Applications in academic research

Lecture 10

Louis SIRUGUE

11/2021

Last time we saw

Causality

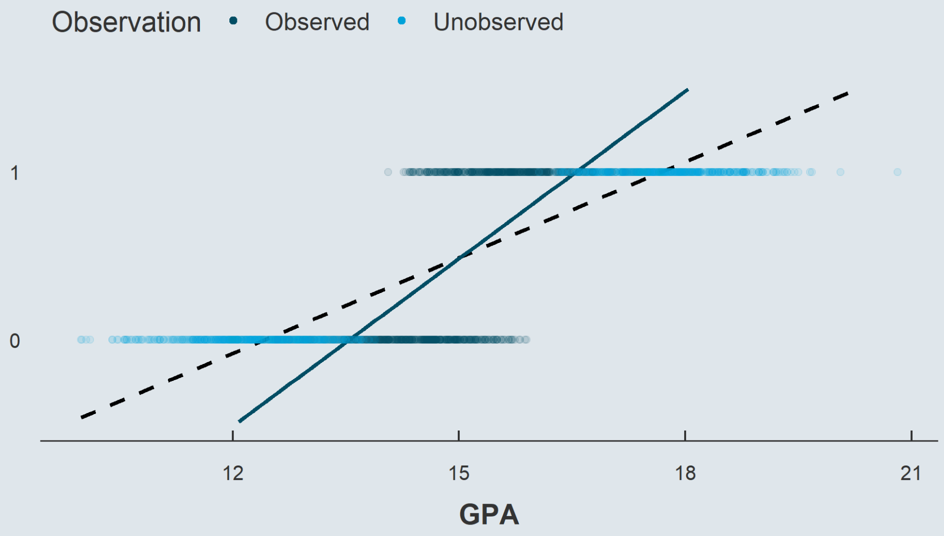
1) Omitted variable bias

- Regressing Earnings on Sex = Male without controls yields $\hat{\beta} = 21612.33$
- But controlling for weekly hours we obtain $\hat{\beta} = 13794.39$

→ Variables that are correlated to both x and y should be controlled for

→ And the coefficients must unambiguously be interpreted *ceteris paribus*

2) Selection bias



- Self-selection selection into the population studied causes problems of **external validity**
- Self-selection into the treatment variable causes problems of **counterfactual validity**

Last time we saw

Theoretical vs. empirical moments

	Theoretical moment	Empirical moment
First moment:	$\mathrm{E}(X_{\text{discrete}}) = \sum_{i=1}^{k} x_i p_i$ $\mathrm{E}(X_{\text{continuous}}) = \int_{\mathrm{R}} x f(x) dx$	$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$
Second moment:	$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \equiv \sigma^2$	$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$

Today: Catch up and Applications in academic research

1. Catch up: Randomness

- 1.1. β vs. $\hat{\beta}$

2. Catch up: Causality from randomness

- 2.1. Randomized Controlled Trials
- 2.2. Types of randomization
- 2.3. Multiple testing

3. Applications in academic research

- 3.1. Labor market discrimination (Behaghel et al., 2015)
- 3.2. Intergenerational mobility (Chetty et al., 2014)

4. Wrap up!

Today: Catch up and Applications in academic research

1. Catch up: Randomness

- 1.1. β vs. $\hat{\beta}$

1. Catch up: Randomness

1.1. β vs. $\hat{\beta}$

- Last time we simulated two normally distributed random variables X and Y of 1,000 observations according to the following theoretical moments:
  - $N = 1000$
  - $\mathrm{E}\left[X\right] = 5$
  - $\mathrm{E}\left[Y\right] = 30$
  - $\mathrm{Var}(X) = 2$
  - $\mathrm{Var}(Y) = 10$
  - $\mathrm{Cov}(X, Y) = 4$

```r
library(MASS)
data <- as_tibble(mvrnorm(1000, c(5, 30), matrix(c(2, 4, 4, 10), 2)))
colnames(data) <- c("x", "y")
c(mean(data$x), mean(data$y))
```

```
## [1]  4.956508 29.980248
```

