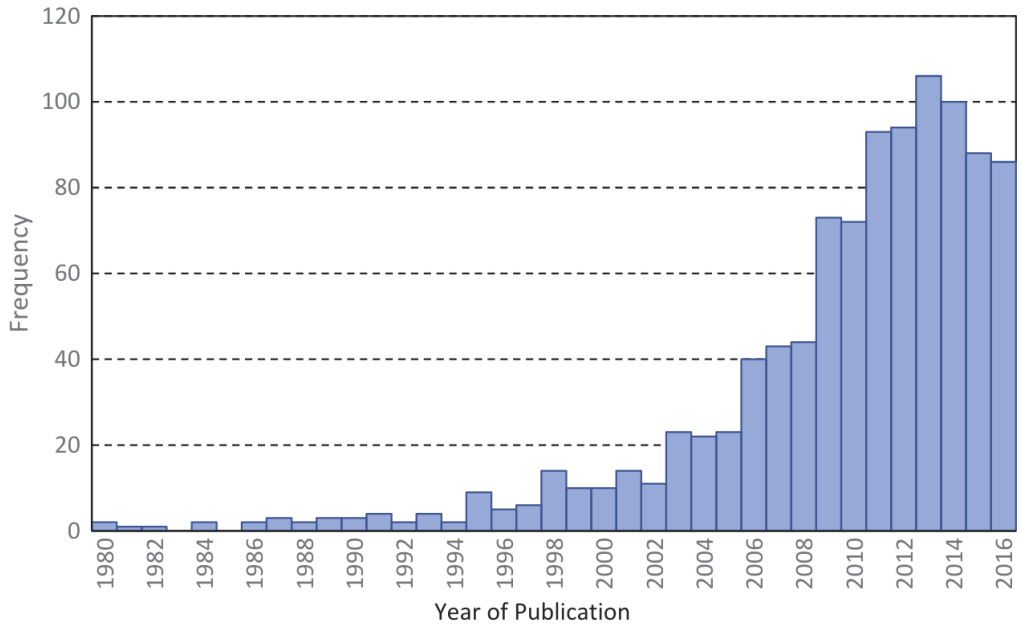
Origin and diffusion

- The method is usually attributed to Fischer (1935) and Neyman (1934)
- When the method is implemented properly, average differences in future outcomes for experimental groups provide unbiased estimates of the impacts of the treatments offered
- From 1945 to 2000 over 350 000 randomized clinical trials have been conducted
- The use of randomized experiments for social research has greatly increased since the U S War on Poverty in the 1960's

Randomisation: An introduction

Experiments in social sciences

Figure 1: Number of randomised controlled trials in education completed internationally between 1980 and 2016 reviewed by Connolly, Keenan, and Urbanska (2018)



Randomisation: An introduction

Experiments in social sciences

Some of my favorite experiments

- **Early childhood education**:
  - Heckman, Pinto, and Savelyev 2013
  - Orla Doyle. 2020. "The First 2,000 Days and Child Skills." *Journal of Political Economy* 128, no. 6 (June): 2067–2122
- **Cash transfers**:
  - Lisa Gennetian et al. 2022. *Unconditional Cash and Family Investments in Infants: Evidence from a Large-Scale Cash Transfer Experiment in the U.S.* w30379. Cambridge, MA: National Bureau of Economic Research, August
  - Cornelius Christian and Christopher Roth. 2016. "Can Cash Transfers Prevent Suicides? Experimental Evidence from Indonesia." *SSRN Journal*
- **Labor market**
  - Crépon et al. 2013a (To analyze next week !)
  - Luc Behaghel, Bruno Crépon, and Marc Gurgand. 2014. "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment." *American Economic Journal: Applied Economics* 6, no. 4 (October 1, 2014): 142–174

Randomisation: An introduction

Experiments in social sciences

Some of my favorite experiments

- **Reducing hate**:
  - Dominik Hangartner et al. 2021. "Empathy-Based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment." *Proceedings of the National Academy of Sciences* 118, no. 50 (December 14, 2021): e2116310118
  - David Broockman and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352, no. 6282 (April 8, 2016): 220–224
- **Toss coin for major life decision**
  - Steven D Levitt. 2021. "Heads or Tails: The Impact of a Coin Toss on Major Life Decisions and Subsequent Happiness." *The Review of Economic Studies* 88, no. 1 (January 1, 2021): 378–405
- **Psychology: poverty and cognitive functioning**
  - A. Mani et al. 2013. "Poverty Impedes Cognitive Function." *Science* 341, no. 6149 (August 30, 2013): 976–980

Randomisation : a mansplaining story

The lady tasting tea

- In the 1920s, statistician Ronald Fisher introduced a new method for statistical inference called *randomization inference*
- One famous example of this method is Fisher's "Lady Tasting Tea" experiment, reported in his book "The design of experiments" (Fischer 1935)
- **Context:** It's a summer party in Cambridge, and Fisher and other academics are drinking tea when a woman made an bold claim.
- Muriel Bristol, a British phycologist claimed to be able to tell **whether the tea or the milk was added first** to a cup.
- "Nonsense," returned Fisher, smiling, "Surely it makes no difference." But she maintained, with emphasis, that of course it did. and her soon-to-be-husband added "Let's test her."
- Fisher randomly arranged eight cups of tea, four made with milk added first and four made with tea added first, and presented them to Lady Bristol.
- Lady Bristol tasted the 8 cups and **correctly classified** the 4 cups with milk poured first (the way she prefers)
- Would that be convincing enough for R. Fischer ?

Randomisation : a mansplaining story

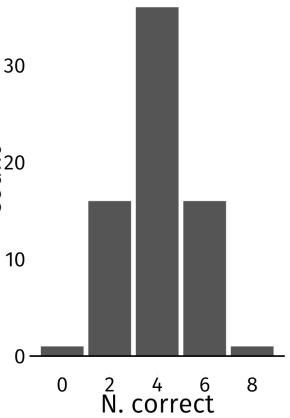
The lady tasting tea

1. First, we need to determine the **null hypothesis**, which is the hypothesis that lady Bristol is **no better** at distinguishing between the two types of tea **than chance alone** would allow.
2. Next, we need to determine the **probability of obtaining the observed result** (i.e., correctly identifying all four cups of tea made with milk added first) **under the null hypothesis**. In this case, the probability is simply the number of ways to arrange the cups such that she would get all four correct divided by the total number of possible arrangements.
   There are $\binom{8}{4} = 70$ ways of selecting 4 cups among 8. Under the null hypothesis, there is thus **one** set of 4 cups chosen among 8 cups **out of 70** possible combinations of 4 cups among those 8 cups that she could pick by chance.
   $\frac{1}{70}$ =1.43 %.
3. The probability that she gave the right combination by sheer luck is less than 2 %, which correspond to the exact p-value of the test (Fisher's exact p-value)

Randomisation : a mansplaining story

Figure 2: The statistical test is permutation under the null

Cup	*Guess*	*Truth*	P_1	P_2	P_3	...	P_{70}
1	m	m	m	m	m		t
2	t	t	~~m~~	~~m~~	~~m~~		t
3	t	t	~~m~~	~~m~~	~~m~~		t
4	m	m	m	t	t		t
5	m	m	t	m	t		m
6	t	t	t	t	~~m~~		~~m~~
7	t	t	t	t	t		~~m~~
8	m	m	t	t	t		m
		8/8	4/8	4/8	2/8		4/8



Randomisation : a mansplaining story

How does that relate to impact evaluation ?

