Session III
More advanced analyses of randomised experiments

Evaluating public policies

Arthur Heim (PSE & Cnaf)

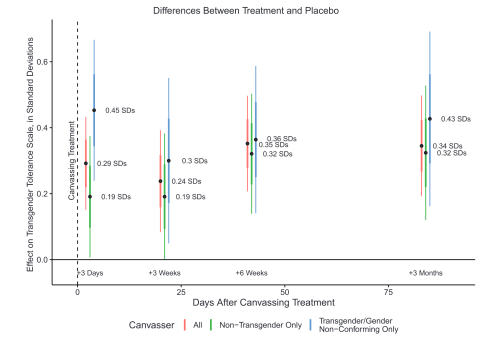


Fig. 1. Complier average causal effects on transgender tolerance scale. The 95% confidence intervals surround point estimates; the thicker lines represent one standard error. Both transgender and nontransgender canvassers produced large and lasting increases in tolerance.

February 13, 2023*

Outline

Introduction

Forwords

* Source: Broockman and Kalla (2016) ; effect of a door-to-door experiment to reduce anti-LGBTQIA+ sentiments

What we have seen so far

- Roadmap of evaluation :
  1. Identification strategy
  2. Estimation strategy
  3. Inference
- Defined the Rubin 1974 causal model and randomisation as an identification strategy
- Randomisation remove selection bias, allowing identification of causal effects.
- 2nd task after identification is estimation.
- With RCT, estimation of average treatment effect can be obtained by simple a difference in mean.
- Regressions can also be used to estimate the treatment effect of interests under some conditions.

Introduction

Today's agenda

- The conditional independence assumption
- Paper I: The STAR experiment
- Last step: inference
- Paper II: Displacement effect of job-search assistance

Outline

From randomisation to regression

Reminder

The conditional expectation function

- The Law of Iterated Expectations (LIE):

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

- Brings us to this decomposition theorem:

$$Y_i = \mathbb{E}[Y_i|\boldsymbol{X_i}] + \varepsilon_i \tag{1}$$

  Where ε_i is en error term that's mean independent of $\boldsymbol{X_i}$ and thus uncorrelated with any function $\boldsymbol{X_i}$
- Consider the following population linear equation:

$$Y_i = \alpha + \boldsymbol{X_i'}\boldsymbol{\beta} + \varepsilon_i \tag{2}$$

- This equation can be estimated by Least Square with solution

$$\hat{\beta} = [\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'Y$$

- ♭ However, Regression methods were not originally developed for analyzing data from randomized experiments,
- The attempts to fit the appropriate analyses into the regression framework requires some subtleties.

From randomisation to regression

Reminder

Wisdom from Athey and Imbens 2017

> *« In particular there is a disconnect between the way the conventional assumptions in regression analyses are formulated and the implications of randomization. As a result it is easy for the researcher using regression methods to go beyond analyses that are justified by randomization, and end up with analyses that rely on a difficult-to-assess mix of randomization assumptions, modelling assumptions, and large sample approximations. »*

Let's make the connexion clear.

Conditional independance assumption and regressions

From randomisation to regression

- 2 conceptual differences:
  1. In Neyman's analysis (finite population), **potential outcomes are fixed** and assignment varies
  2. In regression analysis, **realized outcomes and assignment are fixed** but different units, with different **error** (but same treatment status) are sampled
- May seem like some "geeky jargon" details but in many settings, it is very important especially when we leave the experimental ideal.
- Example: Comparing the effect of a policy whose adoption was staggered in different US States: Your PSU are US states, it's a finite population of 51 units !
- Consider a pure randomized control trial, treatment D, outcome Y, individual attributes $\mathbf{X}$.
- Let us consider our analysis sample as a random sample from an infinite population.
- This allows us to think of all variables as random variables with finite moments (e.g. population averages and standard deviation).
- In particular, define $\beta = \mathbb{E}[Y_i(1) - Y_i(0)]$ and $\alpha = \mathbb{E}[Y_i(0)]$.

Conditional independance assumption and regressions

From randomisation to regression

- Consider the population regression:

$$Y = \alpha + \beta D + \eta$$

- η is the **individual error**. In the OLS regression over our sample:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

- ε_i is the **residual**. The least squares estimator for β is based on minimizing the sum of squared residuals over α and β

$$\left(\hat{\beta}_{\text{ols}}, \hat{\alpha}_{\text{ols}}\right) = \arg\min_{\beta,\alpha} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \beta \cdot D_i\right)^2,$$

- With solutions

$$\hat{\beta}_{\text{ols}} = \frac{Cov(D_i, Y_i)}{S_N^2} = \frac{\sum_{i=1}^{N} \left(D_i - \bar{D}\right) \cdot \left(Y_i^{\text{obs}} - \bar{Y}^{\text{obs}}\right)}{\sum_{i=1}^{N} \left(D_i - \bar{D}\right)^2} = \bar{Y}_{\text{t}}^{\text{obs}} - \bar{Y}_{\text{c}}^{\text{obs}}$$

- and

$$\hat{\alpha}_{\text{ols}} = \bar{Y}^{\text{obs}} - \hat{\beta}_{\text{ols}} \cdot \bar{D}$$

Conditional independance assumption and regressions

Reminder

From randomisation to regression

- 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. η_i is a (population) error term and is independent of all regressors. Formally, it has conditional mean zero given the regressors, i.e.,

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

  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 In smaller sample, it follows a student distribution

Conditional independance assumption and regressions

From randomisation to regression

- Consider the population regression:

$$Y = \alpha + \beta D + \eta$$

- The least squares estimate of β over our sample is identical to the simple difference in means, so by the Neyman results the least squares estimator is unbiased for the average causal effect.
- The expected value of the outcome conditional on treatment is: $\mathbb{E}[Y_i \mid D_i = 1] = \alpha + \beta + \mathbb{E}[\eta_i \mid D_i = 1]$
- For the control: $\mathbb{E}[Y_i \mid D_i = 0] = \alpha + \mathbb{E}[\eta_i \mid D_i = 0]$
- The difference between treatment and control is:

$$\beta + \mathbb{E}[\eta_i \mid D_i = 1] - \mathbb{E}[\eta_i \mid D_i = 0]$$

- So β is a measure of treatment effect provided

$$\mathbb{E}[\eta_i \mid D_i = 1] = \mathbb{E}[\eta_i \mid D_i = 0]$$

- In the regression framework, randomisation is linked with the 3rd condition: Under random assignment, the average **error** for treated and control units are 0:

$$\mathbb{E}[\eta_i | D = 0] = 0 \quad \text{and} \quad \mathbb{E}[\eta_i | D_i = 1] = 0$$

Conditional independance assumption and regressions

From randomisation to regression

- Note that random assignment of the treatment does **not** imply that the **error term** η is independent of D_i.
- In fact, in general there will be heteroskedasticity, and we need to use the Eicker-Huber White robust standard errors to get valid confidence intervals.
- Mean-independence of the treatment and population error is sufficient.
- The **error term** in the population regression also has a clear interpretation. With simple notation manipulation you can show:

$$\begin{aligned} \eta_i &= Y_i(0) - \alpha + D_i\left(Y_i(1) - Y_i(0) - \beta\right) \\ &= (1 - D_i) \cdot \underbrace{\left(Y_i(0) - \mathbb{E}[Y_i(0)]\right)}_{\text{Individual deviation in } Y_i(0)} + D_i \cdot \underbrace{\left(Y_i(1) - \mathbb{E}[Y_i(1)]\right)}_{\text{Individual TE Heterogeneity}} \end{aligned}$$

Conditional independance assumption and regressions

Reminder

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

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

- 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 β_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 $\boldsymbol{X_1}$.

- In the Appendix, I give you an illustration of the FWL theorem in action.

▸ Go to illustration

Conditional independance assumption and regressions

What is conditional independence ?

- So far, we manipulated **conditional expectations** $\mathbb{E}[Y_i|D_i]$ by treatment status and used independence of treatment and potential outcomes $(Y_i(1), Y_i(0) \perp D_i)$ to define average treatment effects.
- In some settings, random assignment depend on other factors $\boldsymbol{X}$ (e.g. block randomisation) and the independence hold true **conditional** on the value of these other factors.
- In other settings, it may be plausible that an explanatory variable of interest (e.g. a treatment or policy) is independent of the outcomes conditional on some characteristics.
- This is the **conditional independence assumption**
- You have probably already read papers making causal claim saying things like "all other things being equal", "controlling for X, we find..." or "we matched observation based on the following variables"...
- Sometimes, we see the latin expression *Ceteris paribus*. That's it. That's the conditional independence assumption.

Conditional independance assumption and regressions

What is conditional independence ?

- The conditional independence assumption (CIA) is usually written in the form:

$$(Y_i(1), Y_i(0)) \perp D_i | \boldsymbol{X_i} \tag{3}$$

- *"beyond covariates X, there are no characteristics of the individual associated both with the potential outcomes and the treatment"*
- Conditional on X, assignment is "as good as random"
- The "as good as random" **hypothesis** then allows you to make causal claim. $\Rightarrow$ The CIA can also be an **identification strategy**
- We also call it selection on observables, exogeneity, ignorability or Unconfoundedness.
- ♭ Unless you actually randomly assign a variable conditional on some other, the CIA is a **non-refutable assumption** i.e. you can't formally test its validity[2].

2. Actually, there are methods to assess the plausibility of the CIA in some

What is conditional independence ?

Few remarks

- You don't need to observe all variables that determine participation/assignment nor all those that determine potential outcomes.
- You need to observe **all** variables that determine **both participation and outcomes**: The **confounders**
- The ignorability of treatment assignment says that if you can't control for confounders, **your statistical model is showing a correlation and not causation**
- if there's a variable that determines participation, but not outcomes, it's not a confounder: that's called an instrument, and you should use it as such (More on that in lecture VI).
- The CIA Is Everywhere (in empirical papers). But unless there are very good reason to believe that there are no latent factor that confound the results (such as a randomized experiment), it is a **strong assumption**

What is conditional independence ?

Careful with the CIA

- What you read in the paper:

  "In our preferred model, the set of control variables includes gender, age, country of residence during childhood, marital status, type of residence, wealth and occupation dummies."

- What it means in the data:

Figure 1: Source: Somewhere on twitter, probably `@KhoaVuUmn`, saved on my phone for this moment

- I mean... Sure they are a fairly good match but whatever the "treatment" condition we would be comparing between the two, I am pretty sure there would be some unobserved factor correlated with treatment and outcomes.

Conditional independance assumption and regressions

CIA in randomized studies

- In a pure RCT, there is **no need** for covariates.
  But, there are three main reason why we (almost) always use regressions with covariates (in our preferred specifications):
  1. **Design features** like block-randomisation are easy to account for with fixed-effects in regressions ;
  2. Covariates may make **analyses more informative** by removing variation associated with covariates, making the residual variance smaller.
  3. Covariates can help correct design issues (remaining baseline unbalance, attrition), one **need** to account for in the analysis.
- ♭ When covariates are correlated with treatment, i.e., when people with certain attributes are more or less likely to be treated, it has two consequences
  1. Omitting the variable in the model creates **omitted variable bias**
  2. If treatment effect is heterogeneous w.r.t. these same covariates, the regression of Y on D and X will generally not correspond to the ATE. (Sł oczyński 2022)
- Some variables are **bad controls** and **must not** be included in the model (e.g. Mediating variables, post exposure attributes)
- **Common support**[3]: your data should "overlap" i.e. you should have individuals from both treatment and controls over the whole covariate distribution.

3. The support of a random variable is the set of realizations that occur with

Conditional independance assumption and regressions

Omitted variable bias

- Consider for the sake of the example that we want to estimate the impact of schooling S_i on income Y_i.
- Say there exist an unobserved latent factor called "Ability" noted A_i such that the population model is of the form:

$$Y_i = \alpha + \delta S_i + \gamma A_i + \mu_i$$

- Since A_i cannot be observed, say we estimate the following regression using OLS

$$\hat{Y_i} = \hat{\alpha} + \hat{\delta} S_i + \mu_i$$

Conditional independance assumption and regressions

Omitted variable bias

- The OLS estimator is $\hat{\delta} = \frac{Cov(Y_i, S_i)}{\sigma_S^2}$

$$\Rightarrow \hat{\delta} = \frac{Cov(\alpha + \delta S_i + \gamma A_i + \mu_i, S_i)}{\sigma_S^2}$$

$$\Rightarrow \hat{\delta} = \delta + \gamma \frac{Cov(A_i, S_i)}{\sigma_S^2}$$

- If $\gamma > 0$ (The higher the ability the higher the income) and if $Cov(A_i, S_i) > 0$ (Those with higher ability stay longer in school), then the OLS estimate is biased and over-estimate the true effect of schooling because it's contaminated by omitted variable bias.
- however, if A can be accurately measured, or, if we leave our example, if we observe A and the conditional assumption holds true, adding A in the regression will retrieve the causal parameter of interest.