1. Catch up: Randomness

1.1. β vs. $\hat{\beta}$

- Because we know the joint DGP of X and Y, we do know the actual values of α and β from the regression

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

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{4}{2} = 2$$

$$\alpha = \text{E}\left[Y\right] - \beta \times \text{E}\left[X\right] = 30 - 2 \times 5 = 20$$

- And the coefficients $\hat{\alpha}$ and $\hat{\beta}$ we can compute using observed data are estimates of these true parameters

```
summary(lm(y ~ x, data))$coefficients
```

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 20.033043 0.16304323 122.86951        0
## x            2.006898 0.03150799  63.69488        0
```

1. Catch up: Randomness

1.1. β vs. $\hat{\beta}$

- The higher the number of observations, the closer $\hat{\beta}$ from β on expectation:

```
beta_hat <- function(n){
  data <- as_tibble(mvrnorm(n, c(5, 30), matrix(c(2, 4, 4, 10), 2)))
  colnames(data) <- c("x", "y")
  return(summary(lm(y ~ x, data))$coefficients[2, 1])
}
c(beta_hat(10), beta_hat(1000), beta_hat(100000))
detach("package:MASS", unload = TRUE)
```

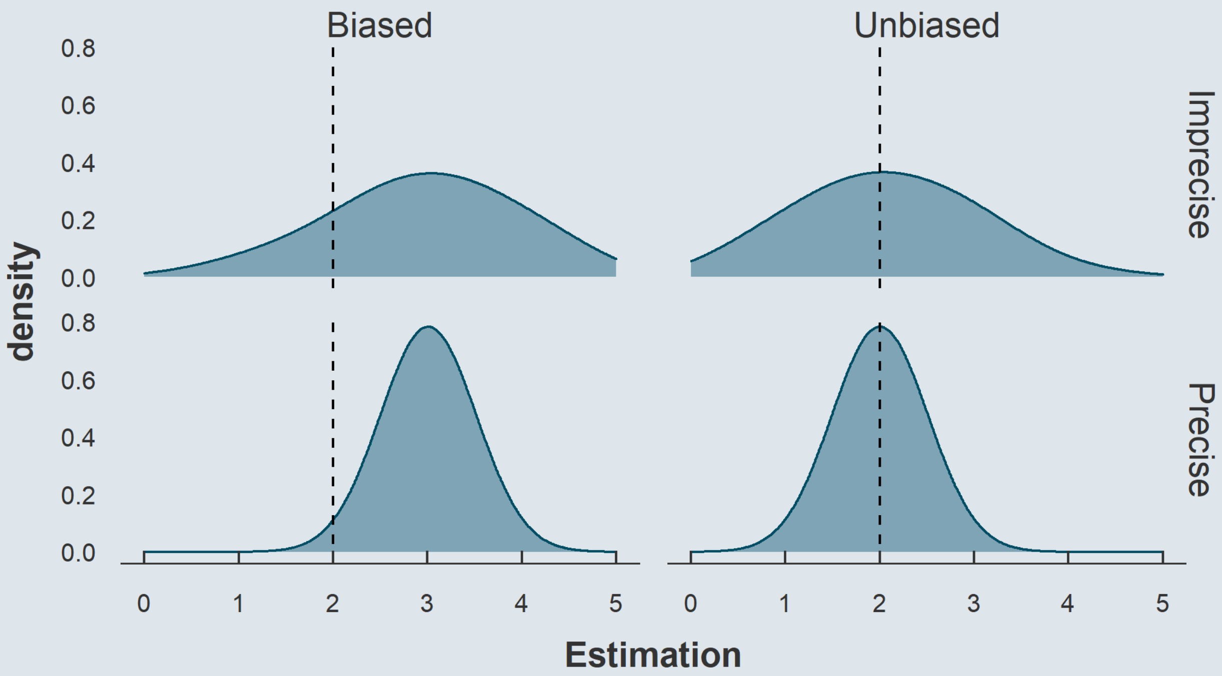
```
## [1] 1.029040 1.986450 1.999659
```

- This is what we call *consistency*
  - With this DGP the OLS estimator is consistent
  - But this is not always the case
  - You'll see the conditions for that next year

1. Catch up: Randomness

1.1. β vs. $\hat{\beta}$

- Keep in mind that consistency, unbiasedness, and precision, are very distinct concepts
  - Consider these 4 cases where we compare the distribution of estimations $\hat{\beta}$ from 1,000 randomly drawn samples to the true β



- An estimator is **unbiased** if on expectation it gives the true value we want to estimate
- An estimator is **precise** if the estimations it provides are close to each other (low variance)
- An estimator is **consistent** if the larger the sample size the higher the probability that we obtain the true value we want to estimate

Overview

1. Catch up: Randomness ✓

- 1.1. β vs. $\hat{\beta}$

2. Catch up: Causality from randomness

- 2.1. Randomized Controlled Trials
- 2.2. Types of randomization
- 2.3. Multiple testing

3. Applications in academic research

- 3.1. Labor market discrimination (Behaghel et al., 2015)
- 3.2. Intergenerational mobility (Chetty et al., 2014)

4. Wrap up!

Overview

1. Catch up: Randomness ✓

- 1.1. β vs. $\hat{\beta}$

2. Catch up: Causality from randomness

- 2.1. Randomized Controlled Trials
- 2.2. Types of randomization
- 2.3. Multiple testing

2. Causality from randomness

2.1. Randomized Controlled Trials

- A Randomized Controlled Trial (RCT) is a type of experiment in which the thing we want to know the impact of (called the treatment) is randomly allocated in the population
  - It is a way to obtain causality from randomness

Take for instance the `asec_2020.csv` dataset we've been working with:

```r
asec_2020 %>% group_by(n()) %>%
  summarise(`Mean Earnings` = mean(Earnings),
           `% Female` = 100 * mean(Sex == "Female"),
           `% Black` = 100 * mean(Race == "Black"),
           `% Asian` = 100 * mean(Race == "Asian"),
           `% Other` = 100 * mean(Race == "Other"),
           `Mean Hours` = mean(Hours))
```

```
## # A tibble: 1 x 7
##   `n()` `Mean Earnings` `% Female` `% Black` `% Asian` `% Other` `Mean Hours`
##   <int>           <dbl>      <dbl>     <dbl>     <dbl>     <dbl>        <dbl>
## 1 64336          62132.       48.1      10.6      7.04      3.76         39.5
```

2. Causality from randomness

2.1. Randomized Controlled Trials

- Let's compare the average characteristics for two randomly selected groups:

```
asec_2020 %>%
  mutate(Group = ifelse(rnorm(n(), 0, 1) > 0, 1, 0)) %>%
  group_by(Group) %>%
  summarise(`Mean Earnings` = mean(Earnings),
            `% Female` = 100 * mean(Sex == "Female"),
            `% Black` = 100 * mean(Race == "Black"),
            `% Asian` = 100 * mean(Race == "Asian"),
            `% Other` = 100 * mean(Race == "Other"),
            `Mean Hours` = mean(Hours))
```

```
## # A tibble: 2 x 7
##   Group `Mean Earnings` `% Female` `% Black` `% Asian` `% Other` `Mean Hours`
##   <dbl>           <dbl>      <dbl>     <dbl>     <dbl>     <dbl>        <dbl>
## 1     0          62440.       47.8      10.8      6.92      3.67         39.6
## 2     1          61826.       48.4      10.5      7.16      3.86         39.5
```

2. Causality from randomness

2.1. Randomized Controlled Trials

- Their average characteristics are very close!
  - On expectation their average characteristics are the same
- And just as the two randomly selected populations are comparable in terms of their observable characteristics
  - On expectation they are also **comparable** in terms of their **unobservable characteristics!**
  - Randomization, if properly conducted, thus solves the problem of omitted variable bias

If we assign a treatment to Group 1, Group 2 would then be a valid counterfactual to estimate a causal effect!

- But RCTs are not immune to every problem:
  - If individuals self-select in participating to the experiment their would be a selection bias
  - Even without self-selection, if the population among which treatment is randomized is not representative there is a problem of external validity
  - For the RCT to work, individuals should comply with the treatment allocation
  - The sample must be sufficiently large for the average characteristics across groups to be close enough to their expected value
  - ...

2. Causality from randomness

2.2. Types of randomization

- To some extent their are ways to deal with these problems
- For instance if we want to ensure that a characteristic is well balanced among the two groups, we can **randomize within categories of this variable**
  - Instead of giving the treatment randomly and hoping that we will obtain the same % of females in both groups
  - We assign the treatment randomly among females and among males separately
  - This is called **randomizing by block**
  - *Note that this only works with observable characteristics!*