- Fisher was interested in testing **sharp null hypotheses**, that is, null hypotheses under which we can infer all the missing potential outcomes from the observed ones.
- Going back to Rubin's notation, a sharp null hypothesis for the treatment-control problem is :

$$\mathcal{H}_0 : Y_i(1) = Y_i(0) \ \forall \ i$$

- No treatment effect for anyone. The implicit alternative hypothesis is that there is at least one unit i such that $Y_i(0) \neq Y_i(1)$.
- Given the sharp null hypothesis, we can infer all the missing potential outcomes through permutation.

Outline

Randomisation: Pro, & and How to

Reminder from last week

- Consider a policy that either make people "treated" or "untreated" (e.g. being in a small or large class)
- Let i denote a principal sample unit (PSU) (individual, household, firm...) and let Y be an outcome of interest (e.g. test score at the end of 3^{rd} grade)
- Let D_i be the observed variable indicating treatment status $D_i = \mathbb{1}(\text{Treated})$
- Every individual can theoretically be treated or untreated and for a given individual i, there exist different potential values for their outcomes: $Y_i(1)$, their outcome when treated and $Y_i(0)$ when they aren't.
- When individual i is treated their **observed** outcome is $Y_i = Y_i(1)$, when they are not we observe $Y_i = Y_i(0)$.
- Potential outcomes can be linked to observed outcome and treatment through a switching equation

$$\begin{aligned} Y_i = &\quad D_i Y_i(1) + (1 - D_i) Y_i(0) \\ Y_i = &\quad Y_i(0) + (Y_i(1) - Y_i(0) D_i \end{aligned} \tag{1}$$

Reminder from last week

Average observed differences and selection bias

- We can decompose the simple average difference in ouctome (SDO) by treatment status to extract parameters of interests
- Under SUTVA, observed outcomes Y_i reveal potential outcomes $Y_i(\cdot)$ for the relevant units

$$\underbrace{\mathbb{E}\Big[Y_i|D_i=1\Big]-\mathbb{E}\Big[Y_i|D_i=0\Big]}_{\text{Simple difference (SDO)}}=\mathbb{E}\Big[Y_i(1)|D_i=1\Big]-\mathbb{E}\Big[Y_i(0)|D_i=0\Big]$$

- We add and substract counterfactual values for treated individuals:

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

$$=\underbrace{\mathbb{E}\Big[Y_i(1)-Y_i(0)|D_i=1\Big]}_{\text{ATT}}+\underbrace{\mathbb{E}\Big[Y_i(0)|D_i=1\Big]-\mathbb{E}\Big[Y_i(0)|D_i=0\Big]}_{\text{Selection bias}}$$

- The selection bias measures how different the treated and control groups are (on average) in terms of their potential outcomes **in the absence of treatment**

Randomisation: Pro, & and How to

Why randomisation removes selection bias

- Randomization makes the random treatment status D_i independent of potential outcomes $Y_i(1)$ and $Y_i(0)$: $\Rightarrow D_i \perp Y_i(0), Y_i(1)$
- **Reminder**: If two random variables Y and X are independent, then

$$\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y]$$

- Thus, randomization implies that $\mathbb{E}[Y_i(0) \mid D_i = 1] = \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i(0)]$ which implies in turn that the selection bias vanishes out:

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

- The mean SDO in the two groups is an estimate of the **ATT parameter**, but <u>also</u> of the **Average Treatment Effect (ATE)**:

$$\begin{aligned}\mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] &= \underbrace{\mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1]}_{\text{ATET}} \\ &= \underbrace{\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]}_{ATE}\end{aligned}$$

Strenghs and weakness of randomization

Strenghts

I: Removing selection bias

- Random assignment remove selection biais and retrieve causal estimates when well implemented
- On average, research groups have the same characteristics *ex-ante* (even unobservables)
- ♭ In small samples, random assignment can leave some imbalance across groups. Those are **sampling errors**

Strenghs and weakness of randomization

Strenghts

I: Removing selection bias

- Random assignment remove selection biais and retrieve causal estimates when well implemented
- On average, research groups have the same characteristics *ex-ante* (even unobservables)
- ♭ In small samples, random assignment can leave some imbalance across groups. Those are **sampling errors**

II: Controlling uncertainty

- In a design-based framework we can decompose the total variability between sampling variability and variability due to randomization (Abadie et al. 2020)
- Different approaches for inference

Strenghs and weakness of randomization

Strenghts

I: Removing selection bias

- Random assignment remove selection biais and retrieve causal estimates when well implemented
- On average, research groups have the same characteristics *ex-ante* (even unobservables)
- ♭ In small samples, random assignment can leave some imbalance across groups. Those are **sampling errors**

II: Controlling uncertainty

- In a design-based framework we can decompose the total variability between sampling variability and variability due to randomization (Abadie et al. 2020)
- Different approaches for inference

III: No extra assumption required

- When treatment is truely randomly assign and the experiment is not compromized by attrition or other issues, then there is no need for extra modeling assumption

Strenghs and weakness of randomization

Weaknesses

Acceptability and ethics

- Policy makers may find it hard to select units randomly as some people will necessarily be denied access (legal issues, equality, political cost).
- Population may reject such design if they perceive it as unfair, or fostering injustice, or dangerous
- Ethics of research with human subject (See Resnik 2018): principle of "**clinical equipoise**", information and consents, risk mitigation, etc.

Strenghs and weakness of randomization

Weaknesses

Acceptability and ethics

- Policy makers may find it hard to select units randomly as some people will necessarily be denied access (legal issues, equality, political cost).
- Population may reject such design if they perceive it as unfair, or fostering injustice, or dangerous
- Ethics of research with human subject (See Resnik 2018): principle of "**clinical equipoise**", information and consents, risk mitigation, etc.

Bias in experiments

- The sample from an experiment may be different from a more general population even when randomly sampled because people may react to the experiment itself.
- Replicability, power and sample size
- Bad research practices: threatening participants to increase follow-up, adding observation up to a point where results are satisfactory

Strenghs and weakness of randomization

Weaknesses

Acceptability and ethics

- Policy makers may find it hard to select units randomly as some people will necessarily be denied access (legal issues, equality, political cost).
- Population may reject such design if they perceive it as unfair, or fostering injustice, or dangerous
- Ethics of research with human subject (See Resnik 2018): principle of "**clinical equipoise**", information and consents, risk mitigation, etc.

Bias in experiments

- The sample from an experiment may be different from a more general population even when randomly sampled because people may react to the experiment itself.
- Replicability, power and sample size
- Bad research practices: threatening participants to increase follow-up, adding observation up to a point where results are satisfactory

Substitution bias

- The treatment may be a subsitute to another policy and the counterfactual may not be well identified. Example: Head Start provided formal childcare in the US but the counterfactual is a mixture of parental care and already available informal childcare. See e.g. Kline and Walters (2016)

Strenghs and weakness of randomization

Using real world constraints for experiments

Political advantage of randomisation

- Lotteries are simple, easily understood and may be very transparent
- Useful and legit when there are no other reason to select participants
- May be seen as fair compared to other criteria

Strenghs and weakness of randomization

Using real world constraints for experiments

Political advantage of randomisation

- Lotteries are simple, easily understood and may be very transparent
- Useful and legit when there are no other reason to select participants
- May be seen as fair compared to other criteria

Ressource limitation of staggered adoption

- Many policies have limited ressources and cannot reach the whole population immediately.
- Sometimes there are more eligible people than actual participants, we can manipulate information or incentives to foster participation
- Staggered adoption affect different individual subsequently and provide sources for comparisons
- These constraints can be use to implement policy evaluation, randomly if possible

How to randomise ?