Conditional independance assumption and regressions

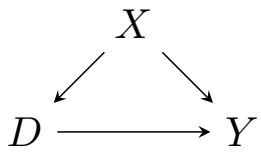
A word on 'bad controls'

- I recommend reading the chapter on **Directed Acyclical Graphs** (DAG) in Scott Cunningham. 2018. *Causal Inference: The Mixtape*
  Click, this is a Link
- Pearl (1995) develop a framework for causal inference using graph theory to represent causal relationships and define conditions for identifications and estimations.
- Very powerful to make some assumption explicit or "discover" causal relationship, but sometimes it's harder to capture other important features (e.g. heterogeneity).
- See Imbens (2020) for a discussion on the links and differences between DAGs and the potential outcome framework, and Heckman and Pinto (2022) for a general discussion on causality.
- With DAGs, If we control for X, We say we close the **backdoor path** between D and Y through X, and the causal relationship is identified.

Conditional independance assumption and regressions

A word on 'bad controls'

Figure 2: Conditionning is "closing the back door"



- A back-door path is any path from D to Y that starts with an arrow pointing **into D**.
- "Backdoor paths" creates **Fork** relationships: $D \leftarrow X \rightarrow Y$. We say X is a **confounder**.
- If one close a backdoor path (by conditioning), then the partial causal effect of D on Y is identified !
- There are two other different path configurations:
  1. **Chains**: $D \rightarrow M \rightarrow Y$: In that case, M is either a **mediator** or D is an **instrument** for M
  2. **Colliders**: $D \rightarrow C \leftarrow Y$: In that case, C is a **collider**
- A **Mediator** or mediating variable **transmit** the effect of D to Y through it. Distinction between total, direct and indirect effect.
- A **Collider** is more counter-intuitive ; In general, it's a variable caused by at least two others (arrows colliding...). In the treatment-effect framework, a collider is a variable that is both caused by the treatment and an outcome.

Conditional independance assumption and regressions

A word on 'bad controls'

A notorious collider bias: Survivorship biais

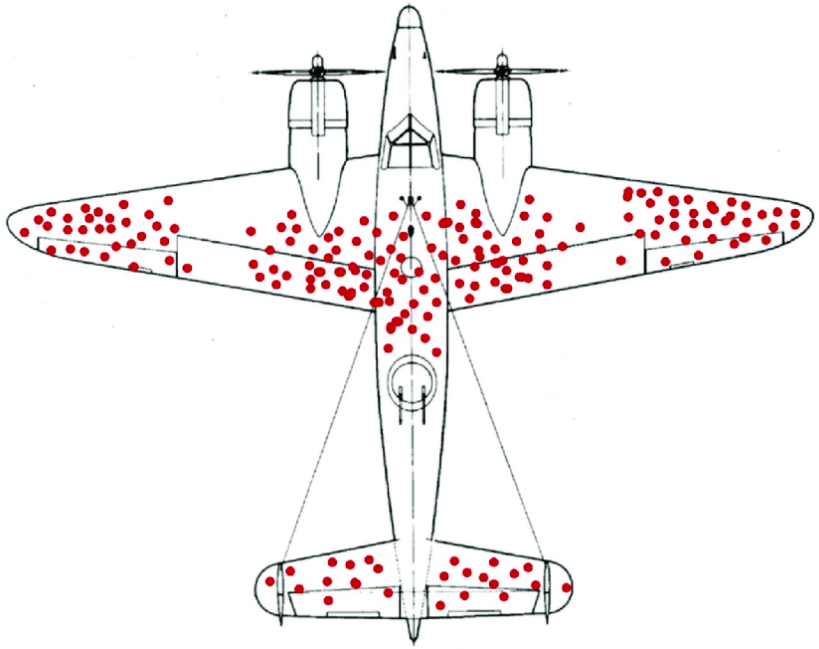


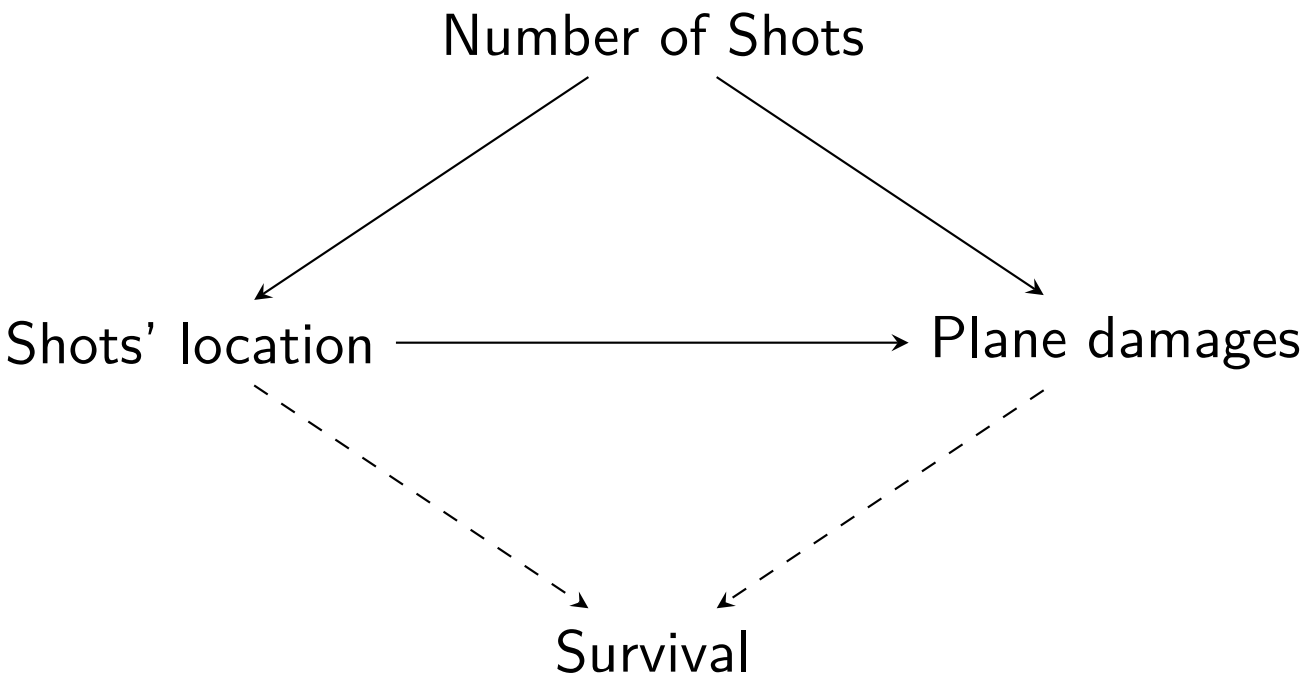
Figure 3: Wald plane as a symbol of survivorship bias

Conditional independance assumption and regressions

A word on 'bad controls'

A notorious collider bias: Survivorship biais

Figure 4: Wald plane in DAG



Conditional independance assumption and regressions

A word on 'bad controls'

How do DAG help us with "bad controls" ?

- From these definition it should be clear that with DAG, you explicitly show which relationships you are modelling and which ones you aren't.
- Clear definition of a causal relationships: identified if there is no backdoor paths left open.
- Also help you think about the role of other variables. **only confounders** should be controlled.
- **General rule:** don't control for post-exposure variables unless you are actually doing a mediation analysis.
- In Pearls word, a confounder is always a parent node, so in case a doubt, map a DAG.

Conditional independance assumption and regressions

A word on 'bad controls'

Collider bias: a high stake example (See the Discussion in Cunningham 2018)

- In a controversial paper, Fryer (2019) Analyse racial bias in police use of force.
- To do that he access a large database of arrest reports and with careful econometric analysis, controlling for many things, he find no evidence of racial bias in these data.
- These conclusions have been heavily criticized for various reasons. One critics from "Mr Selection model" (Pr. J.J. Heckman) is that these police administrative datasets select on officers' **post-treatment decisions to detain civilians**. Decisions that are potentially **also discriminatory**, thus **omitting all data on encounters not resulting in detainments** and potentially severely understating the extent of racial bias in policing (Knox, Lowe, and Mummolo 2020).
- In other words, the observation itself is conditioned on having been arrested, which is affected by the variable of interest (race) and outcome (if they intend to use force on you they are more likely to arrest you).
- It's fairly easy to notice that using a DAG.

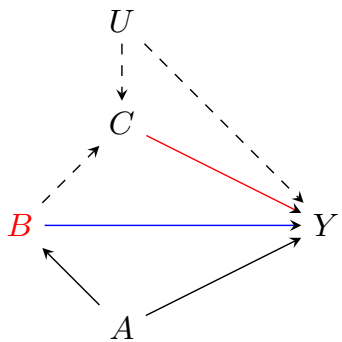
Conditional independance assumption and regressions

A word on 'bad controls'

Collider bias: a high stake example (See the Discussion in Cunningham 2018)

- Denote B for race, Y for use of force and the relationship between the two (in blue) is the one of interest.
- We denote Police controls and arrests C.
- Fryer observe individuals attributes A that affect both the likelihood that a person is from a minority and is brutalized by the police (externalizing behavior, neighborhood, past record),...
- There are unobserved factors U that we call police suspicion (can include many things) that cause both arrests and use of violence.
- From this graph, It's clear that the relationship of interest is not identified, because observations are conditioned on C although B $\rightarrow$ C is not observable.

Figure 5: C is a collider bias



Heterogenous treatment effect

Weird things happening under the hood

- Consider a RCT with block randomization and let P_j denote treatment probabilities across discrete blocks $\mathbf{X}$ (e.g. the share of treated may be higher/lower in some groups).
- Consider the OLS regression with block fixed effects:

$$Y_i = \sum_j \delta_j \mathbb{1}(X_{ij} = 1) + \beta_{OLS} D_i + \varepsilon_i$$

- this regression is **saturated in the covariates**, which means that it is linear in the covariates by construction. It is **not fully saturated** because it doesn't include interactions between treatment and blocks. When is β_{OLS} equal to the ATE or the ATT?

Heterogenous treatment effect

Weird things happening under the hood

1. β_{OLS} equals the ATE if **treatment effect is constant** for everyone or if **conditional treatment probabilities are constant** and equal to $p_j = p_{j'} = .5 \; \forall j, j'$ (Angrist and Pischke 2008, Section 3.3.1). Why ?
   - OLS is a **minimum variance estimator**. Thus, it gives more weight to strata with lower expected variance in their estimates i.e higher weight to more precise within-strata estimates.
   - When are these estimates going to be more precise? When the treatment and control group are roughly the same size and so the variance is maximized.
   - Because $\mathbb{V}[D|\boldsymbol{X} = \boldsymbol{x}] = Pr(D|\boldsymbol{X} = \boldsymbol{x}) \cdot (1 - Pr(D|\boldsymbol{X} = \boldsymbol{x}))$, which is highest when $Pr(D|\boldsymbol{X} = \boldsymbol{x}) = .5$
   - When treatment probabilities vary across blocks/covariates, β_{OLS} produces a **treatment-variance weighted average** of block-specific treatment effects. ▸ Proof
   - The problem is that these weights are **reversed**. If (almost) all units are treated in a block, the regression gives it a (close to) 0 weight as its variance is almost null.

Heterogenous treatment effect

Weird things happening under the hood

2. "Paradox": $\beta_{OLS} = ATT$ when **very few units are treated** and $\beta_{OLS} = ATU$ when **most units are treated**.
   - The very recent (and super clear) paper by Sł oczyński 2022 explains what the problem is (you can see why with the proof in this course appendix).
   - Suppose you have two strata: X_1 is large with few treated units so $P(X_1)$ is small, and a second X_2 that's small but with a lot of treated units.
   - Intuitively, the motivation for using OLS is that the linear projection of Y ond D and X is the best predictor of Y given D and X.
   - So OLS is best at predicting **actual outcomes**.
   - But causal inference is about predicting the **missing outcomes** i.e. the counterfactual values.
   - If we wanted to predict "what is", we would put a lot of weights where we have a lot of precision (so, the big strata with few treated).
   - That's what OLS does.
   - But what we want is to estimate the counterfactal and for that we would need to put more weights where there are lot of treated units because that's where the treatment effect is more precisely estimated.

Heterogenous treatment effect

Comments and solutions

- If the probabilities do not vary much, this is of little importance. But if treatment probabilities vary a lot, the OLS results will be far from the true ATE.
- If we know the conditional treatment probabilities (also called **propensity scores** (by design), we can re-weight the observations (Imbens 2004):
  1. To estimate the ATE, weight treated observations by $w_1 = \frac{1}{(Pr(D_i|\boldsymbol{X_i}=x))}$ and controls by $w_0 = \frac{1}{(1-Pr(D_i|\boldsymbol{X_i}=x))}$;
  2. To estimate ATT, weight treated observations by $w_1 = \frac{(Pr(D_i|\boldsymbol{X_i}=x))}{(1-Pr(D_i|\boldsymbol{X_i}=x))}$ and controls get unit weights ;
  3. To estimate ATU, weight treated observations by $w_1 = \frac{(1-Pr(D_i|\boldsymbol{X_i}=x))}{(Pr(D_i|\boldsymbol{X_i}=x))}$ and controls get unit weights ;
- The idea that you can correct for non-random sampling by weighting by the reciprocal of the probability of selection dates back to Horvitz and Thompson (1952).
- When we don't know the probability, we need consistent estimators.