```
asec_2020 %>%
  group_by(Sex) %>% # Randomize treatment by sex
  mutate(Group = ifelse(rnorm(n(), 0, 1) > 0, 1, 0)) %>%
  ungroup() %>% group_by(Group) %>%
  summarise(...)
```

2. Causality from randomness

2.2. Types of randomization

- Now imagine that you want to estimate the impact of calory intake at the 10am break on pupils grades
  - You regularly give a snack to a sample of randomly selected children and a few months later you test whether there is a significant difference between their grades and that of untreated children
  - Do you expect the estimated effect to reflect the actual impact you aim to measure?

- What if some children shared their snack with untreated children?
  - These *treated children* would have *less calories* and then possibly lower grades than under full compliance
  - And their *untreated* friends would have *more calories* than expected and then possibly higher grades
  - Thus, this ***spillover effect*** would tend to fallaciously shrink the observed effect of the treatment

- One solution to that problem is to **randomize by cluster**
  - Instead of considering the treatment to be at the child level
  - Consider that the treatment is a the school level
  - A treated unit is a school where some/all children are treated
  - An untreated school is a school where no child is treated

Beware that in terms of inference, computing standard errors the usual way while the treatment is at a broader observational level than the outcome would give fallaciously low standard errors, which would need to be corrected

2. Causality from randomness

2.3. Multiple testing

- Another inference issue that RCTs can be subject to is multiple testing
  - If you conduct a well-designed RCT you might be tempted to exploit the causal framework to test a myriad of effects

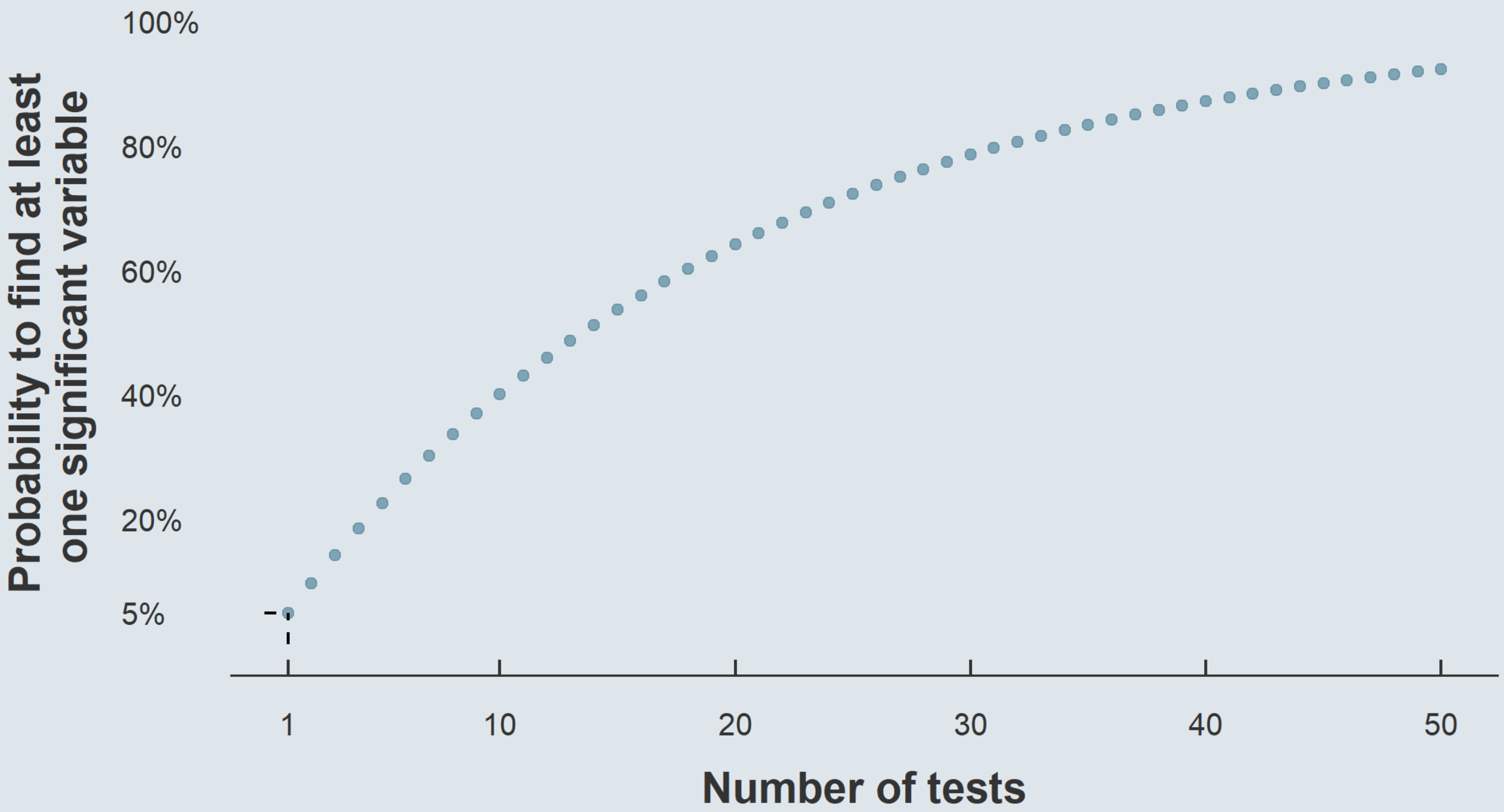
- You randomize your treatment and you compare the averages of many outcomes between treated and untreated individuals
  - You would be tempted to conclude that there is a significant effect for every variable whose corresponding p-value is lower that .05
  - But you cannot do that!

- The probability to have a p-value lower than .05 just by chance for one test is indeed 5%
  - But if you do multiple tests in a row, the probability to have a p-value lower than .05 among these multiple tests is greater than 5%
  - The greater the number of tests you perform, the higher the probability to get a significant result just by chance

This is what we call *multiple testing*

2. Causality from randomness

2.3. Multiple testing



2. Causality from randomness

2.3. Multiple testing

- There are many ways to correct for multiple testing

- The simplest one is called the **Bonferroni** correction
  - It consists in **multiplying the p-value by the number of tests**
  - But it also leads to a large **loss of power** (the probability to find an effect when there is indeed an effect decreases a lot)

- There are more sophisticated ways to deal with the problem, which can be categorized into two approaches
  - **Family Wise Error Rate**: Control the probability that there is at least one true assumption rejected
  - **False Discovery Rate**: Control the share of true assumptions among rejected assumptions

→ *We won't cover these methods in this course but keep the multiple testing issue in mind when you encounter a long series of statistical tests*

Overview

1. Catch up: Randomness ✓

- 1.1. β vs. $\hat{\beta}$

2. Catch up: Causality from randomness ✓

- 2.1. Randomized Controlled Trials
- 2.2. Types of randomization
- 2.3. Multiple testing

3. Applications in academic research

- 3.1. Labor market discrimination (Behaghel et al., 2015)
- 3.2. Intergenerational mobility (Chetty et al., 2014)

4. Wrap up!

Overview

1. Catch up: Randomness ✓

- 1.1. β vs. $\hat{\beta}$

2. Catch up: Causality from randomness ✓

- 2.1. Randomized Controlled Trials
- 2.2. Types of randomization
- 2.3. Multiple testing

3. Applications in academic research

- 3.1. Labor market discrimination (Behaghel et al., 2015)
- 3.2. Intergenerational mobility (Chetty et al., 2014)

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

- Research papers always start with an abstract that briefly describes the study:

Unintended Effects of Anonymous Résumés†

By Luc Behaghel, Bruno Crépon, and Thomas Le Barbanchon*

We evaluate an experimental program in which the French public employment service anonymized résumés for firms that were hiring. Firms were free to participate or not; participating firms were then randomly assigned to receive either anonymous résumés or name-bearing ones. We find that participating firms become less likely to interview and hire minority candidates when receiving anonymous résumés. We show how these unexpected results can be explained by the self-selection of firms into the program and by the fact that anonymization prevents the attenuation of negative signals when the candidate belongs to a minority. (*JEL* J15, J68, J71)

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

Typical structure of an empirical research paper:

- Introduction/literature
- Data/Descriptive statistics
- Empirical framework
- Results
- (Heterogeneity)
- Robustness checks
- Conclusion

Structure of Behaghel et al. (2015) is this one:

- Introduction
- Institutional Background
- Experiment and Data Collection
  - Program and Experimental Design
  - Data Collection
- Impact of Anonymous Résumés
  - Interview Rates
  - Hiring Rates
  - Recruitment Success
  - Robustness Checks
- Mechanisms
  - Firms’ Participation Decision
  - Résumé Valuation by Participating Firms
- Conclusion

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

Program and Experimental Design

1. **Firm entry in the program:** Firms with more than 50 employees posting vacancies lasting at least three months at the public employment service (PES) were offered to enter the program, which consists in having a 50% chance to receive anonymized instead of standard resumes for that vacancy.

2. **Matching of resumes with vacancies:** The PES posts the vacancy on a variety of media, including a public website asking interested job seekers to apply through the PES branch. The PES agent selects resumes from these applicants and from internal databases of job seekers.

3. **Randomization and anonymization:** Resumes are randomly anonymized or not with a 50% probability and sent to the employer.

4. **Selection of resumes by the employer:** The employer selects the resumes of applicants she would like to interview and contact them (through the PES if resumes are anonymized).

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

Data sources

1. **Administrative data**
   - **Coverage:** All firms and all job seekers who used the public employment services in the experimental areas during (and after) the program
   - **Content:** information on the firm (size, sector), on the job position offered (occupation level, type of contract) and limited information on candidates (unless the candidate is filed as unemployed)

2. **Telephone interviews:**
   - **Coverage:** All firms entering the program, a subsample of firms that declined to participate, subsamples of applicants to vacancies posted by these two groups of firms both during and after the experiment
   - **Content:** additional characteristics of the vacancy and of the recruiter (characteristics that could be associated with a differential treatment of candidates), questions on the result of the recruitment (time to hiring and match quality)

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

Sample description

- **1,005 firms entered the program (608 declined):**
  - 385 firms in the control group
  - 366 firms in the treatment group
  - 254 firms not allocated because canceled or job filled too early
- **Sample of 1,268 applicants:**
  - 660 to vacancies from the control group
  - 608 to vacancies from the treatment group
  - 203 to vacancies from firms that withdrew before randomization
- **Main variables:**
  - Whether the candidates is from the minority or the majority
  - Whether the resume was anonymized
  - Whether the employer called back for an interview

- **Authors use sampling weights:**
  - Representativity of the sample
  - Non-response bias correction
  - The weight associated with an individual can be viewed as the number of individuals she represents

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

- Import the data