1. Individual random assignment
   - **Bernoulli trial**:flip coins for each person in the experiment.
     - may not balance group sizes especially in smaller samples.
   - **Completely randomized experiments**: a fixed number of units, say N_t, is drawn at random from the population of N units to receive the active treatment, with the remaining $N_c = N - N_t$ assigned to the control group. Each unit has the same treatment probability
     - Bad luck may bring unbalance on some important characteristics, especially in small samples.

Randomisation in practice

How to randomise

2. More sophisticated random assignments
   - **Block random assignment**:From a set of known attributes, Classify individuals in J mutually exclusive blocks (or strata) and run a completly randomized experiment within block.
     - Ensure balance w.r.t. attributes that define blocks $\rightarrow$ more precise estimate because it control between-block differences and only uses within-block variation.
     - **Conditional ignorability**. outcomes $Y_i(1)$ and $Y_i(0)$ are **conditionally independent** of the assignment variable D_i given that the unit belongs to block $b_j = (B_i = b_j)$
   - ***Clustered randomized design***: If units i belong to a larger structure (e.g. classroom, school, village,...) with similar characteristics or which cannot be separated, one can randomly assign treatment to the cluster, thus treating everyone in treated clusters and nobody in untreated clusters.
     - The principal sampling unit is the cluster, treatment effect comes from variation across clusters. Large precision loss.

Can use more sophisticated design with block-cluster random assignment, or block-pair random assignment

Randomisation in practice

How to randomise

3. Randomisations using practical constrains
   - **Randomisation *at the bubble***:Some policies may be determined by a set of criteria and you want to test wether providing access to those at the margin of these criteria would benefit.
   - Take individuals at the margin at randomly assign treatment with the appropriate design.
     - Only valid for the marginal population
   - ***Phase in design***: Take advantage of staggered adoption by randomly assigning order or waves of treatment
     - everybody gets treated eventually so people may be more willing to accept randomisation
     - If outcomes are measured for those who are treated, we cannot have long term impacts because everybody gets treated.
     - Not a problem when outcomes are measured at another level (for phase-in training for teachers where outcomes are measured over students)

Outline

The analysis of randomised experiments

Exact P-values for Sharp Null Hypotheses

- In order to conduct randomization inference, we need to supply 1) a test statistic, 2) a null hypothesis, and 3) a randomization procedure.
- a sharp null hypothesis for the treatment-control problem is :

$$\mathcal{H}_0 : Y_i(1) = Y_i(0) \ \forall \ i$$

$$T^{\text{ave}} = \bar{Y}_{\text{t}}^{\text{obs}} - \bar{Y}_{\text{c}}^{\text{obs}} = \frac{1}{N_{\text{t}}} \sum_{i:D_i=1} Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}} \sum_{i:D_i=0} Y_i^{\text{obs}}.$$

- We can calculate the probability, over the randomization distribution, of the statistic taking on a value as large, in absolute value, as the actual value given the actual treatment assigned. This calculation gives us the p-value for this particular null hypothesis.
- In other words, we compute the mean difference over all possible assignment of the treatment status (i.e. we re-arrange observations in treatment and control groups), and compare the share of mean differences that are higher/lower than the observed mean differences.
- ♭ The number of possible permutations rise exponentially with sample size and rapidely becomes computationaly infeasible. Instead, we can randomly sample over the permutation distribution.

The analysis of randomised experiments

Exact P-values for Sharp Null Hypotheses

Simulated example

- Imagine a hypothetical experiment in which 2 of 7 villages randomly elect a female council head and the outcome is the share of the local budget allocated to water sanitation per inhabitant (in $) (example in the RI package for R)

City	Z	Y
Village 1	1	25
Village 2	0	15
Village 3	0	20
Village 4	0	20
Village 5	0	10
Village 6	0	15
Village 7	1	30

term	estimate	p.value
Z	11.5	0.047619

The analysis of randomised experiments

Exact P-values for Sharp Null Hypotheses

Simulated example

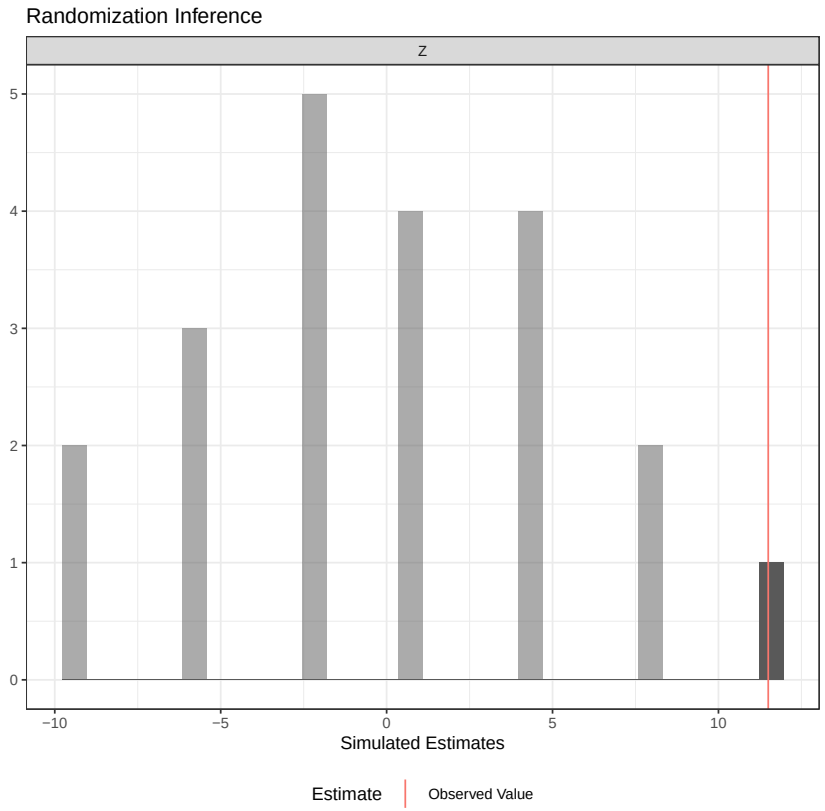


Figure 3: Randomisation inference: illustration

The analysis of randomised experiments

Exact P-values for Sharp Null Hypotheses

On exact p-value

- Although the observed outcomes do not change for any unit under the null hypothesis, the value of the statistic changes because who is in the treatment group and who is in the control group changes.
- The p-value associated with this statistic is 0.048, suggesting we should reject that women-led city council has no effect on sanitation spending.
- Randomisation inference can accommodate more sophisticated designs, larger sample sizes (we sample the permutation distribution instead of computing all possible permutations), different statistics.
- Sharp null test are very restrictive, test the absence of presence of treatment effect for at least one unit, not the average difference. Inference for average treatment effect derives from Neyman (1934) work.