Conditional independance assumption and regressions

Common support

- When the conditional independence assumption holds, we actually need an extra assumption/condition to estimate the effects

$$0 < Pr(D_i = 1|\boldsymbol{X} = \boldsymbol{x}) < 1 \ \forall \ \boldsymbol{x} \tag{4}$$

- *"At all x's, there must be both treatment and control observations"*
- Implies $f(X|D = 1)$ overlaps with $f(X|D = 0)$
- If unmet, restrict sample to observations with overlap
- This is important because OLS will project over the support of X and if one group has no support on some values, then we rely on extrapolation.
- Always check that your observations are balanced over the support of the X. The distribution of the X impacts the weights OLS give to different observations.
- In practice, check for outliers, observe densities, scatter plots etc. Observation without common support should usually be dropped (Lechner and Strittmatter 2017).

Common support

Covariates with RCT: How to ?

1. **Special case: fully saturated regressions**
   - $Y_i = \sum_x \mathbb{1}(X_i = x)\delta_x + \beta D_i + \sum_x \tau_x D_i \times \mathbb{1}(X_i = x) + \varepsilon_i$
   - The regression fits the CEF perfectly (whatever the distribution of Y) because the true CEF is linear in parameters.
   - Thus, the OLS estimate of β is an unbiased estimate of the ATE (Athey and Imbens 2017).
   - The coefficients τ estimate treatment effect heterogeneity interpreted as deviation from the ATE.
2. **General case where X may be continuous**:
   - **Pooled regression**: The regression of Y, D and $\boldsymbol{X}$ yields consistent estimate of the ATE provided D and $\boldsymbol{X}$ are uncorrelated, which follows under random assignment (Negi and Wooldridge 2021).
   - It estimates the ATE if we assume constant treatment effect or a random sample of a large population (where heterogeneity is in the error term).
   - **(Lin 2013) regressions**: Estimate $Y_i = \alpha + \beta D_i + \dot{\boldsymbol{X}}_{\boldsymbol{i}}\gamma + D_i \times \dot{\boldsymbol{X}}_{\boldsymbol{i}}\tau + \varepsilon_i$
   - Where $\dot{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{X}_{\boldsymbol{i}} - \bar{\boldsymbol{X}}_{\boldsymbol{i}}$ is the deviation from the population average (in practice, use sample mean).
   - The demeaning of the covariates ensures that the coefficient on D is the treatment effect.

Common support

Advices for practice

- **Check for balance** in x over the whole distribution. Your covariates should be "well-behaved" - not much skewness or outliers.
- If in fact the covariates have very skewed distributions, the finite sample bias in the linear regression estimates may be substantial
- In case of imbalance, **drop the observations without overlap**.
- Always show your estimate without covariates (unless they reflect the design)
- in an RCT, adding covariates **shouldn't change the coefficient** much, only standard errors should be (slightly) lower. If it does, worry and investigate !
- The treatment heterogeneity parameters can be interpreted in the fully saturated regression. However, in the Lin regression, these parameters are meaningful only if the relationship between the outcome and the covariate is linear in parameter as included in the regression.
- **Converting covariates into indicator variables** is usually a good idea (e.g. quantile dummies) to get fully saturated regression with clear interpretations.

Common support

More advanced methods

- **Inverse propensity score weighting**: If we can access the conditional treatment probability (or propensity score) to construct weights, we can use weighted regression methods based on the inverse propensity score (see e.g. Imbens (2004)).
- **doubly robust methods**: We can use IPW in regressions with covariates. This inverse-probability-weighting regression adjustment (IPWRA) has the doubly robust property: we only need either the propensity score model to be true or the outcome model to be true to recover unbiased estimates. (See the recent paper by (Sł oczyński, Uysal, and Wooldridge 2022))
- **Matching**: more on that in lecture 11.
- **Machine learning**: not this year, but if you are curious : Susan Athey and Guido W Imbens. 2019. "Machine Learning Methods Economists Should Know About." *Annual Review of Economics* 11:62. Important contribution when the set of potential covariates is large (Belloni, Chernozhukov, and Hansen 2014; Chernozhukov, Hansen, and Spindler 2015; Chernozhukov et al. 2017), or to identify treatment effect heterogeneity (Imai and Ratkovic 2013; Athey and Imbens 2016; Demirer et al. 2017; Wager and Athey 2018; Athey, Tibshirani, and Wager 2019)

Outline

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

A large randomized experiment: the STAR project

- Alan Krueger (Quarterly Journal of Economics 1999) re-analyzed a randomized experiment of the effect of class size on student achievement
- The project is known as the Tennessee **Student/Teacher Achievement Ratio** (STAR) Project and was run in the 1980's.
- 11 600 students and their teachers were randomly assigned to one of three groups
  1. Small classes (13 to 17 students)
  2. Regular classes (22 to 25 students)
  3. Regular classes (22 to 25 students) with a full time teacher's aide
- After the assignment, the design called for students to remain in the same class type for four years
- Randomization occurred within schools

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

The main equation in Krueger (1999)

Figure 6: The estimation strategy from Krueger (1999)[p. 510]

We begin analyzing the STAR data by estimating the following regression equation for students in each grade level:

$$(2) \quad Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \beta_3 X_{ics} + \alpha_s + \epsilon_{ics},$$

where Y_{ics} is the average percentile score on the SAT test of student i in class c at school s, $SMALL_{cs}$ is a dummy variable indicating whether the student was assigned to a small class that year, REG/A_{cs} is a dummy variable indicating whether the student was assigned to a regular-size class with an aide that year, and X_{ics} is a vector of observed student and teacher covariates (e.g., gender). The independence between class-size assignment and other variables is only valid within schools, because randomization was done within schools. Consequently, a separate dummy variable is included for each school to absorb the school effects, α_s.

The equation is estimated by ordinary least squares (OLS). In calculating the standard errors, however, the error term ϵ_{ics} is modeled in a components-of-variance framework. Specifically, ϵ_{ics} is assumed to consist of two components: $\epsilon_{ics} = \mu_{cs} + \epsilon'_{ics}$, where μ_{cs} is a class-specific random component that is common to all members of the same class, and ϵ'_{ics} is an idiosyncratic error term.

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Comments on the estimation strategy

- Estimate together the two treatment effects
- Condition on school fixed effect to account for stratification
- Use covariates to improve precision in a pooled regression.
- Adjust standard error for clustering at the classroom level

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

The effect of small class size on IQ in Kindergarten

Figure 7: OLS estimates of the effect of class size on average percentile of Stanford Binet IQ test from Krueger (1999)[p. 512]

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
			A. Kindergarten	
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Master's degree	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Comments on the results

- First column estimate the first difference without school fixed effects $\rightarrow$ doesn't correctly account for the design
- Adding fixed effect changes the estimated treatment effect a little bit. Why ? Because it removes between-school variations and identify the within-school average difference (with weights)
- Adding covariates does not change the coefficient but improve precision as expected.
- Reducing class size from 23 to 15 students on average increases the IQ rank by 5 percentiles in first grade. i.e. someone who would have have a median IQ in a normal class size would be ranked at the 55^{th} percentile in small classroom on average.

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

The effect of small class size on IQ in frist grade

Figure 8: OLS estimates of the effect of class size on average percentile of Stanford Binet IQ test from Krueger (1999)[p. 512]

	B. First grade			
Small class	8.57	8.43	7.91	7.40
	(1.97)	(1.21)	(1.17)	(1.18)
Regular/aide class	3.44	2.22	2.23	1.78
	(2.05)	(1.00)	(0.98)	(0.98)
White/Asian (1 = yes)	—	—	6.97	6.97
			(1.18)	(1.19)
Girl (1 = yes)	—	—	3.80	3.85
			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61
			(.87)	(.87)
White teacher	—	—	—	−4.28
				(1.96)
Male teacher	—	—	—	11.82
				(3.33)
Teacher experience	—	—	—	.05
				(0.06)
Master's degree	—	—	—	.48
				(1.07)
School fixed effects	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Problem 1: Dealing with attrition

- A common problem in randomized experiments: non-response and attrition
- If attrition is **random** and affects the treatment and control groups in the same way, estimates would remain unbiased
- Here the attrition is likely to be non random especially good students from large classes may have enrolled in private schools creating a selection bias problem
- Krueger addresses this concern by imputing test scores (from their earlier test scores) for all children who leave the sample and then re estimates the model including students with imputed test scores

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Consistent estimate with imputed missing outcomes

Figure 9: Comparison of raw estimates with specifications with imputed missing outcomes Krueger (1999)[p. 512]

TABLE VI
EXPLORATION OF EFFECT OF ATTRITION DEPENDENT VARIABLE: AVERAGE PERCENTILE SCORE ON SAT

	Actual test data		Actual and imputed test data	
Grade	Coefficient on small class dum.	Sample size	Coefficient on small class dum.	Sample size
K	5.32	5900	5.32	5900
	(.76)		(.76)	
1	6.95	6632	6.30	8328
	(.74)		(.68)	
2	5.59	6282	5.64	9773
	(.76)		(.65)	
3	5.58	6339	5.49	10919
	(.79)		(.63)	

Estimates of reduced-form models are presented. Each regression includes the following explanatory variables: a dummy variable indicating initial assignment to a small class; a dummy variable indicating initial assignment to a regular/aide class, unrestricted school effects; a dummy variable for student gender; and a dummy variable for student race. The reported coefficient on small class dummy is relative to regular classes. Standard errors are in parentheses.

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Problem 2: Students changed class after random assignments

- Subjects moved between treatment and control groups
- Call this problem "imperfect compliance"
- Krueger reports reduced form results where he uses initial assignment and not current status as explanatory variable
- In Kindergarten, OLS and reduced form estimates are the same because students remained in their initial class for at least one year
- In 1st and 2nd grade, OLS (column 1 4) and reduced form (columns 5 8) are different

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Compliance to treatment assignment in 1st grade

Figure 10: Transition between class-size in adjecent grades ; number of students in each class Krueger (1999)[p. 515]

B. First grade to second grade

	Second grade			
First grade	Small	Regular	Reg/aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Reduced-form coefficient smaller than OLS estimates

Figure 11: Comparison of raw estimates with specifications with imputed missing outcomes Krueger (1999)[p. 513]

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C. Second grade								
Small class	5.93	6.33	5.83	5.79	5.31	5.52	5.27	5.26
	(1.97)	(1.29)	(1.23)	(1.23)	(1.70)	(1.16)	(1.10)	(1.10)
Regular/aide class	1.97	1.88	1.64	1.58	.47	1.44	1.16	1.18
	(2.05)	(1.10)	(1.07)	(1.06)	(1.23)	(0.87)	(0.81)	(0.81)
White/Asian (1 = yes)	—	—	6.35	6.36	—	—	6.27	6.29
			(1.20)	(1.19)			(1.21)	(1.20)
Girl (1 = yes)	—	—	3.48	3.45	—	—	3.48	3.44
			(.60)	(.60)			(.60)	(.60)
Free lunch (1 = yes)	—	—	−13.61	−13.61	—	—	−13.75	−13.77
			(.72)	(.72)			(.73)	(.73)
White teacher	—	—	—	.39	—	—	—	.43
				(1.75)				(1.76)
Male teacher	—	—	—	1.32	—	—	—	.82
				(3.96)				(4.23)
Teacher experience	—	—	—	.10	—	—	—	.10
				(.06)				(.07)
Master's degree	—	—	—	−1.06	—	—	—	−1.16
				(1.06)				(1.05)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.22	.28	.28	.01	.21	.28	.28

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

What's important

- Large experiment with high compliance rate and fairly clean design shows that reducing class size in early grades improve cognive development
- Teacher assistant don't seem to work that well thus testing the hypothesis that it's less about teacher/student ratio but maybe the learning environment ?
- This paper was very influential and opened the path of a 10-year academic debate on the effects of class size.
- At that time, the consensus was that class size has little to no effect on students' achievement (See Krueger, Hanushek, and Rice (2002))
- The puzzle with teaching assistants also nurtured a broad strand of literature (See the work of Peter Blatchford for instance)

Case study: The STAR experiment by Krueger (1999)

Does class size or teacher/student ratio inprove student achievement ?

Where is the debate now ?

- Filges, Sonne-Schmidt, and Nielsen (2018) do a systematic review and meta-analysis of the literature on class size with very strict inclusion criteria.
- They collected 127 studies analysing 55 different populations from 41 different countries. A large number of studies (45) analysed data from the Student Teacher Achievement Ratio (STAR) experiment
- Overall, the evidence suggests at best a small effect on reading achievement. There is a negative, but statistically insignificant, effect on mathematics.
- For the non-STAR studies the primary study effect sizes for reading were close to zero but the weighted average was positive and statistically significant.
- There was some inconsistency in the direction of the primary study effect sizes for mathematics and the weighted average effect was negative and statistically non-significant.
- The STAR results are more positive, but do not change the overall finding. All reported results from the studies analysing STAR data indicated a positive effect of smaller class sizes for both reading and maths, but the average effects are small.

Outline

Inference: a history of variance

What are we talking about

- **Definition**: Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability.