```r
library(haven)
data_rct <- read_dta("data_candidates_mainsample.dta")
View(data_rct)
```

data_rct

Filter Cols: 1 - 50

	ID_OFFRE Identifier vacancy	ID_CANDIDAT Identifier candidate	CVA Treatment: anonymous resume	REFUSAL Recruiter refused the experiment	ENTRETIEN Interviewed	RECRUTE Hired	POIDS_SEL Sampling weights (within and out of the experiment)	PREN_MUSULMAN Muslim souding name	ZUS_CUCS Deprived neighborhood	ORIGINE_IM_1 Immigrant	ORIGINE_IM_2 Child of immigrant (father)
1	1	1759	1	NA	0	0	5.352863	0	0	0	[illegible]
2	1	6177	1	NA	0	0	5.352863	0	0	0	[illegible]
3	2	371	0	NA	0	0	2.676431	0	0	0	[illegible]
4	3	230	0	NA	0	0	2.676431	1	0	1	[illegible]
5	4	1885	0	NA	0	0	5.352863	0	0	0	[illegible]
6	5	3407	1	NA	0	0	4.014647	1	1	0	[illegible]
7	5	5546	1	NA	0	0	4.906791	0	0	0	[illegible]
8	5	7081	1	NA	0	0	4.906791	0	0	0	[illegible]
9	5	4608	1	NA	0	0	4.906791	0	0	0	[illegible]
10	6	20	1	NA	0	0	28.102530	1	0	1	[illegible]
11	8	5943	1	NA	0	0	2.007324	0	0	0	[illegible]
12	8	6568	1	NA	0	0	2.007324	0	0	0	[illegible]
13	9	4408	0	NA	0	0	21.411451	0	0	0	[illegible]
14	11	6384	1	NA	0	0	21.411451	0	0	0	[illegible]

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

- Recode the data