The analysis of randomised experiments

Randomization Inference for Average Treatment Effects

A simple difference in mean estimator

- Two conceptual views of an experiment:
  1. Finite population: analysis of completely randomized experiments, taking as fixed the potential outcomes in the population and the variability only comes from the randomisation.
  2. Infinite population: random sample from an infinite population, use large sample approximation for inference
- Usual approach: large sample approximation.
- Consider a sample of size $N = N_0 + N_1$ where the N_1 individuals were randomly assigned to treatment and comply with their assignment, and N_0 act as controls.
- The average treatment effect on the sample is :

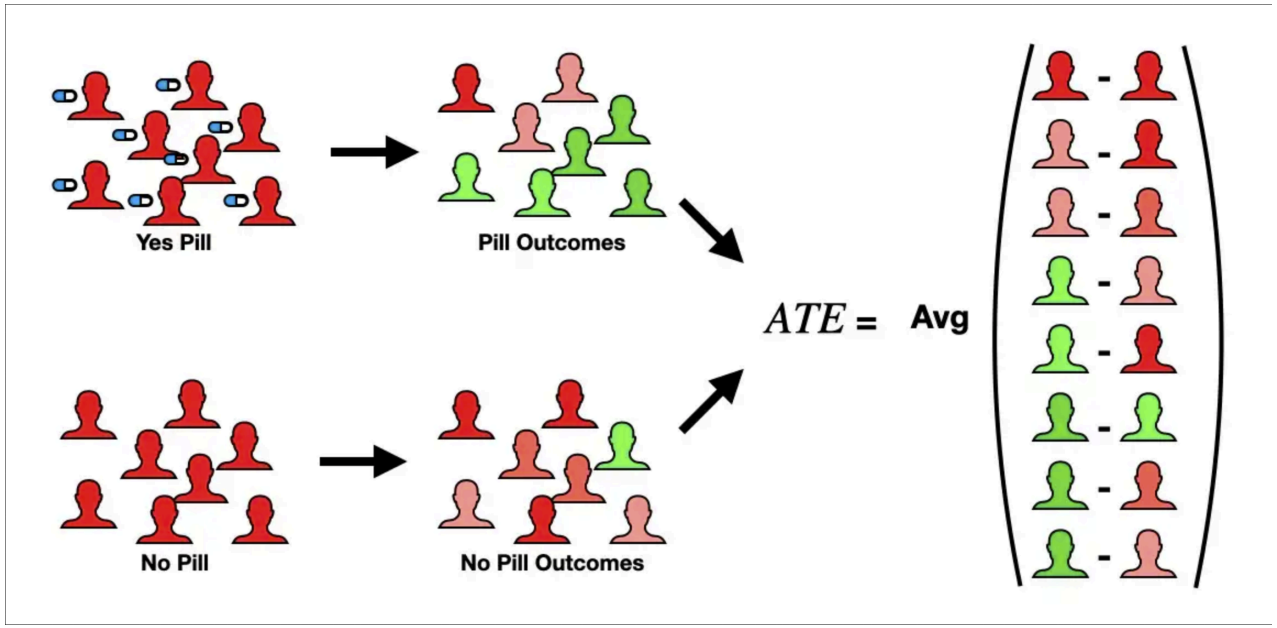
$$\begin{aligned} \hat{\tau} = \bar{Y_1} - \bar{Y_0} &= \frac{1}{N_1}\sum_{N_1} Y_i - \frac{1}{N_0}\sum_{N_0} Y_i \\ &= \frac{1}{N_1}\sum_{N_1} Y_i(1) - \frac{1}{N_0}\sum_{N_0} Y_i(0) \end{aligned}$$

The analysis of randomised experiments

Randomization Inference for Average Treatment Effects

A simple difference in mean estimator

Figure 4: Simple randomisation: illustration



The analysis of randomised experiments

Randomization Inference for Average Treatment Effects

Variance in finite population

Imbens and Rubin (2015) show that sampling variance over the randomisation distribution is:

$$\mathbb{V}(\hat{\tau}) = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{1,0}^2}{N}$$

where S_0^2 and S_1^2 are the variances of $Y_i(0)$ and $Y_i(1)$ in the sample, defined as:

$$S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0) - \bar{Y}(0)\right)^2, \quad \text{and} \quad S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \bar{Y}(1)\right)^2$$

and $S_{1,0}^2$ is the **impossible to observe** sample variance of the unit-level treatment effects, defined as:

$$S_{1,0}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0))\right)^2$$

The analysis of randomised experiments

Randomization Inference for Average Treatment Effects

Neyman Variance and population variance

- In practice researchers therefore use the estimator for $\mathbb{V}(\hat{\tau})$ based on estimating the first two terms by S_0^2 and S_1^2, and ignoring the third term:

$$\mathbb{V}_{NEYMAN} = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1}$$

- This leads in general to an upwardly biased estimator for $\mathbb{V}(\hat{\tau})$, and thus too conservative confidence intervals
- There are two important cases where the bias vanishes
  1. if the treatment effect is constant the third term is zero
  2. if we view the sample at hand as a random sample from an infinite population, then $\mathbb{V}_{NEYMAN}$ is unbiased for the variance of $\mathbb{V}(\hat{\tau})$ viewed as an estimator of the population average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$
- Assuming a large population, the **standard error** of the treatment effect is:

$$SE_{\hat{\tau}} = SE(\bar{Y}_1 - \bar{Y}_0) = \sqrt{\frac{S_0^2}{N_0} + \frac{S_1^2}{N_1}} \tag{2}$$

▸ Proof

The analysis of randomised experiments

Randomization Inference for Average Treatment Effects

Confidence intervals

- We base the confidence interval on a normal approximation to the randomization distribution of $\hat{\tau}$.
- If we wish to construct a central confidence interval with nominal confidence level $(1-\alpha) \times 100\%$, as usual we look up the $\frac{\alpha}{2}$ and $\frac{1-\alpha}{2}$ quantiles of the standard normal distribution, denoted by $z_{\alpha/2}$, and construct the confidence interval:

$$CI_{1-\alpha}(\hat{\tau}) = (\hat{\tau} - z_{\alpha/2} \cdot SE_{\hat{\tau}}, \hat{\tau} + z_{\alpha/2} \cdot SE_{\hat{\tau}}).$$

- When $\alpha = 5$ %, $z_{.05/2} \approx 1.96$, when $\alpha = 10$ %,$z_{.1/2} \approx 1.645$
- This approximation applies when using any estimate of the sampling variance, and, in large samples, the resulting intervals are valid confidence intervals under the same assumptions that make the corresponding estimator for the sampling variance an unbiased or upwardly biased estimator of the true sampling variance.

Outline

Estimations: Special focus on regressions

Regressions: the horseshoe of impact evaluation

- Regression analysis is a set of statistical methods for estimating the relationships between a dependent variable and one or more independent variables
- You are (normally) familiar with bivariate linear regressions: drawing a straight line that fit a scatter plot by minimizing vertical distance between the points and the line: the Ordinary least square (OLS) or *Moindres carrés (ordinaires)* (MCO) in French.
- There are many other methods (Generalized least square, method of moments, maximum likelihood, non-parametric regressions, semi-parametric regressions,...)
- Regressions are **Estimation methods** that can, under certain conditions, retrieve the target parameters of our identification strategy.