Table 1: Terminology for measures of precision

Notations	Names
σ	Population standard deviation
S, s	Sample standard deviation
$\frac{\sigma}{\sqrt{n}}$	Sampling standard deviation of $\overline{X}$
$\frac{S}{\sqrt{n}}, \frac{s}{\sqrt{n}}$	standard error of $\overline{X}$.

- σ describes the population and does not depend on sample size. S and s are the estimator and estimate of σ.
- $\frac{\sigma}{\sqrt{n}}$ measures the sampling variability in $\overline{X}$ and $\frac{S}{\sqrt{n}}$ is our best estimator of this variability.

Monte carlo simulations

1. Let us illustrate consistency and asymptotic properties of OLS estimands using Monte Carlo simulations. Consider the simple bivariate regression

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

2. We generate a sample of 1000 $i.i.d$ observations and a true β=0 and estimate this equation using OLS and collect the estimate of $\hat{\beta}$
3. We do that 10 000 times with a new random sample and plot the distribution of the estimated $\hat{\beta}$

This idea and pieces of the code were taken from **Yuki Yanai (2016) seminar**

Monte carlo simulations

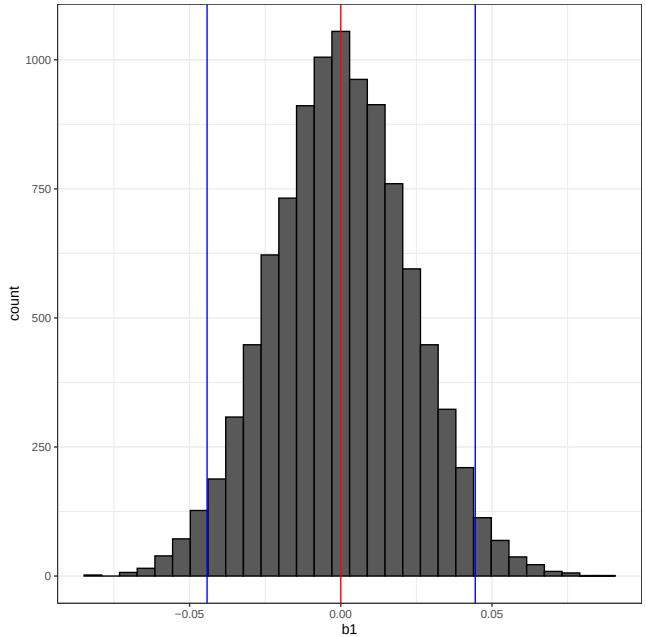


Figure 12: Distribution of the estimated β over 10 000 simulated random samples

- Our simulations verify all OLS hypotheses, hence $\mathbb{E}[\beta_{OLS}] = \beta = 0$ which is shown in this graph.
- The distribution of the 10 000 estimations of this OLS regression has a well behaved bell-curve shape with 2.5 % of the estimations below -0.044 and above 0.044

Monte carlo simulations

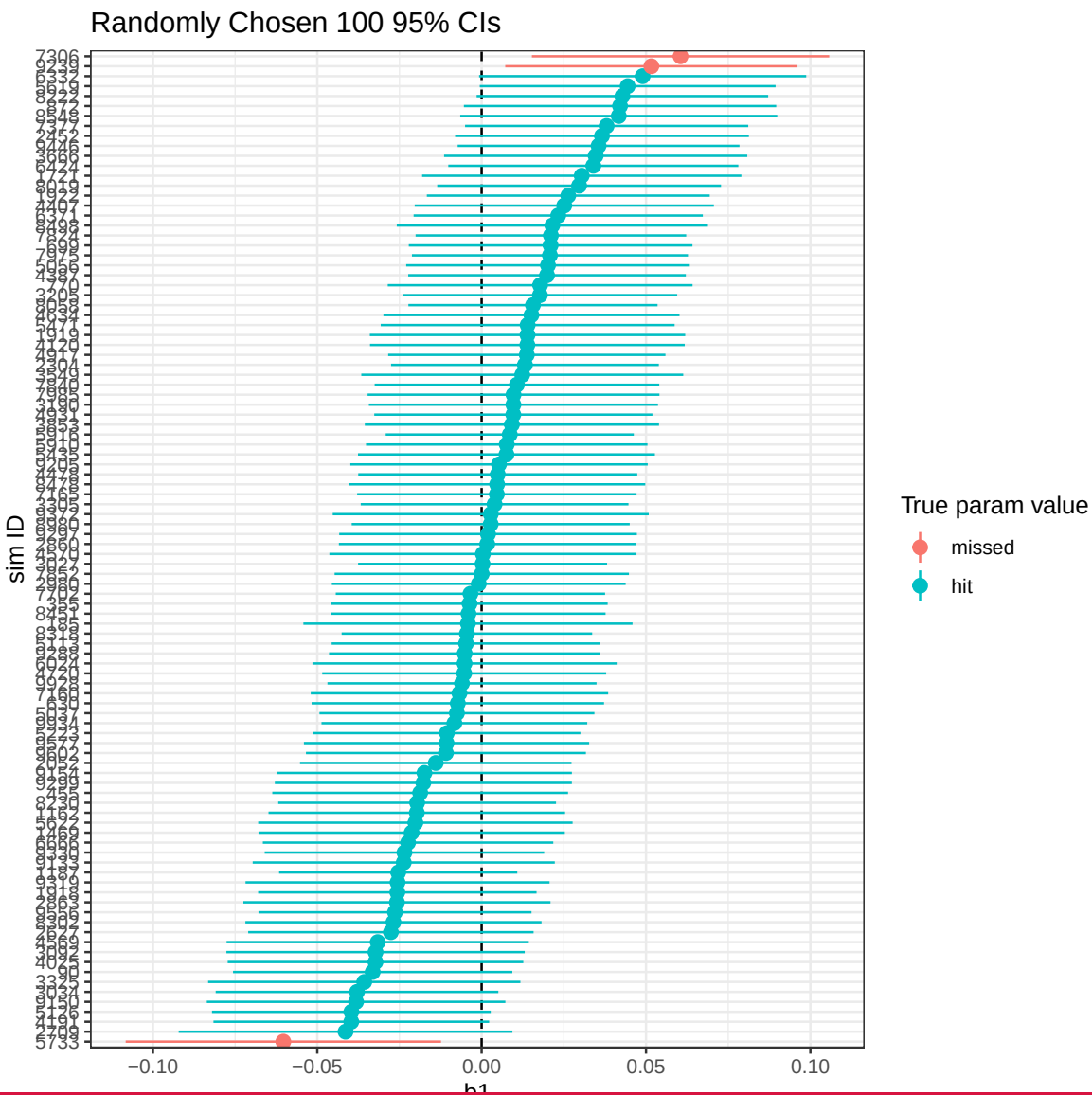
Confidence interval formula and cover range

- Without accessing 10 000 samples, we would build confidence interval with a normality assumption to use fractiles of a normal distribution (1.96) and the estimated standard error of the OLS coefficient.

$$\begin{aligned} IC_{95\ \%} &= \hat{\beta} \pm 1.96 \times S\hat{E}_{\beta} \\ &= [-0.044; 0.044] \end{aligned}$$

- With simulation, we can observe the empirical quantiles of the sampling distribution and simply took those excluding 2.5 % on each side.
- Knowing that the true effect is 0 in our simulations, if we take 100 random estimations among the 10 000 and plot the estimates and their confidence interval, we should find approximately 5 % of these estimates that do not cover 0 over their 95 % CI based on a normality assumption.
- Let's see this:

What are we talking about



The problems with clusters

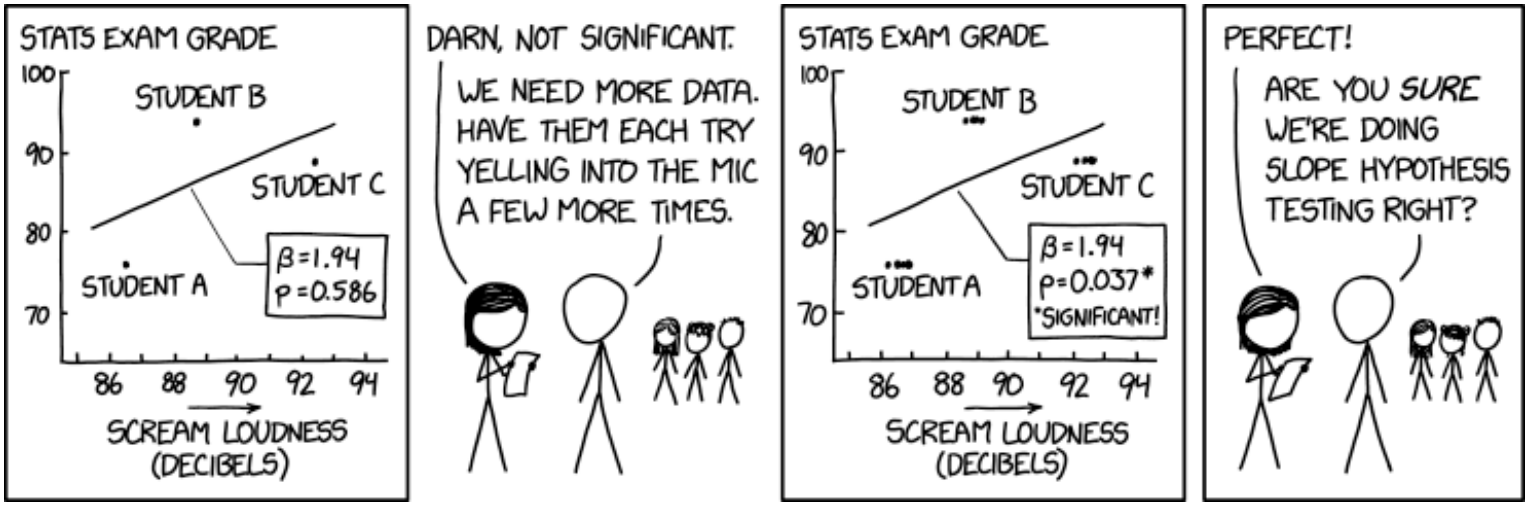


Figure 14: Comics for intuition (from phdcomics.com)

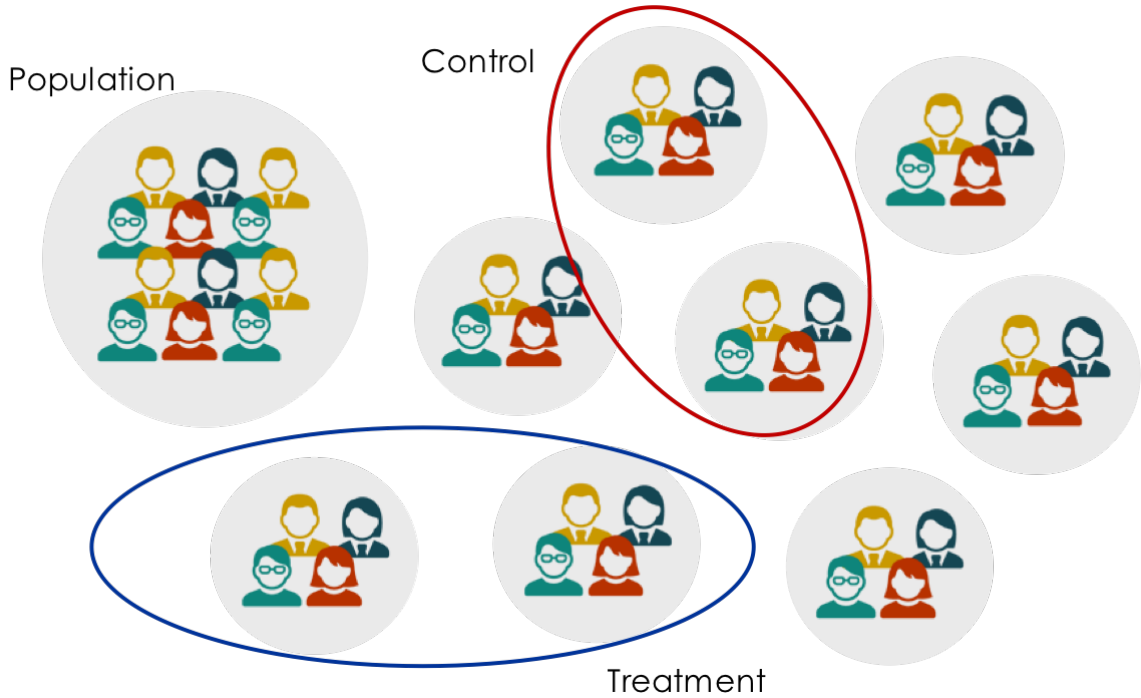
The problems with clusters

What's that "cluster" thing ?