```
data_rct <- data_rct %>% filter(!is.na(CVA)) %>%                # Keep participating firms
   mutate(treatment = ifelse(CVA == 1, "Treatment", "Control"),   # Recode treatment variable
          minority = ifelse(ZouI == 1, "Minority", "Majority")) %>% # Recode minority variable
  rename(interviewed = ENTRETIEN, weight = POIDS_SEL)            # Rename outcome and weight

head(data_rct %>% select(treatment, minority, interviewed, weight), 5)
```

```
## # A tibble: 5 x 4
##   treatment minority interview weight
##   <chr>     <chr>        <dbl>  <dbl>
## 1 Treatment Majority         0   5.35
## 2 Treatment Minority         0   5.35
## 3 Control   Majority         0   2.68
## 4 Control   Minority         0   2.68
## 5 Control   Majority         0   5.35
```

→ We want to know whether anonymizing resume helped reducing labord market discrimination toward the minority group

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

- Authors use the following notations
  - An indicates whether the resume is anonymous
  - D indicates whether the candidate is from the minority
  - Y indicates whether the candidate obtained an interview

- The parameter of interest then writes:

$$\delta = \underbrace{(\overline{Y}^{An=1,D=1} - \overline{Y}^{An=1,D=0})}_{\substack{\text{Difference in interview rates} \\ \text{between the majority and the minority} \\ \text{when resumes are anonymized}}} - \underbrace{(\overline{Y}^{An=0,D=1} - \overline{Y}^{An=0,D=0})}_{\substack{\text{Difference in interview rates} \\ \text{between the majority and the minority} \\ \text{when resumes are \underline{not} anonymized}}}$$

➜ What sign do you expect for δ?

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

```r
means <- data_rct %>% group_by(treatment, minority) %>%
  summarise(means = weighted.mean(interview, weight))
```

treatment	minority	means
Control	Majority	0.12
Control	Minority	0.09
Treatment	Majority	0.18
Treatment	Minority	0.05

```r
means %>%
  summarise(discrim = means[minority == "Minority"] - means[minority == "Majority"]) %>%
  summarise(delta = discrim[treatment == "Treatment"] - discrim[treatment == "Control"])
```

```
## [1] -0.1067092
```

Practice

1) Estimate this parameter of interest using a regression

Hint: To apply weights in a regression you can indicate the weighting variable in the `weights` *argument*

```
lm(y ~ x1 + x2 + ..., data, weights = )
```

- Reminder:

```
library(tidyverse)
library(haven)
data_rct <- read_dta("data_candidates_mainsample.dta") %>%            # read .dta data
  filter(!is.na(CVA)) %>%                                              # Keep participating firms
  mutate(treatment = ifelse(CVA == 1, "Treatment", "Control"),         # Recode treatment variable
         minority = ifelse(ZouI == 1, "Minority", "Majority")) %>%     # Recode minority variable
  rename(interview = ENTRETIEN, weight = POIDS_SEL)                    # Rename outcome and weight
```

$$\delta = \underbrace{(\overline{Y}^{An=1,D=1} - \overline{Y}^{An=1,D=0})}_{\begin{array}{c}\text{Difference in interview rates}\\ \text{between the majority and the minority}\\ \text{when resumes are anonymized}\end{array}} - \underbrace{(\overline{Y}^{An=0,D=1} - \overline{Y}^{An=0,D=0})}_{\begin{array}{c}\text{Difference in interview rates}\\ \text{between the majority and the minority}\\ \text{when resumes are }\underline{\text{not}}\text{ anonymized}\end{array}}$$

Solution

- To see how the difference in means between the minority and the majority varies between the treatment and the control group, these two variables should be interacted:

$$Y_i = \alpha + \beta D_i + \gamma An_i + \delta D_i \times An_i + \varepsilon_i$$