Estimations: Special focus on regressions

The ordinary least square (OLS)

- The OLS regression plays a special role in econometrics and causal inference
- Gauss-Markov theorem: OLS is the Best Linear Unbiased Estimator if errors are uncorrelated with mean zero and homoscedastic with finite variance (see Cunningham (2018, Sections 2.10 to 2.24))
- Multivariate regressions allow to "control for" other characteristics
- Well implemented regressions can sometime exactly estimate the target parameter defined in the identification strategy
- OLS Regressions or variations are used for almost all identification strategies we'll see in this class

Estimating models are super easy with modern softwares. **Understand what's under the hood** when you run a command !

We cover intuition here but it's probably not enough, read the textbooks !

The ordinary least square (OLS)

The conditional expectation function

- The CEF for a dependent variable, Y_i given a $k \times 1$ vector of covariates, $\boldsymbol{X_i}$ (with elements X_{ik}) is the expectation, or population average of Y_i with $\boldsymbol{X_i}$ held fixed.
- For a specific value of X_i, say $\mathrm{X}_i = x$, we write $\mathbb{E}[\mathrm{Y}_i \mid \mathrm{X}_i = x]$.
- For continuous Y_i with conditional density $f_y(\cdot \mid \mathrm{X}_i = x)$, the CEF is

$$\mathbb{E}[\mathrm{Y}_i \mid \mathrm{X}_i = x] = \int t \, f_y(t \mid \mathrm{X}_i = x) \, dt$$

- If Y_i is discrete, $\mathbb{E}[\mathrm{Y}_i \mid \mathrm{X}_i = x]$ equals the sum $\sum_t t f_y(t \mid \mathrm{X}_i = x)$.

The ordinary least square (OLS)

The conditional expectation function

- Another important property is the Law of Iterated Expectations (LIE):

$$\mathbb{E}[Y_i] = \mathbb{E}\Big[\mathbb{E}[Y_i|X]\Big]$$

- Which brings us to this decomposition theorem:

$$Y_i = \mathbb{E}[Y_i|\boldsymbol{X_i}] + \varepsilon_i \tag{3}$$

Where ε_i is en error term that's mean independent of $\boldsymbol{X_i}$ and thus uncorrelated with any function $\boldsymbol{X_i}$

- Considere the following population linear equation:

$$Y_i = \alpha + \boldsymbol{X_i'}\boldsymbol{\beta} + \varepsilon_i \tag{4}$$

- If the conditional expectation function is linear, then the OLS estimator it is !
- Linear models doesn't mean linear relationships. OLS can accomodate many non-linear relationships (splines, polynomials, discontinuities, categorical variables...) and variable transformations

The ordinary least square (OLS)

Let's change X in D, see the implication

- Consider a randomized experiment of a random sample where treatment D_i is randomly assigned to $1/2$ of the sample and let Y_i be an outcome of interest.
- Average treatment effect is $ATE = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$
- using the decomposition theorem:

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

- Hence estimating the regression $Y_i = \alpha + \beta D_i + \varepsilon_i$ gives $\mathbb{E}[Y_i|D_i] = \alpha + \beta D_i$
- Thus, $\mathbb{E}[Y_i|D_i = 0] = \alpha$ and $\mathbb{E}[Y_i|D_i = 1] = \alpha + \beta$ and under random assignment, random sampling and SUTVA:

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

- The OLS regressions of a randomly assigned treatment on the dependent variable give the ATE

The ordinary least square (OLS)

The Least Squares Assumptions in the Multiple Regression Model

- Consider the population regression model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- The OLS regression over a sample of size n yield unbiased estimates of the coefficients if: (Wooldridge 2012)
  1. The relationship between Y and $\boldsymbol{X}$ are linear in parameters
  2. All variables $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, are independent and identically distributed, randomly drawn from the population.
  3. u_i is a (population) error term and is independent of all regressors. Formally, it has conditional mean zero given the regressors, i.e.,

$$\mathbb{E}[u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}] = 0$$

  4. There is some sample variation in the explanatory variable (or no perfect multicollinearity).
- If these assumptions hold, the OLS estimator is unbiased. In large samples[1], $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta_K}$ are jointly normally distributed. Further, each $\hat{\beta}_k \sim \mathcal{N}\left(\beta_k, \sigma^2_{\hat{\beta}_k}\right)$.

[1]

The ordinary least square (OLS)

Homoscedasticity

- A 5^{th} hypothesis is often invoked of constant variance of the error: $\mathbb{E}[u_i^2|\boldsymbol{X}] = \sigma^2 = \mathbb{E}[u^2]$
- This assumption stipulates that our population error term, u, has the same variance given any value of the explanatory variable, x.
- In other words, the variance of the errors conditional on the explanatory variable is simply some finite, positive number. And that number is measuring the variance of the stuff other than x that influence the value of y itself.
- Under homoscedasticity, the variance of the error is estimated by $\widehat{\sigma^2} = \frac{\boldsymbol{e}\boldsymbol{e}'}{N-K}$ where $\boldsymbol{e}$ is the residual of the OLS regression over the K variables.
- Standard error are the estimate of the sampling variability of the estimator: $SE(\hat{\beta}_k) = \frac{\widehat{\sigma^2}}{\sqrt{N}}$

The ordinary least square (OLS)

Hypotheses testing and confidence intervals

- We can test the estimated value $\hat{\beta}_k$ against a null hypothesis β_{k0} and compute the T-stat:

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k - \beta_{k0}}{SE(\hat{\beta}_k)}$$

- We can compute confidence interval using asymptotics of this t-stat : $IC_{1-\alpha} = \hat{\beta}_k \pm \Phi^{-1}(\frac{1-\alpha}{2}) \times SE(\hat{\beta}_k)$
- Correct estimates of $SE(\hat{\beta}_k)$ are as important as the coefficients, and yet it is sometimes not well considered[2]
- The homoscedasticity assumption is not used to show unbiasedness of the OLS estimators
- Without homoscedasticity, OLS no longer has the minimum mean squared errors, which means that the estimated standard errors are biased. In other words, the distribution of the coefficients is probably larger than we thought.
- It's a problem for test and prediction for we may be overconfindent.

2

The ordinary least square (OLS)

Hypotheses testing and confidence intervals

- Heteroskedasticity robust standard errors assume that the $(N \times N)$ matrix $\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}\right]$ is diagonal, meaning there is no correlation between errors accross observations (White 1980).
- You may have groups of observations that belong to certain groups which may mean that there is dependence in the error within groups.
- We can correct for these "clustering" effect using Moulton (1986) adjustment if you assume homoscedasticity or Liang and Zeger (1986) for heteroskedasticity-cluster robust standard error.
- Using R, we obtain these correction using $e.g.$ the sandwich package or estimatr easy command

```
library(estimatr)
lm_robust(dep ~ cov1 + cov2, data = mydatabase, clusters = "myclustervar")
```

More on that next week

Estimations: Special focus on regressions

The Frisch Waugh Lovell (FWL) theorem

- Consider a dependent variable $\boldsymbol{Y}$ and two sets of regressors $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ and the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}$$

- Frisch and Waugh (1933) then Lovell (2010) prove the following results (Greene 2012, p.73):

Theorem

In the linear least squares regression of vector $\boldsymbol{Y}$ on two sets of variables, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, the subvector $\boldsymbol{\beta_2}$ is the set of coefficients obtained when the residuals from a regression of $\boldsymbol{Y}$ on $\boldsymbol{X_1}$ alone and regressed on the set of residuals obtained when each column of $\boldsymbol{X_2}$ is regressed on $\boldsymbol{X_1}$.