- "Clusters" essentially mean "groups" in your data that share some common traits
- In the data it means (some) variables are correlated within these groups
  - In panel data, individuals are literally observed several times so their observations are likely correlated
  - students in a classroom have the same teacher hence their test scores may be correlated
  - Individuals from certain location or with certain characteristics may benefit more from a policy than some others
- This matter in the estimations of your standard errors (not necessarly the estimates) because errors are correlated

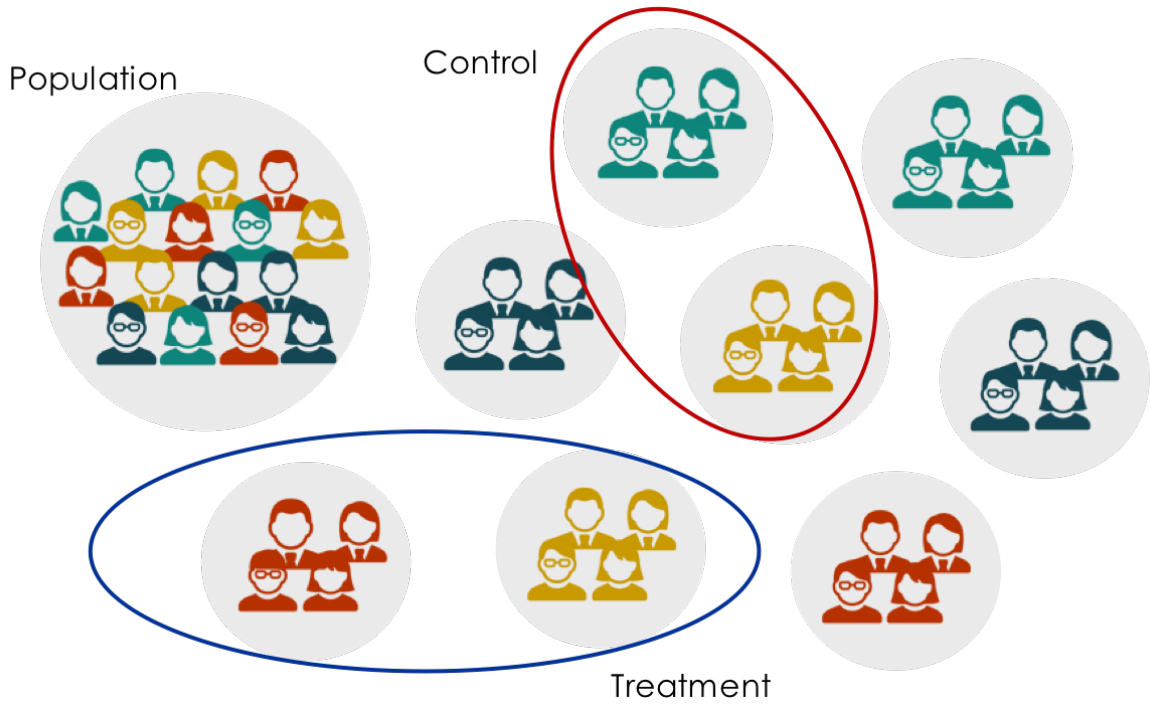
The problems with clusters

Sampling/Treatment with no intracluster correlation



The problems with clusters

Sampling/treatment with high intracluster correlation



The problems with clusters

What is usually meant when one talks about clusters

- Most econometrics texbooks[4] approches the clustering issue as something close to omitted variable bias where, the initial model:

$$Y_{ic} = \alpha + \boldsymbol{X}'\beta + \mu_{ic}$$

actually hides the fact that the error term μ_c has a group structure s.t.:

$$\mu_{ic} = \upsilon_c + \varepsilon_{ic}$$

- And thus, estimating the model without accounting for that yields biased standard errors because $\mathbb{E}[\mu_{ic}\mu_{jc}] = \rho\sigma_\mu^2 > 0$

♭ This presentation, although pedagogical, reinforce the confusion between fixed effect and clustering.

$$(Y_{ic} - \bar{Y}_c) = (\boldsymbol{X_{ic}} - \boldsymbol{\bar{X}_c})'\boldsymbol{\beta} + \mu_{ic} - \bar{\mu_c}$$

4. For instance Cameron and Trivedi 2005; Angrist and Pischke 2008; Wooldridge 2010; Wooldridge 2012

The problems with clusters

What is usually meant when one talks about clusters

- The second approach is usually through panel data and especially Dif in Dif issues.
- The very influential paper by Bertrand, Duflo, and Mullainathan 2004 (QJE) emphasizes the issue of serial correlation in DiD models such as the classic group-time fixed effect estimand:

$$Y_{ict} = \gamma_c + \lambda_t + \boldsymbol{X'}\boldsymbol{\beta} + \varepsilon_{ict}$$

- The problem is that individuals in a given group are likely to suffer from common shocks at some time t such that there is another component hiden in the error above:

$$\varepsilon = \upsilon_{ct} + \eta_{ict}$$

- If these group-time shocks are (assumed) independents, then the situation is closed to the one before and one could cluster by group-time.
- Yet, this is often not true (e.g. if groups are states or region, a bad situation one period is likely to be bad too the next period)

The almost forgotten reason for clustering

"How were your data collected ?"

- "Textbook cases" discussed before are what one may call "model-based" cases for clustering
- These examples implicitely assume that data are collected randomly, or randomly enough.
- However, surveys often use more sophisticated sampling methods with nested structures (e.g. sampling cities, then neighborhoods, then households), stratification and/or weightings.

The first clustering issue should be survey design effect

$\Rightarrow$ Clustering at the primary survey unit (PSU) at the minimum.

Conventional wisdom about standard errors

When to cluster according to Colin Cameron and Miller 2015

Until recently, the conventional wisdom was sumed up as follows:

> *"There are settings where one may not need to use cluster-robust standard errors. We outline several though note that in all these cases it is always possible to still obtain cluster-robust standard errors and contrast them to default standard errors. If there is an appreciable difference, then use cluster robust standard errors". (p.334)*

This is actually wrong according to Abadie et al. (2022)

What does Abadie et al. 2022 change ?

- Adjusting SE for clustering effect is often misunderstood
- Usual recommandations are often too conservatives
- We should cluster:
  - In the presence of heterogenous treatment effect and small number of clusters compared to the overall population
  - when there is correlation between treatment and clusters (cluster assignment)
- We should not cluster:
  - In pure randomized control trial (or any situation without sampling clustering or assignment clustering)
  - when there is constant treatment effect and no clustering in the assignment.

♭ Convincing model but specific to the stated configurations.

♭ Less usefull for less RCT-like designs (e.g. the infamous serial correlation in DID)

Clustering: illustration through simulations

Simulation with $\rho = .5$ intra-cluster correlation

- We consider again the simple bivariate regression

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

- We generate a sample of 1000 observations allocated across 50 clusters (e.g. municipalities) with $\rho = .5$ correlation within cluster.
- Like before, we generate a true β=0
- We estimate this equation using OLS and collect the estimate of $\hat{\beta}$
- We do that 10 000 times with a new random sample and plot the distribution of the estimated $\hat{\beta}$

Clustering: illustration through simulations

Distribution of estimations with clustering

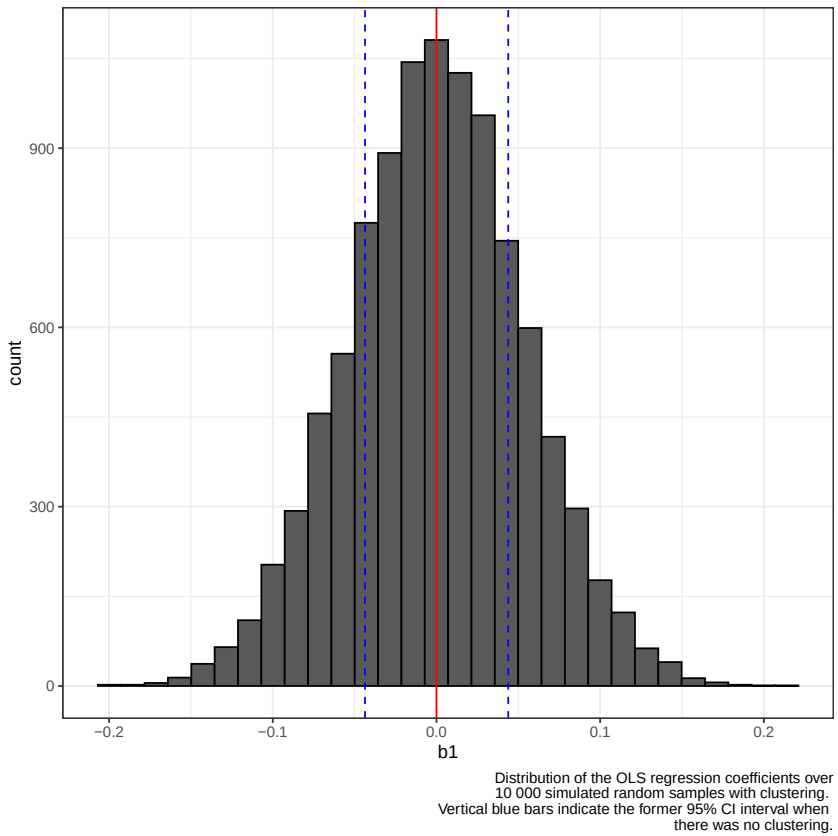


Figure 15: Density of the β estimates with clustering

Clustering: illustration through simulations

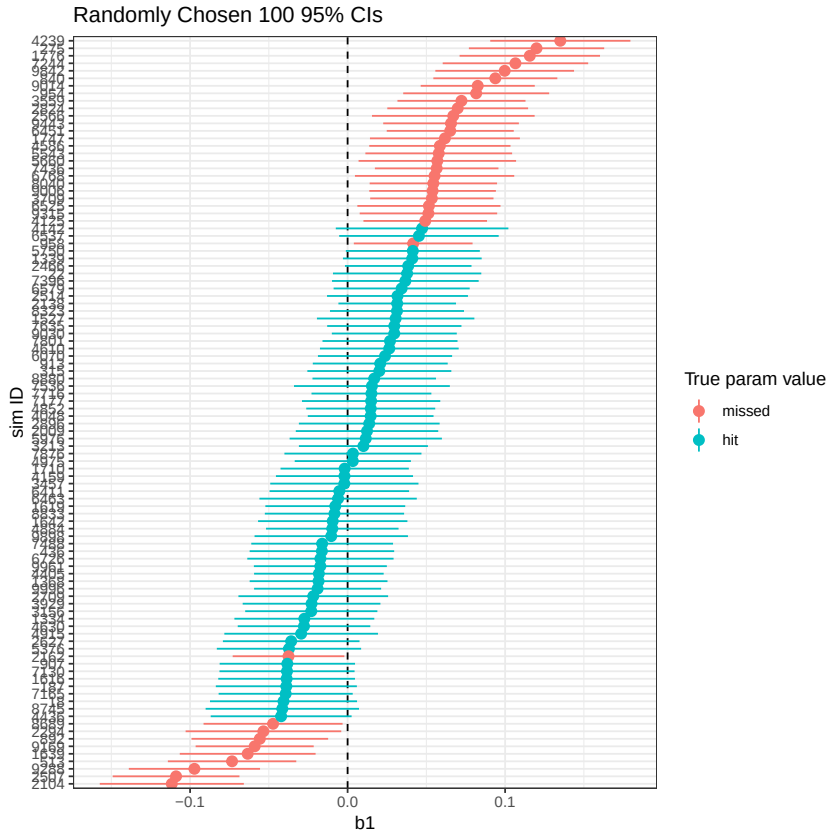


Figure 16: 95 % IC with standard errors assuming independent error and homoscedasticity

Clustering: illustration through simulations

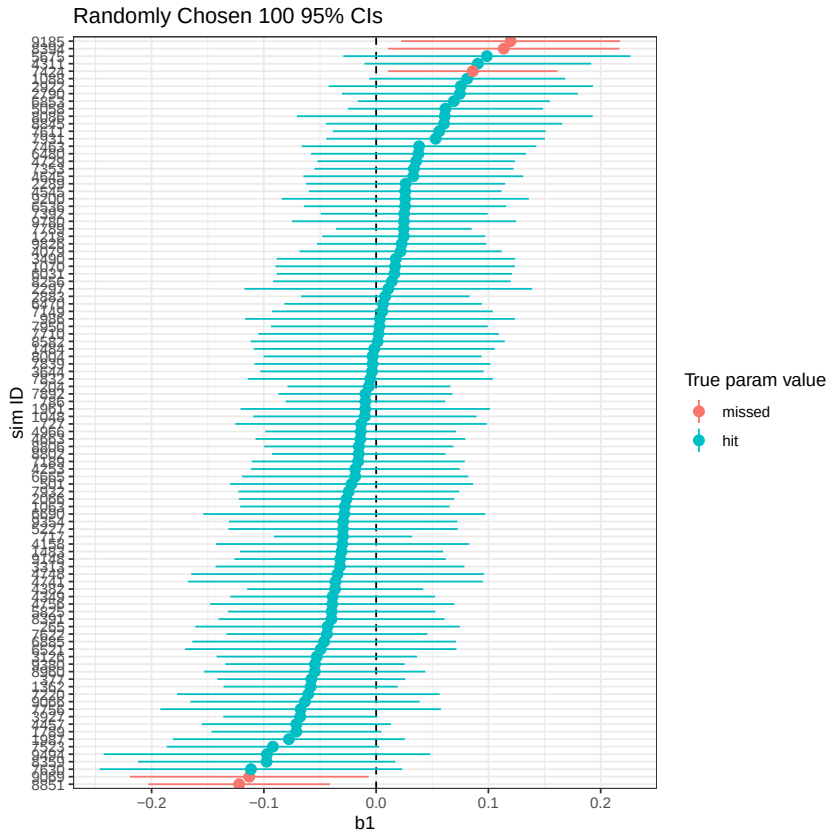


Figure 17: 95 % IC with heteroskedasicity-cluster robust standard errors

Inference: a history of variance

Clustered design: What consequences for estimations

- If nothing else change, OLS regressions are still unbiased BUT standard errors that assume homoscedasticity and independent errors are usually way too small $\Rightarrow$ the false-positive rate in statistical tests is much higher !
- With large samples, the cluster-robust adjustment works well
- With few clusters or few observations within clusters, these standard errors may be too conservatives
- Still very active literature on appropriate inference, especially for more advanced method where no analytical formula exist (yet).