- This theorem is fundamental to modern econometrics

The Frisch Waugh Lovell (FWL) theorem

Implications

- The Frisch-Waugh-Lowell theorem has several important implications for econometric practice, including:
  1. It provides a theoretical justification for the use of OLS in multiple linear regression analysis.
  2. It shows that the OLS estimators are consistent, asymptotically normal, and unbiased under certain assumptions.
  3. It demonstrates the importance of proper specification of the regression model, including the choice of independent variables and functional form.
- The Frisch-Waugh-Lowell theorem is a fundamental result in econometrics that provides a basis for the use of OLS in multiple linear regression analysis. Its implications for econometric practice underscore the importance of careful model specification and the need for rigorous testing of the underlying assumptions.

The Frisch Waugh Lovell (FWL) theorem

Estimating treatment effect using OLS in randomized experiments

- Consider the triplet $(\boldsymbol{Y_i}, D_i, \boldsymbol{X_i})$ of observables from a randomized control trial where Y_i are outcomes, D_i is the treatment, $\boldsymbol{X_i}$ a set of covariates, and let ε_i denote a mean-zero random error component.
- There are several regressions you can run to estimate treatment effects.
  1. **Pooled regression adjustment** (Athey and Imbens 2017b): Estimate the equation $Y_i = \alpha + \beta D_i + \boldsymbol{X'}\gamma + \varepsilon_i$. Adding the variables $\boldsymbol{X_i}$ to the simple regression does not change the probability limit provided $\boldsymbol{D_i}$ and $\boldsymbol{X_i}$ are uncorrelated, which follows under random assignment.
  2. **Saturated regression** (Angrist and Pischke 2008):A saturated regression model is one in which there is a parameter for each unique combination of the covariates. In this case, the regression fits the CEF perfectly (whatever the distribution of Y) because the CEF is a linear function of the dummy categories. We estimate $Y_i = \sum_x D_i \times X_i \alpha_x + \beta D_i + \varepsilon_i$
  3. **Linear projection** (Lin 2013): Estimate two separate regressions for treated and controls: $\hat{\mu_1}(\dot{\boldsymbol{X}}) = \hat{\alpha_1} + \boldsymbol{X_i'}\hat{\gamma_1}$ and $\hat{\mu_0}(\dot{X}_i) = \hat{\alpha_0} + \dot{\boldsymbol{X}}_{\boldsymbol{i}}\hat{\gamma_0}$, then the treatment effect is $\hat{\beta} = \hat{\mu_1}(\dot{\boldsymbol{X}}) - \hat{\mu_0}(\dot{\boldsymbol{X}})$ where $\dot{\boldsymbol{X}} = \boldsymbol{X} - \bar{\boldsymbol{X}}$
  4. **Full regression adjustment** (Negi and Wooldridge 2021): Estimate $Y_i = \alpha + \beta D_i + \dot{\boldsymbol{X}}_{\boldsymbol{i}}\gamma + D_i \times \dot{\boldsymbol{X}}_{\boldsymbol{i}}\delta + \varepsilon_i$ The demeaning of the covariates ensures that the coefficient on D is the treatment effect. This regression is also convenient for obtaining (cluster) heteroscedasticity-robust standard errors.

The Frisch Waugh Lovell (FWL) theorem

Estimating treatment effect using OLS in randomized experiments

- Regressions usually do not estimate ATE or ATT but **variance-weighted** treatment effects.
- Under the assumption of saturated in the covariates, the coefficient on the treatment in a linear regression is a weighted average of the within-stratum effects.
- Why? OLS is a minimum-variance estimator. Thus, it gives more weight to strata with lower expected variance in their estimates. That is, it gives higher weight to more precise within-strata estimates.
- Lessons from Negi and Wooldridge (2021)
  - OLS estimation using a random sample always consistently estimates the parameters in a population linear projection (subject to the mild finite second moment assumptions and the non-singularity of X). This is true regardless of the nature of Y(D) or X.
  - Estimating separate regressions for the control and treated groups is guaranteed to do no worse than both the simple difference-in-means estimator and just including the covariates in additive fashion.
  - Usually, the estimator that includes a full set of interactions strictly improves asymptotic efficiency.

Outline

Wrap-up

From Randomisation to regressions and policy evaluation

- The formal analysis of randomized control trial derived from the work of Fischer, Neyman and Rubin and have been widely used across the world to estimate the effect of various policies and treatment.
- Well-conducted RCT remove selection bias and retrieve the ATE on the population of interest.
- The conditional independence assumption supplement or mimic RCTs to estimate causal parameters conditional on a set of observables
- Regressions can be used to estimate causal parameters and the Frisch-Waugh-Lovell theorem provides a basis for the use of OLS in multiple linear regression analysis for conditional independence

Next week: More advanced stuff on RCT

- **To read: mandatory**:S. Athey and G. W. Imbens. 2017a. "Chapter 3 - The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments,* edited by Abhijit Vinayak Banerjee and Esther Duflo, 1:73–140. Handbook of Field Experiments. North-Holland, January 1, 2017
- **To read: mandatory**: Bruno Crépon et al. 2013b. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment *." *The Quarterly Journal of Economics* 128, no. 2 (May 1, 2013): 531–580

Bibliography I

- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–296.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.
- Athey, S., and G. W. Imbens. 2017a. "Chapter 3 - The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments,* edited by Abhijit Vinayak Banerjee and Esther Duflo, 1:73–140. Handbook of Field Experiments. North-Holland, January 1, 2017.
- Athey, S., and G.W. Imbens. 2017b. "Chapter 3 - the Econometrics of Randomized Experiments." In *Handbook of Field Experiments,* edited by Abhijit Vinayak Banerjee and Esther Duflo, 1:73–140. Handbook of Economic Field Experiments. North-Holland.
- Behaghel, Luc, Bruno Crépon, and Marc Gurgand. 2014. "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment." *American Economic Journal: Applied Economics* 6, no. 4 (October 1, 2014): 142–174.
- Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352, no. 6282 (April 8, 2016): 220–224.
- Christian, Cornelius, and Christopher Roth. 2016. "Can Cash Transfers Prevent Suicides? Experimental Evidence from Indonesia." *SSRN Electronic Journal.*
- Connolly, Paul, Ciara Keenan, and Karolina Urbanska. 2018. "The Trials of Evidence-Based Practice in Education: A Systematic Review of Randomised Controlled Trials in Education Research 1980–2016." *Educational Research* 60, no. 3 (July 3, 2018): 276–291.

Bibliography II

- Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013b. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment *." *The Quarterly Journal of Economics* 128, no. 2 (May 1, 2013): 531–580.
- ———. 2013a. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics* 128 (2): 531–580.
- Cunningham, Scott. 2018. *Causal Inference: The Mixtape.*
- Doyle, Orla. 2020. "The First 2,000 Days and Child Skills." *Journal of Political Economy* 128, no. 6 (June): 2067–2122.
- Fischer, Ronald A. 1935. *The Design of Experiments.* HAFNER PRESS. London: COLLIER MACMILLAN PUBLISERS.
- Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1, no. 4 (October): 387.
- Gennetian, Lisa, Greg Duncan, Nathan Fox, Katherine Magnuson, Sarah Halpern-Meekin, Kimberly Noble, and Hirokazu Yoshikawa. 2022. *Unconditional Cash and Family Investments in Infants: Evidence from a Large-Scale Cash Transfer Experiment in the U.S.* w30379. Cambridge, MA: National Bureau of Economic Research, August.
- Greene, William H. 2012. *Econometric Analysis.* 7th Edition. PEARSON.

Bibliography III

- Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, et al. 2021. "Empathy-Based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment." *Proceedings of the National Academy of Sciences* 118, no. 50 (December 14, 2021): e2116310118.
- Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103, no. 6 (October): 2052–2086.
- Huntington-Klein, Nick. 2021. *The Effect: An Introduction to Research Design and Causality | The Effect.* December 21, 2021.
- Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.
- Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *The Quarterly Journal of Economics* 131, no. 4 (November): 1795–1848.
- Kloek, Tuenis. 1981. "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated." *Econometrica* 49 (1): 205–207.
- Levitt, Steven D. 2021. "Heads or Tails: The Impact of a Coin Toss on Major Life Decisions and Subsequent Happiness." *The Review of Economic Studies* 88, no. 1 (January 1, 2021): 378–405.
- Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

Bibliography IV

- Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7, no. 1 (March 1, 2013).
- Lovell, Michael C. 2010. "A Simple Proof of the FWL Theorem." *The Journal of Economic Education* (August 7, 2010).
- Mani, A., S. Mullainathan, E. Shafir, and J. Zhao. 2013. "Poverty Impedes Cognitive Function." *Science* 341, no. 6149 (August 30, 2013): 976–980.
- Moulton, Brent. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32 (3): 385–397.
- Negi, Akanksha, and Jeffrey M. Wooldridge. 2021. "Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects." *Econometric Reviews* 40, no. 5 (May 28, 2021): 504–534.
- Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55, no. 3 (May): 703.
- Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625.
- Resnik, David. 2018. *The Ethics of Research with Human Subjects.* Vol. 74. International Library of Ethics, Law and the New Medicine. Springer, January 1, 2018.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.

Bibliography V

- White, Albert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Discret Test for Heteroskedasticiy." *Econometrica* 48 (4): 817–838.
- Wooldridge, Jeffrey M. 2012. "Introductory Econometrics: A Modern Approach," 910.

Outline

Outline

Derivating the Neyman variance forumula for the ATE

Some reminders

- The variance of a random variable X is defined as:

$$\mathbb{V}[X] = \mathbb{E}\Big[X - \mathbb{E}[X]\Big]^2$$

- T he variance is used to quantify the **amount of variation or dispersion** of a set of data values Literally speaking, the variance is the average of the squared difference between all the possible values of the random variable and its expected value In a way, it can be viewed as "the square of the distance to the mean" The standard deviation is the square root of the variance It is expressed in the same unit as the mean
- The sample analogue of the variance is:

$$\hat{Var}(X) \equiv S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y_i})^2$$

Derivating the Neyman variance forumula for the ATE

Estimated variance of outcomes

- The estimated variance of the outcome $Y(1)$,in the treatment group is (by definition):

$$\hat{Var}\left(Y(1)\right) \equiv S_1^2 = \frac{1}{N_1}\sum_{i=1}^{N_1}(Y_i(1) - \bar{Y_i}(1))^2$$

- The estimated variance of the mean outcome $\bar{Y_i}(1)$ in the treatment group is:

$$\begin{aligned}
\hat{Var}\left(\bar{Y_i}(1)\right) &= \hat{Var}\left(\tfrac{1}{N_1}\textstyle\sum_{i=1}^{N_1} Y_i(1)\right) \\
&= \tfrac{1}{N_1^2}\hat{Var}\left(\textstyle\sum_{i=1}^{N_1} Y_i(1)\right) \text{ because } \mathbb{V}[aX] = a^2\mathbb{V}[X] \\
&= \tfrac{1}{N_1^2}\textstyle\sum_{i=1}^{N_1}\hat{Var}\left(Y_i(1)\right) \text{ if observations are } i.i.d. \\
&= \tfrac{S_1^2}{N_1^2}
\end{aligned}$$

- Same computation for $Y(0)$

Derivating the Neyman variance forumula for the ATE

Estimated variance of outcomes

- The variance of a linear combination of two random variables X and Y is:

$$\mathbb{V}[aX + bY] = a^2\mathbb{V}[(]X) + b^2\mathbb{V}[Y] + 2ab\ Cov(X, Y)$$

  where $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- When X and Y are independent, $Cov(X, Y) = 0$ Since $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
- Therefore,

$$\hat{Var}\left(\bar{Y}(1) - \bar{Y}(0)\right) = \hat{Var}\left(\bar{Y}(1)\right) + \hat{Var}\left(\bar{Y}(0)\right) + 2\hat{Cov}\left(\bar{Y}(1), \bar{Y}(0)\right)$$

- In a fixed population framework, with the number of treated units fixed at N_1, the two events – unit i being treated and unit i being treated – are not independent.
- If we consider the population ATE and our sample as random i.i.d. from the population, $\hat{Cov}\left(\bar{Y}(1), \bar{Y}(0)\right) = 0$.
- considering the N observed units as a simple random sample from an infinite super-population, the Neyman estimator is an unbiased estimate of the sampling variance of the estimator of the super-population average treatment effect (See formal proof in Imbens and Rubin (2015)[Chapter 6, Appendix B].

Back to Variance

Outline

Derivation of the Least Square estimator and its variance

Finding β

Consider the general equation:

$$Y_i = \boldsymbol{X_i'}\beta + \mu_i \tag{5}$$

where Y is an $n \times 1$ outcome vector, $\boldsymbol{X}$ is an $n \times p$ matrix of covariates, β is an $p \times 1$ vector of coefficients, and μ is an $n \times 1$ vector of errors. Our purpose was then to estimate the theoretical value of β and thus, find the solution to the problem:

$$\beta = \arg\min_b \mathbb{E}[(Y_i - \boldsymbol{X_i'}b)^2]$$

Using the first order condition : $\mathbb{E}[\boldsymbol{X_i}(Y_i - X_i'b)] = 0$ the solution for b can be written :

$$\mathbb{E}[\boldsymbol{X_i}\boldsymbol{X_i'}]^{-1}\mathbb{E}[\boldsymbol{X_i}Y_i] \tag{6}$$

Derivation of the Least Square estimator and its variance

Finding β

In the simplest case where $\boldsymbol{X}$ only contains one variable x and the constant, the estimation of the coefficient is $\beta_x = \frac{cov(Y_i, x_i)}{\mathbb{V}[x_i]}$ and the intercept is $\alpha = \mathbb{E}[(]Y_i) - \beta_x \mathbb{E}[X_i]$.
Because matrix notation is not always easy, it can be useful to write that the coefficient of the k^{th} regressor is equal to:

$$\beta_k = \frac{cov(Y_i, \tilde{x}_{pi})}{\mathbb{V}[\tilde{x}_{pi}]}$$

Where $\tilde{x}_{pi}$ is the residual from a regression of x_{ki} on all the other covariates. This formula reveals that each coefficient in a multivariate regression is the bivariate slope coefficient of the corresponding regressor after "partialling out" all the other variables (Angrist and Pischke 2008) and can be interpreted *ceteris paribus*.