Outline

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Context and motivation

Bruno Crépon et al. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics* 128 (2): 531–580

- In France, unemployment rate is 17.5 % for age 15 30 against 9.2 % in the whole population
- Higher education has traditionally been somewhat protective
  - In France, unemployment rate is 9.4 % for college graduates vs 21.4 % for the others
  - However, even educated youth may experience unemployment and long term unemployment
  - 20 to 30 % of young high school/college graduates have been unemployed for more than 6 months, and around 10 % have been unemployed for more than 12 months

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Context and motivation

- One common policy response is to provide hard to place jobseekers with **reinforced counseling scheme**
- Provide assistance with writing resume, searching for job offers and answering to them, preparing for interviews
- Reinforced counseling programs are **costly** as they mean more frequent meetings with the caseworker
- Intensive support caseworkers have about 30 unemployed in their caseload, instead of 120 in the normal situation
- One strong orientation of the public employment policy was to **use services of private operators** instead of Pôle emploi
- End of the monopoly of the Employment Agency is a key component of the Employment policy in France
- Work through contracts with placement agencies

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A program for young unemployed with tertiary education

- In 2006 the Ministry of Employment launched such a program for 10 000 young people in 10 regions in France
- A program with private operators
- The total fee ranges from 1 600 to 2 100 euros
- Private operators paid in three parts (strong incentives):
  - 1/3 when the youth joins the program 533 to 700 euros)
  - 1/3 when the youth gets (and takes) a job within 6 months with a contract lasting at least 6 months
  - 1/3 if the job lasts indeed at least 6 months
- The objective is to put quickly youth into "stable" jobs

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A program for young unemployed with tertiary education

- Target population:
  - Less than 30 years old
  - Unemployed for more than 6 months (or cumulating more than 12 months of unemployment over the last 18 months)
  - Diploma after 2 years of college
- A usual criticism made to such programs is that they help the participants **at the expense of others**
- Focus on this last issue

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A clever experimental design

Figure 18: The experimental design from Crépon et al. (2013)

4.1 Experimental design

The randomization took place at both the labor market and individual level. It was organized in the areas covered by 235 public unemployment agencies, scattered across 10 administrative regions (about half of France). Each agency represents a small labor market, within which we may observe treatment externalities. On the other hand, the agencies cover areas that are sufficiently large, and workers in France are sufficiently immobile, that we can assume that no spillovers take place across areas covered by different agencies.[15] Migration or spillover would lead us to underestimate the magnitude of externalities. The results we present below are robust to the exclusion of one region (Nord Pas de Calais), which is dominated by a large city (Lille), where treatment and control areas are contiguous.

In order to improve precision, we first formed groups of five agencies that covered areas similar in size and with comparable local populations; we obtained 47 such quintuplets. Within

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A clever experimental design

Figure 19: The experimental design from Crépon et al. (2013)

each of these strata, we randomly selected one permutation assigning the five labor markets to five fractions of treated workers: $P \in \{0, 0.25, 0.50, 0.75, 1\}$.

Every month from September 2007 to October 2008, job seekers who met the criteria for the target population (aged below 30, with at least a two-year college degree, and having spent either 12 out of the last 18 months or six months continuously unemployed or underemployed) were identified by the national ANPE office, using the official unemployment registries.

The list of job seekers was then transmitted to us and we randomly selected a fraction of workers following the assigned proportion into treatment within each agency area. The list of individuals that we selected to be potential beneficiaries of the program was then passed on to the contracted counseling firm in the area, which was in charge of contacting the youth and offering them entry into the job placement program. Entry was voluntary, and the youth could elect to continue receiving services from the local public unemployment agency instead, or no service at all. No youth from the control group could be approached by the firm at any time, and none of them were treated.

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A clever experimental design

- **Block-random assignment** of **clusters** i.e. labor markets.
- Principal statistical unit is the labor market
- Each cluster in a strata has 1/5 of being in one of the 5 **treatment arms**
  - Full Control
  - Full treatment
  - 1/4 treated, 3/4 controls
  - 50-50
  - 3/4 treated, 1/4 controls
- Individuals within clusters are randomly assigned to treatment or control with treatment probability given by the treatment arm.
- Participation is not mandatory $\rightarrow$ *encouragement design*, main results are **intention to treat analysis**.
- Need extra assumption to estimate the average treatment effect on the treated (More on that in session 7 on instrumental variables).

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A clever experimental design

- A "super control group" eligible unemployed workers in 0 assignment areas
- Comparing those assigned to **control groups** and those assigned to the **super control group** identify **displacement effects**. Why ?

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

A clever experimental design

- A "super control group" eligible unemployed workers in 0 assignment areas
- Comparing those assigned to **control groups** and those assigned to the **super control group** identify **displacement effects**. Why ?
- Without displacement effects, exit rates of unemployment should not vary by treatment intensity.
- Comparing those assigned to **treatment groups** and those assigned to the **super control group** identify the average effect on the treated.

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Estimation

Figure 20: Main regressions in Crépon et al. (2013)

We estimate a fully unconstrained reduced form model, and test whether the effect of being assigned to treatment or to control varies by assignment probability. The specification we consider is the following:

$$\begin{aligned} y_{ic} &= \beta_{25} Z_{ic} P_{25c} + \beta_{50} Z_{ic} P_{50c} + \beta_{75} Z_{ic} P_{75c} + \beta_{100} Z_{ic} P_{100_c} \\ &+ \delta_{25} P_{25c} + \delta_{50} P_{50c} + \delta_{75} P_{75c} \\ &+ X_{ic}\gamma_4 + u_{ic} \end{aligned} \tag{6}$$

where Z_{ic} is the assignment to treatment variable and P_{xc} is a dummy variable at the area level indicating an assignment rate of $x\%$. ZP_{25} is thus a dummy for being assigned to treatment in a labor market with a rate of 25% assignment. As before, control variables are individual characteristics (gender, education, etc.) and the set of 47 dummy variables for city quintuplets (our randomization strata). Standard errors are clustered at the local area level. The parameter β_x measures the effect of being assigned to treatment in an area where $x\%$ of the eligible population was assigned to treatment, compared to being unassigned in an area of the same type (or, for β_{100}, compared to the super-control). Coefficient δ_x measures the effect of being assigned to the control group in an area where $x\%$ of the eligible population was assigned to treatment, compared to being in the super-control group in which no one was assigned to treatment. Note

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Estimation

Figure 21: Main table in Crépon et al. (2013)

	Labor market outcome: Long term fixed contract			
		Not employed		
	All workers	All	Men	Women
	(1)	(2)	(3)	(4)
Assigned to treatment in 25% areas	0.016	0.021	0.037	0.015
	(0.012)	(0.014)	(0.027)	(0.016)
Assigned to treatment in 50% areas	0.009	0.013	0.021	0.008
	(0.012)	(0.013)	(0.021)	(0.020)
Assigned to treatment in 75% areas	-0.015	0.007	0.061**	-0.016
	(0.016)	(0.019)	(0.030)	(0.021)
Assigned to treatment in 100% areas	0.010	0.025**	0.021	0.028**
	(0.009)	(0.010)	(0.014)	(0.014)
25% areas	-0.002	-0.015	-0.041**	-0.001
	(0.010)	(0.011)	(0.019)	(0.013)
50% areas	-0.002	-0.014	-0.026	-0.005
	(0.010)	(0.013)	(0.018)	(0.017)
75% areas	0.016	-0.006	-0.055**	0.014
	(0.016)	(0.020)	(0.027)	(0.024)
Control Mean	0.199	0.167	0.150	0.178
F-test for equality of all assigned to treatment coefficients to zero	0.34	0.05	0.07	0.22
F-test for equality of all areas coefficients to zero	0.72	0.48	0.04	0.92
F-test for equality of all areas coefficients	0.52	0.90	0.59	0.77
Number of observations	21431	11806	4387	7419

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Estimation

- Noting significant among all workers, so we focus among those who were unemployed at the time of randomisation.
- The main ITT is in the 4th line. For the unemployment, receiving the treatment instead of not increases fixed-term employment by 2.5pp, slightly more for women.
- Coefficients by treatment intensity and their relative displacement effects are almost perfectly symmetric suggesting that, indeed, job search assistance **impose negative externalities to the untreated**, especially when there are more treated units.
- The F tests at the bottom of the table test the joint hypotheses of null effects or that the effects are equal.
- However, there is little power due to the clustering setting, imperfect compliance and the number of hypotheses to test.
- After that, the regressions **impose some structure** on the parameters and move away from the ITT. Parameters are still causal but their interpretation is not straightforward (OLS weighting and all that).

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Alternative model

Figure 22: Rational for another model Crépon et al. (2013)

Due to the relatively low take-up (and hence the relatively small direct reduced form impact of program assignment on the probability to find a job), and the fact that a sizable fraction of the target sample was in fact already employed when the experiment started, the power of the experiment to detect difference between cities with different assignment is relatively low. Moreover, the average κ (the share of eligible among all young job seekers) is only 19%, which implies that the difference in share of treated between a zone treated at 75% and a zone treated at 25% is only 19*75%-19*25%=9.5%. As a result, even for men alone (where we do find a significant negative impact of being in a treated labor market, for example, and where the pattern has generally the right shape) we cannot reject equality between the dummies indicating different assignment rules. Since our next tests involve subsamples, this will further affect power.

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Alternative model

Figure 23: Alternative regression equation Crépon et al. (2013)

For this reason, we estimate a simpler regression, which exploits the presence of the super-control (with zero probability of treatment assignment), and pools all those who were assigned to control in an area in which some were treated on the one hand, and all those who were assigned to treatment on the other hand. This regression does not allow us to estimate the slope of program effects with respect to the share treated, but has more power against the null that there are no externalities.

The reduced form specification is:

$$y_{ic} = \alpha_2 + \beta_2 Z_{ic} P_c + \delta_2 P_c + X_{ic}\gamma_2 + \omega_{ic} \tag{7}$$

where P_c is a dummy for being in any treatment area (*i.e.* an area with positive share treated). In this specification, β_2 is the difference between those assigned to treatment (whether treated or not), and those who are in treatment zones but are not themselves assigned to treatment. δ_2 is the effect of being untreated in a treated zone (compared to being untreated in an untreated zone). The sum $\beta_2 + \delta_2$ is the effect of being assigned to treatment (compared to being in an entirely unaffected labor market).

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Alternative model

Figure 24: Alternative regression equation Crépon et al. (2013)

Table 5: Reduced form: Impact of the program, accounting for externalities

	By job type: share of job seekers who are eligible for program					
	Not employed			Not employed, above third quartile		
	All	Men	Women	All	Men	Women
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Long term fixed contract					
Assigned to program (β)	0.023***	0.043***	0.013	0.040**	0.072**	0.021
	(0.008)	(0.013)	(0.010)	(0.016)	(0.029)	(0.022)
In a Program area (δ)	-0.013	-0.036***	-0.001	-0.040*	-0.086**	-0.013
	(0.009)	(0.013)	(0.012)	(0.021)	(0.035)	(0.027)
Net effect	0.010	0.007	0.012	0.000	-0.014	0.008
of program assignment (β+δ)	(0.008)	(0.011)	(0.011)	(0.019)	(0.031)	(0.024)
Control Mean	0.16	0.131	0.177	0.19	0.161	0.204

Case study: displacement effects of job search assistance (Crepon Et. Al. 2013)

Main authors' conclusion:

"After eight months, eligible, unemployed youths who were assigned to the program were significantly more likely to have found a stable job than those who were not. But these gains are transitory, and they appear to have come partly at the expense of eligible workers who did not benefit from the program, particularly in labor markets where they compete mainly with other educated workers, and in weak labor markets. Overall, the program seems to have had very little net benefits."

Outline

Wrap-up

Fairly advanced econometric stuff today

- The formal link between RCT and regression analysis
- Case study : typically example of what could be in a test
- Inference: what it means, and the impact of clustering
- Advanced design: A double-nested randomisation
- Most of the things we'll see from now on try to mimic settings that are akin to RCT. This was the core, this you **have to master**.

Next week: Difference in differences

- **To read: mandatory**: Card and Krueger 1993
- **To read: mandatory**: Bertrand, Duflo, and Mullainathan 2004
- **Very good DiD paper**: The impact of Glyphosate on children birth outcomes in Brazil Mateus Dias, Rudi Rocha, and Rodrigo R Soares. 2023. "Down the River: Glyphosate Use in Agriculture and Birth Outcomes of Surrounding Populations." *The Review of Economic Studies* (February 6, 2023): rdad011

Wrap-up

TA session in 15 minutes

1. Power analysis: how to compute the appropriate sample size in your experiment ?
2. Replication crisis: Towards reproducible research
3. How to use R, Rmarkdown, LaTeX to make transparent and reproducible research
4. Replicating claims of Jere R Behrman et al. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *journal of political economy*, 41

See you 28, rue des Saints-Pères room H101

Bibliography I

- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2022. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* 138, no. 1 (December 15, 2022): 1–35.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.
- Athey, S., and G. W. Imbens. 2017. "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, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113, no. 27 (July 5, 2016): 7353–7360.
- Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods Economists Should Know About." *Annual Review of Economics* 11:62.
- Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1148–1178.
- Behrman, Jere R, Susan W Parker, Petra E Todd, and Kenneth I Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *journal of political economy,* 41.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81, no. 2 (April 1, 2014): 608–650.

Bibliography II

- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249–275.
- 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.
- Cai, Zhanrui, Runze Li, and Yaowu Zhang. 2022. "A Distribution Free Conditional Independence Test with Applications to Causal Discovery." *Journal of Machine Learning Research* 23 (February): 1–41.
- Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics : Methods and Applications.* Cambridge University Press.
- Card, David, and Alan Krueger. 1993. *Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania.* w4509. Cambridge, MA: National Bureau of Economic Research, October.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107, no. 5 (May): 261–265.
- Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach." *Annual Review of Economics* 7, no. 1 (August): 649–688.
- Colin Cameron, A., and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372.

Bibliography III

- Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "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.*
- Demirer, Mert, Esther Duflo, Ivan Fernandez-Val, and Victor Chernozhukov. 2017. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments.* The IFS, December 30, 2017.
- Dias, Mateus, Rudi Rocha, and Rodrigo R Soares. 2023. "Down the River: Glyphosate Use in Agriculture and Birth Outcomes of Surrounding Populations." *The Review of Economic Studies* (February 6, 2023): rdad011.
- Filges, Trine, Christoffer Scavenius Sonne-Schmidt, and Bjørn Christian Viinholt Nielsen. 2018. "Small Class Sizes for Improving Student Achievement in Primary and Secondary Schools: A Systematic Review." *Campbell Systematic Reviews* 14 (1): 1–107.
- Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1, no. 4 (October): 387.
- Fryer, Roland G. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* 127 (3): 1210–1261.
- Greene, William H. 2012. *Econometric Analysis.* 7th Edition. PEARSON.
- Heckman, James J., and Rodrigo Pinto. 2022. "The Econometric Model for Causal Policy Analysis." *Annual Review of Economics* 14, no. 1 (August 12, 2022): 893–923.

Bibliography IV

- Horvitz, D. G., and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47, no. 260 (December): 663–685.
- Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *The Annals of Applied Statistics* 7, no. 1 (March): 443–470.
- Imbens, Guido W. 2004. "NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY: A REVIEW." *Review of Economics and Statistics* 86 (1): 4–29.
- Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58, no. 4 (December): 1129–1179.
- Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020. "Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection?" *SSRN Electronic Journal.*
- Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114 (2): 497–532.
- Krueger, Alan B, Eric A Hanushek, and Jennifer King Rice. 2002. *The Class Size Debate.* Mischel and Rothstein. Economic Policy Institute. Washington.
- Lechner, Michael, and Anthony Strittmatter. 2017. "Practical Procedures to Deal with Common Support Problems in Matching Estimation."
- Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7, no. 1 (March): 295–318.

Bibliography V

- Lovell, Michael C. 2010. "A Simple Proof of the FWL Theorem." *The Journal of Economic Education* (August 7, 2010).
- 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.
- Pearl, Judea. 1995. "Causal Diagrams for Empirical Research (with Discussion)." *Biometrika* 82 (4): 662–710.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.
- Sł oczyński, Tymon. 2022. "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights." *The Review of Economics and Statistics* 104, no. 3 (May 9, 2022): 501–509.
- Sł oczyński, Tymon, S. Derya Uysal, and Jeffrey M. Wooldridge. 2022. *Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment,* arXiv:2208.01300, August 2, 2022.
- Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113, no. 523 (July 3, 2018): 1228–1242.
- Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.
- Wooldridge, Jeffrey M. 2012. "Introductory Econometrics: A Modern Approach," 910.

Outline

Outline

The Frisch-Waugh-Lovell theorem in action

Basic illustration: life expectancy and GDP per capita

- Use GAPMINDER package and data to illustrate regressions with the good old relationship between life expectancy and GDP per capita.

```
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

The Frisch-Waugh-Lovell theorem in action

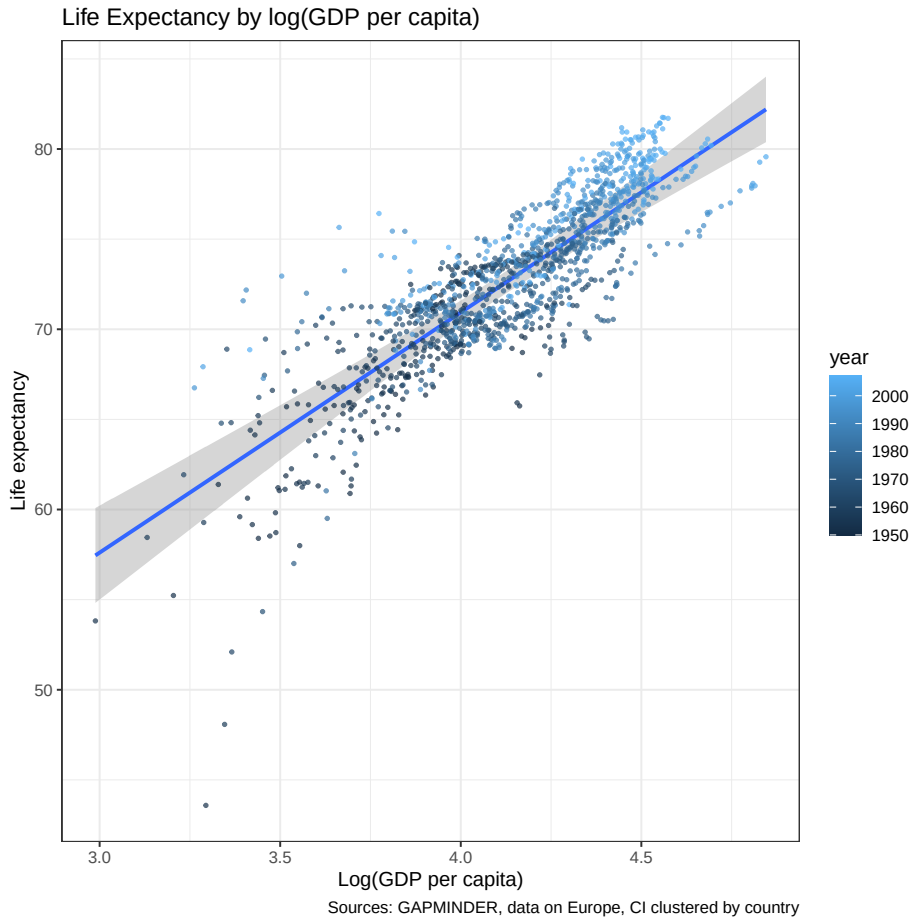


Figure 25: Linear regression of life expectancy on log GDP per capita

The Frisch-Waugh-Lovell theorem in action

What did we do ? What can we do ?

- In the previous slide we estimated the regression:

$$Y_{it} = \alpha + \beta D_{it} + \epsilon_{it}$$

- We want to account for systematic differences between countries and common evolution over years. One way to do that is to define 2 extra sets of dummies $C_i = \mathbb{1}(Country = i)$ and $T_t = \mathbb{1}(year = t)$ that we put in a matrix $\boldsymbol{X}$.
- Consider the regression:

$$Y_{it} = \alpha + \beta D_{it} + \boldsymbol{X}'\delta + \epsilon_{it}$$

- By the FWL theorem, estimating this regression gives the same estimate for $\hat{\beta}$ as estimating subsequently:
  1. $Y_{it} = \alpha_0 + \boldsymbol{X}'\rho + \mu_{it}$ removing the time and country fixed effects in the outcome
  2. $D_{it} = \alpha_1 + \boldsymbol{X}'\eta + \upsilon_{it}$ removing the time and country fixed effects in GDP per capita
  3. $\mu_{it} = \alpha_2 + \beta\upsilon_{it} + \varepsilon_{it}$ the residualized outcome on the residualized "treatment"
- Let's see that

The Frisch-Waugh-Lovell theorem in action

Frisch-Waugh-Lovell in action in R !

```
FW_S1 <- lm_robust(lifeExp ~ factor(year) + factor(country), data = mygapminder,
    cluster = country)

# Retrieve the residual from this first regression, call it res_Life
mygapminder$res_Life <- mygapminder$lifeExp - FW_S1$fitted.values
# Then regress the log of GDP per capita over the same year dummies
FW_S2 <- lm_robust(log(gdpPercap) ~ factor(year) + factor(country), data = mygapminder,
    cluster = country)
# get the residual, call them res_gdp
mygapminder$res_gdp <- log(mygapminder$gdpPercap) - FW_S2$fitted.values