Derivation of the Least Square estimator and its variance

Finding β

In practice we generally do not have access to population data and therefore draw statistical inference using samples.
The sample equivalent of this quantity is given by :

$$\hat{\beta} = \sum_{i=1}^{N} \boldsymbol{X_i} \boldsymbol{X_i'}^{-1} \sum_{i=1}^{N} \boldsymbol{X_i} Y_i = [\boldsymbol{X'X}]^{-1}[\boldsymbol{X'Y}]$$

We can derive an explicit function that represents the variance of our estimates, $\mathbb{V}[\beta|\boldsymbol{X}]$, given that $\boldsymbol{X}$ is fixed.
What we are interested in is $\mathbb{V}[\beta|\boldsymbol{X}]$, which is the variance of all of our estimated coefficients $\hat{\beta}$ and the covariance between our coefficients. We can represent this as:

$$\mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \begin{pmatrix} \mathbb{V}[\hat{\beta}_0|\boldsymbol{X}] & cov(\hat{\beta}_0, \hat{\beta}_1|\boldsymbol{X}) & \cdots & cov(\hat{\beta}_0, \hat{\beta}_P|\boldsymbol{X}) \\ cov(\hat{\beta}_1, \hat{\beta}_0|\boldsymbol{X}) & \mathbb{V}[\hat{\beta}_1|\boldsymbol{X}] & \cdots & cov(\hat{\beta}_1, \hat{\beta}_P|\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_P, \hat{\beta}_0|\boldsymbol{X}) & cov(\hat{\beta}_P, \hat{\beta}_1|\boldsymbol{X}) & \cdots & \mathbb{V}[\hat{\beta}_P|\boldsymbol{X}] \end{pmatrix}$$

Derivation of the Least Square estimator and its variance

Finding β

Our goal is to estimate this matrix as it contains some interesting elements especially the first diagonal which contains the variance of each estimated coefficient which will allow us to compute standards errors :

$$SE(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \begin{pmatrix} \sqrt{\mathbb{V}[\hat{\beta}_0|\boldsymbol{X}]} \\ \sqrt{\mathbb{V}[\hat{\beta}_1|\boldsymbol{X}]} \\ \vdots \\ \sqrt{\mathbb{V}[\hat{\beta}_P|\boldsymbol{X}]} \end{pmatrix}$$

To get there, we start by rewritting the OLS estimand of the β matrix :

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\boldsymbol{X'X}]^{-1}[\boldsymbol{X'Y}] \\ &= [\boldsymbol{X'X}]^{-1}[\boldsymbol{X'}(\boldsymbol{X\beta} + \boldsymbol{\mu})] \\ &= \underbrace{[\boldsymbol{X'X}]^{-1}[\boldsymbol{X'X\beta}]}_{=I\boldsymbol{\beta}} + [\boldsymbol{X'X}]^{-1}[\boldsymbol{X'\mu}] \end{aligned}$$

Derivation of the Least Square estimator and its variance

Finding β

While we have $\boldsymbol{X}$, we do not have $\mathbb{E}\left[\boldsymbol{X}'\boldsymbol{\mu}\boldsymbol{\mu}'\boldsymbol{X}|\boldsymbol{X}\right]$, which is the variance-covariance matrix of the errors. This matrix represents all of the unobserved errors correlate with each other and their variance.
Moreover, so far we didn't use the properties of the OLS estimated regarding the correlation of the residual with regressors etc. This matrix here is not identified for it has $N \times N$ unknown parameters that define the variance of each error and the correlation of errors. In theory, we could know the correlation between the error across observations, known as serial correlation, or whether variance of the errors is constant across observations, known as homoscedasticity.

Finding β

The group structure problem

- Heteroskedasticity robust standard errors assume that the $(N \times N)$ matrix $\mathbb{E}\left[\varepsilon\varepsilon' | \boldsymbol{X}\right]$ is diagonal, meaning there is no correlation between errors accross observations. ▸ Memo
- This assumption is false in many settings among which:
  - Non-stationary time series or panel data
  - Identical values of one or more regressors for groups of individuals $=$ clusters
  - ...
- From a setting where potentially all errors are correlated together, we cannot use the estimated residuals as in the robust SE (White 1980) (because $\sum \hat{X}_i \hat{\epsilon}_i = 0$ by construction)
- Hence, one has to allow correlation up to a certain point: in time (Newey and West 1987), or among members of a group (Kloek

Finding β

The group structure problem

- Assuming homoscedasticity:

$$\mathbb{E}\Big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}|\boldsymbol{X}\Big] \equiv \Omega_{ij} = \begin{cases} 0 & \text{if } C_i \neq C_j \\ \rho\sigma^2 & \text{if } C_i = C_j,\ i \neq j \\ \sigma^2 & \text{if } i = j \end{cases}$$

- Suppose just 2 groups, this matrix looks something like:

$$\Omega_{ij} = \begin{pmatrix} \sigma^2_{(1,1)1} & \cdots & \rho\sigma^2_{(1,n_1)1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2_{(n_1,1)1} & \cdots & \sigma^2_{(n_1,n_1)1} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma^2_{(n_1+1,n_1+1)2} & \cdots & \rho\sigma^2_{(n_1+1,N)2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho\sigma^2_{(N,1)2} & \cdots & \sigma^2_{(N,N)2} \end{pmatrix}$$

Finding β

Assuming homoscedasticity & group size

- Assuming homoscedasticity & same group size:

$$\mathbb{V}_{kloek}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \mathbb{V}_{OLS} \times \left(1 + \rho_{\varepsilon}\rho_{X}\frac{N}{C}\right) \tag{8}$$

- Where ρ_{ε} is the within cluster correlation of the errors
- Where ρ_{X} is the within cluster correlation of the regressors

Relaxing homoscedasticity

- The cluster adjustment by Liang and Zeger 1986 used in most statistical packages:

$$\mathbb{V}_{LZ}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{c=1}^{C}\boldsymbol{X}_{\boldsymbol{c}}'\boldsymbol{\Omega}_{\boldsymbol{c}}\boldsymbol{X}_{\boldsymbol{c}}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{9}$$

Finding β

Estimated versions

- The estimated version of the so called robust (EHW) variance is:

$$\hat{\mathbb{V}}_{EWH}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\Big(\sum_{i=1}^{N}(\boldsymbol{Y_i} - \hat{\boldsymbol{\beta}}'\boldsymbol{X_i})^2\boldsymbol{X_i}\boldsymbol{X_i'}\Big)(\boldsymbol{X'X})^{-1} \tag{10}$$

- The estimated version of the cluster robust (LZ) variance is:

$$\hat{\mathbb{V}}_{LZ}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\Big(\sum_{c=1}^{C}\big(\sum_{i:C_i=c}\underbrace{(\boldsymbol{Y_i} - \hat{\boldsymbol{\beta}}'\boldsymbol{X_i})\boldsymbol{X_i}'}_{\hat{\varepsilon}X_i}\big)$$
$$\big(\sum_{i:C_i=c}\underbrace{(\boldsymbol{Y_i} - \hat{\boldsymbol{\beta}}'\boldsymbol{X_i})\boldsymbol{X_i}}_{\hat{\varepsilon}X_i}\big)'\Big)(\boldsymbol{X'X})^{-1} \tag{11}$$

- These are the main estimators used by applied researchers between which one has to choose.