# Now we regress the first residual on the second:
FW <- lm_robust(res_Life ~ res_gdp, data = mygapminder, cluster = country)
# Compare with the regression with controls
Controls <- lm_robust(res_Life ~ res_gdp + factor(year) + factor(country), data = mygapminder,
    cluster = country)
# plot the residual over years to see there's no correlation anymore, but still
# remaining variation with gdp in particular, that we can color code to add a
# dimension ;-)
restime <- ggplot(mygapminder) + geom_point(aes(y = res_Life, x = year, color = log(gdpPercap))) +
    geom_smooth(aes(y = res_Life, x = year), method = "lm_robust", method.args = list(clusters = mygapmind
    scale_color_gradient(low = "blue", high = "orange")
restime
```

The Frisch-Waugh-Lovell theorem in action

We removed time and country fixed variation

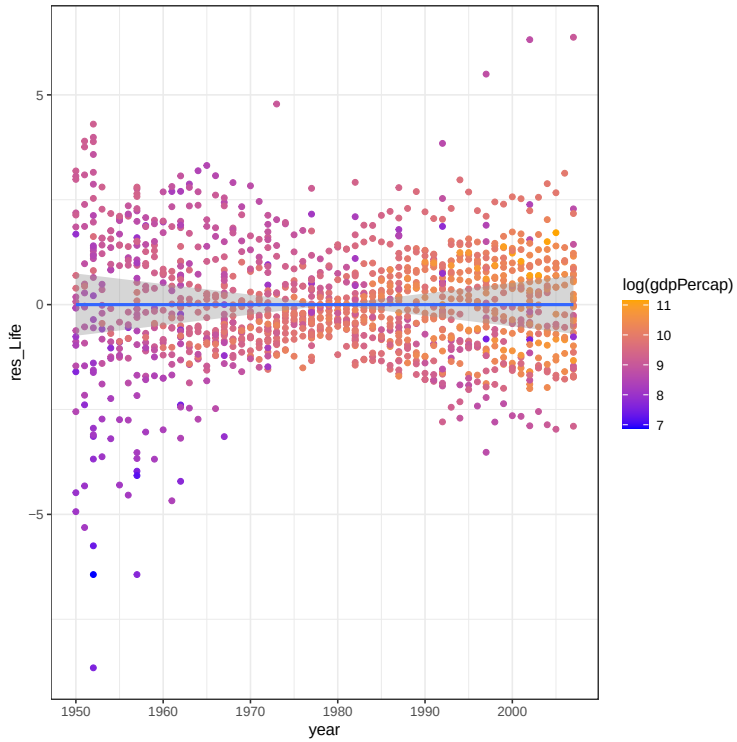
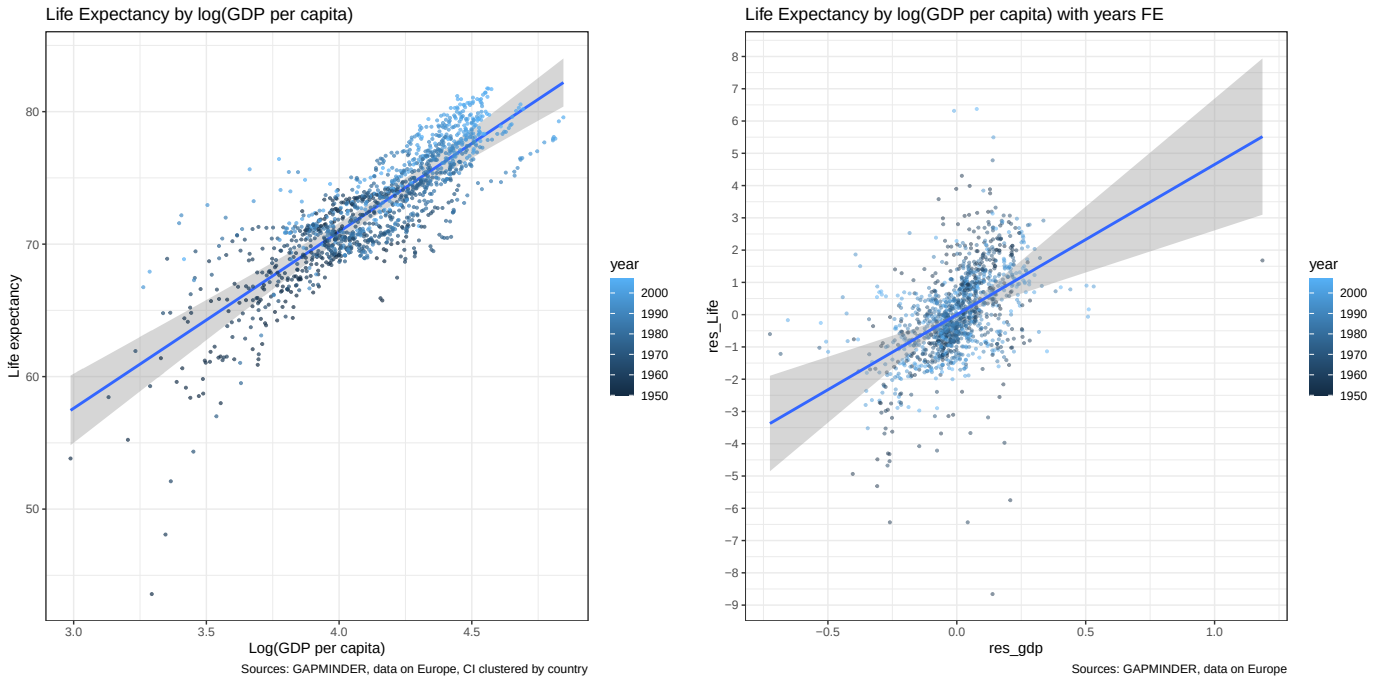


Figure 26: Remaining variation after controlling for time and country fixed effects

The Frisch-Waugh-Lovell theorem in action



(a) No control

(b) After controlling for time and country fixed effects

Figure 27: Evolution of the correlation between GDP and life expectancy after controlling for time and country fixed effects

The Frisch-Waugh-Lovell theorem in action

How to interpret OLS models with log transformations

Table 2: Summary of Functional Forms Involving Logarithms

Model	**Dependent Variable**	**Independent Variable**	**Interpretation of β_1**
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\,\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\,\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Sources:(Wooldridge 2012, p.72)

The Frisch-Waugh-Lovell theorem in action

By FWL the estimated semi-elasticities are the same

Table 3: Illustration of the Frisch-Waugh-Lowell theorem

	Simple Correlation	Controls	Residualized
Log(GDP)	5.788***	4.653***	
	(0.509)	(1.067)	
Log(GDP), residualized			4.653***
			(1.041)
Constant	X	X	X
Year+country Fixed effect		X	
Num.Obs.	1302	1302	1302
R2	0.720	0.927	0.244
R2 Adj.	0.720	0.922	0.243
AIC	5964.9	4395.5	4213.5
BIC	5980.4	4881.6	4229.0
RMSE	2.39	1.22	1.22
Std.Errors	by: country	by: country	by: country

† It's a *level-log* model in base 10, so we interpret the coefficient on gdp by saying "when GDP increases by 1 point, the life expectancy increases by $\hat{\beta}/100$ years

Back to CIA

Outline

Refresher on conditional independence

Independence

- This is a good time for a quick refresher on independence. Two random variables are independent if and only if: $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.
- For discrete random variables: $P(X = x, Y = y) = P(X = x)P(Y = y)$
- In terms of events: $P(A \cap B) = P(A)P(B)$. These definitions are not that intuitive but: What is the conditional probability if two events are **independent**?

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

- So the probability of of A given than B occurs is just $P(A)$. In words, B happening does not affect $P(A)$ (and vice versa)
- Better: knowing one doesn't tell you anything about the other event chances of happening

Refresher on conditional independence

Conditional independence

- Conditional independence is an important concept and closely related to regression models and the conditional independence assumption
- Events A and B are conditionally independent if $P(A \cap B \mid Z) = P(A \mid Z)P(B \mid Z)$
- More useful: If A and B are conditional independent given Z, then $P(A \mid B, Z) = P(A \mid Z)$
- In words, knowing B doesn't tells us anything about $P(A)$ once we know $\mathbf{Z}$

Outline

Estimations with heterogeneous treatment effect

Derivating the ATT and ATU

- First, it's useful to remember that the ATE is a weighted sum of the conditional ATEs: $\beta_X = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$. We compute the overall ATE as $\beta = \sum_x \beta_X Pr[X_i = x]$.
- Note that, by ignorability,

$$\begin{aligned} \beta_X &= \mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, D_i = 0] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] \end{aligned}$$

- There is a similar derivation for the ATT:

$$\begin{aligned} \beta_{ATT} &= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \\ &= \mathbb{E}\Big[\mathbb{E}[Y_i(1) - Y_i(0)|X_i, D_i = 1]|D_i = 1\Big] \text{ by the LIE} \\ &= \mathbb{E}\Big[\big(\mathbb{E}[Y_i(1)|X_i, D_i = 1] - \mathbb{E}[Y_i(0)|X_i, D_i = 1]\big)|D_i = 1\Big] \\ &= \mathbb{E}\Big[\big(\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]\big) \mid D_i = 1\Big] \text{ by ignorability} \\ &= \mathbb{E}[\beta_X \mid D_i = 1] \\ &= \sum_x \beta_X Pr[X_i = x \mid D_i = 1] \end{aligned}$$

Estimations with heterogeneous treatment effect

Derivating the ATT and ATU

- Remember Bayes's formula:
  $\Pr[X_i = x \mid D_i = 1] \Pr[D_i = 1] = \Pr[D_i = 1 \mid X_i = x] \Pr[X_i = x]$, so we can rewrite the ATT as a propensity score-weighted function of the CATEs (with a normalizing factor):

$$\beta_{ATT} = \frac{\sum_x \beta_X \cdot \Pr[D_i = 1 \mid X_i = x] \Pr[X_i = x]}{\sum_x \Pr[D_i = 1 \mid X_i = x] \Pr[X_i = x]}$$

- In words, the Average treatment effect on the treated is the weighted average of block-specific ATE, with weights equal to the conditional treatment probability in the block time the probability of being in this block.
- So the weight is estimated by the share of treated in a block times the share of the sample in the block.
- We make the same derivation with the ATU.

Estimations with heterogeneous treatment effect

What the saturated-in-X regression gives

- Consider the saturated-in-X regression:

$$Y_i = \sum_j \delta_j \mathbb{1}(X_{ij} = 1) + \beta_{OLS} D_i + \varepsilon_i$$

- And the auxiliary regression of treatment over block indicators:

$$D_i = \sum_j \pi_j \mathbb{1}(X_{ij} = 1) + \upsilon_i$$

- This equation is fully saturated and thus estimates $\mathbb{E}[D_i|\boldsymbol{X_i}]$ and thus $\upsilon_i = D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]$.
- By the FWL theorem, β_{OLS} is equivalent to the regression of Y_i on the residual of the previous auxiliary regression:

$$\begin{aligned} \beta_{OLS} &= \frac{Cov(Y_i, \upsilon_i)}{\mathbb{V}[\upsilon_i]} \\ &= \frac{\mathbb{E}\left[Y_i \cdot \left(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\right)\right]}{\mathbb{E}\left[\left(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\right)^2\right]} \end{aligned} \tag{5}$$

Estimations with heterogeneous treatment effect

What the saturated-in-X regression gives

- Remember, estimating the regression of Y on X and D is the same as estimating Y on $\mathbb{E}[Y_i|\boldsymbol{X_i}, D_i]$, so in the big expectation in the numerator, we can substitute Y_i by $\mathbb{E}[Y_i|\boldsymbol{X_i}, D_i]$:

$$\beta_{OLS} = \frac{\mathbb{E}[Y_i|D_i, \boldsymbol{X_i}] \cdot \mathbb{E}[\Big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\Big)]}{\mathbb{E}\Big[(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}])^2\Big]} \tag{6}$$

- We can expand the CEF $\mathbb{E}[Y_i|\boldsymbol{X_i}, D_i]$ further to get:

$$\mathbb{E}[Y_i|\boldsymbol{X_i}, D_i] = \mathbb{E}[Y_i|D_i = 0, X_i] + \beta_X D_i$$

- We then plug this expression in the numerator of the previous equation

$$\begin{aligned}\mathbb{E}[Y_i|D_i, \boldsymbol{X_i}] \cdot \mathbb{E}[\big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\big)] &= \mathbb{E}\Big[\big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\big)\mathbb{E}[Y_i|D_i = 0, \boldsymbol{X_i}]\Big] \\ &+\mathbb{E}\Big[D_i\big(D_i - \mathbb{E}[D_i|X_i]\big)\beta_X\Big]\end{aligned}$$

Estimations with heterogeneous treatment effect

What the saturated-in-X regression gives

- Because $\mathbb{E}[Y_i|D_i = 0, \boldsymbol{X_i}]$ is a function of $\boldsymbol{X_i}$, it is uncorrelated with $\big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\big)$, so the first hand term on the right-hand side is zero ;

$$\mathbb{E}\Big[\big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\big)\mathbb{E}[Y_i|D_i = 0, \boldsymbol{X_i}]\Big] = 0$$

- For the same reason, D_i is uncorrelated with $\big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\big)$ so the numerator actually becomes:

$$\begin{aligned}\mathbb{E}[Y_i|D_i, \boldsymbol{X_i}] \cdot \mathbb{E}[\big(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}]\big)] &= \mathbb{E}\Big[D_i\big(D_i - \mathbb{E}[D_i|X_i]\big)\beta_X\Big] \\ &= \mathbb{E}\Big[\big(D_i - \mathbb{E}[D_i|X_i]\big)^2\beta_X\Big]\end{aligned}$$

- At this point, we have shown:

$$\beta_{OLS} = \frac{\mathbb{E}\Big[\big(D_i - \mathbb{E}[D_i|X_i]\big)^2\beta_X\Big]}{\mathbb{E}\Big[(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}])^2\Big]} = \underbrace{\frac{\mathbb{E}\Big[\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|\boldsymbol{X_i}]\beta_X\Big]}{\mathbb{E}\Big[\mathbb{E}[(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}])^2|\boldsymbol{X_i}]\Big]}}_{\text{Using the LIE again}} \tag{7}$$

Estimations with heterogeneous treatment effect

What the saturated-in-X regression gives

- There are components in this expression we can make sens, in particular:

$$\mathbb{E}[(D_i - \mathbb{E}[D_i|\boldsymbol{X_i}])^2 \, |\boldsymbol{X_i}] = \sigma^2_D(\boldsymbol{X_i})$$

- Is the conditional variance of the treatment given $\boldsymbol{X}$. So:

$$\beta_{OLS} = \frac{\mathbb{E}\Big[\sigma^2_D(\boldsymbol{X_i})\beta_X\Big]}{\mathbb{E}\Big[\sigma^2_D(\boldsymbol{X_i})\Big]} \tag{8}$$

- This establishes that the regression model of Y on block fixed effect and a treatment produces a treatment-variance weighted average of block specific average treatment effects β_X.
- Finally, because D_i is binary

$$\sigma^2_D(\boldsymbol{X_i}) = Pr(D_i|\boldsymbol{X_i}) \cdot (1 - Pr(D_i|\boldsymbol{X_i}))$$

- So,

$$\begin{aligned} \beta_{OLS} &= \frac{\mathbb{E}\Big[Pr(D_i|\boldsymbol{X_i})\cdot(1-Pr(D_i|\boldsymbol{X_i}))\beta_X\Big]}{\mathbb{E}\Big[Pr(D_i|\boldsymbol{X_i})\cdot(1-Pr(D_i|\boldsymbol{X_i}))\Big]} \\ &= \frac{\sum_x \beta_x \Big(Pr(D_i|\boldsymbol{X_i}=x)\cdot(1-Pr(D_i|\boldsymbol{X_i}=x))\Big)Pr(\boldsymbol{X_i}=x)}{\sum_x \Big(Pr(D_i|\boldsymbol{X_i}=x)\cdot(1-Pr(D_i|\boldsymbol{X_i}=x))\Big)Pr(\boldsymbol{X_i}=x)} \end{aligned} \tag{9}$$

Estimations with heterogeneous treatment effect

What the saturated-in-X regression gives

- This shows that the regression estimand weights each covariate-specific treatment effect by
$\big(Pr(D_i|\boldsymbol{X_i} = x) \cdot (1 - Pr(D_i|\boldsymbol{X_i} = x))\big)Pr(\boldsymbol{X_i} = x)$
- The OLS regression put more weights where the treatment variance is highest, not where the treatment probability is highest.
- Comparing the weights for ATT and that of OLS, we clearly see that the weights are "polluted" by 1 - the probability of treatment and if we looked at the weights for the ATU that would be the pollution of the treatment probability.
- Therefore, unless treatment effect is constant or the conditional treatment probabilities are constant and equal to .5, the OLS estimand is not the ATE, nor the ATT, nor the ATU, but a conditional treatment-variance weighted parameter of the ATE.

▸ Back to Heterogeneous treatment effects