```
summary(lm(interview ~ minority + treatment + minority*treatment,
           data_rct, weights = weight))$coefficients[, c(1, 4)]
```

```
##                                        Estimate     Pr(>|t|)
## (Intercept)                          0.11638530 1.575140e-12
## minorityMinority                    -0.02365790 3.180243e-01
## treatmentTreatment                   0.06101349 1.181630e-02
## minorityMinority:treatmentTreatment -0.10670915 2.210982e-03
```

  - The constant is the interview rate for individuals in both reference groups (majority/control), and interview rates for each group can be retrieved by adding the relevant coefficients to the constant
  - The coefficient associated with the minority variable is thus the difference in means between the minority and the majority group
  - And the coefficient associated with the interaction is how this difference in means differ between the treatment and the control group

3. Applications in academic research

3.1. Labor market discrimination (Behaghel et al., 2015)

- Why the effect is negative?

TABLE 7—INTERVIEW AND HIRING RATES IN FIRMS ACCEPTING TO PARTICIPATE (*and Randomized in the Control Group*) AND IN REFUSING FIRMS

	All	Minority (D)	Majority (ND)	Gap (D–ND)
Panel A. Interview rates				
Participating (only controls) firms (c)	0.143	0.146	0.140	0.006
	(0.015)	(0.021)	(0.022)	(0.031)
Nonparticipating firms (r)	0.146	0.073	0.210	–0.137**
	(0.038)	(0.028)	(0.059)	(0.063)
Difference (c-r)	–0.004	0.073**	–0.070	0.143**
	(0.041)	(0.035)	(0.063)	(0.070)
Number of candidates	1,378	759	619	1,378
Number of vacant jobs	507	374	376	507

- Self-selection issue:
  - There is a large difference between participating and non-participating firms in how their interview rates differ between the minority and the majority group
  - Participating firms are not less likely to interview a candidate from the minority group, while it is the case for non-participating firm
  - The difference in the interview gap between participating and non-participating firms amount to 14pp

3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

Abstract

We use administrative records on the incomes of more than 40 million children and their parents to describe three features of intergenerational mobility in the United States. First, we characterize the joint distribution of parent and child income at the national level. The conditional expectation of child income given parent income is linear in percentile ranks. On average, a 10 percentile increase in parent income is associated with a 3.4 percentile increase in a child's income. Second, intergenerational mobility varies substantially across areas within the U.S. For example, the probability that a child reaches the top quintile of the national income distribution starting from a family in the bottom quintile is 4.4% in Charlotte but 12.9% in San Jose. Third, we explore the factors correlated with upward mobility. High mobility areas have (1) less residential segregation, (2) less income inequality, (3) better primary schools, (4) greater social capital, and (5) greater family stability. While our descriptive analysis does not identify the causal mechanisms that determine upward mobility, the publicly available statistics on intergenerational mobility developed here can facilitate research on such mechanisms.

3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

- How to characterize the joint distribution of parent and child income?

The intergenerational elasticity:

$$\log(y_i^c) = \alpha + \beta_{IGE} \log(y_i^p) + \varepsilon_i$$

→ $\hat{\beta}$ would be the expected percentage increase in child income for a 1% increase in parent income

The rank-rank correlation:

$$\text{percentile}(y_i^c) = \alpha + \beta_{RRC}\text{percentile}(y_i^p) + \varepsilon_i$$

- In this particular case, because the dependant and the independant variables have the same variance, the regression coefficient equals the correlation coefficient

3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

$$\beta = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$$

$$= \frac{\text{Cov}(x,y)}{\text{SD}(x)\times\text{SD}(x)} \times \frac{\text{SD}(y)}{\text{SD}(y)}$$

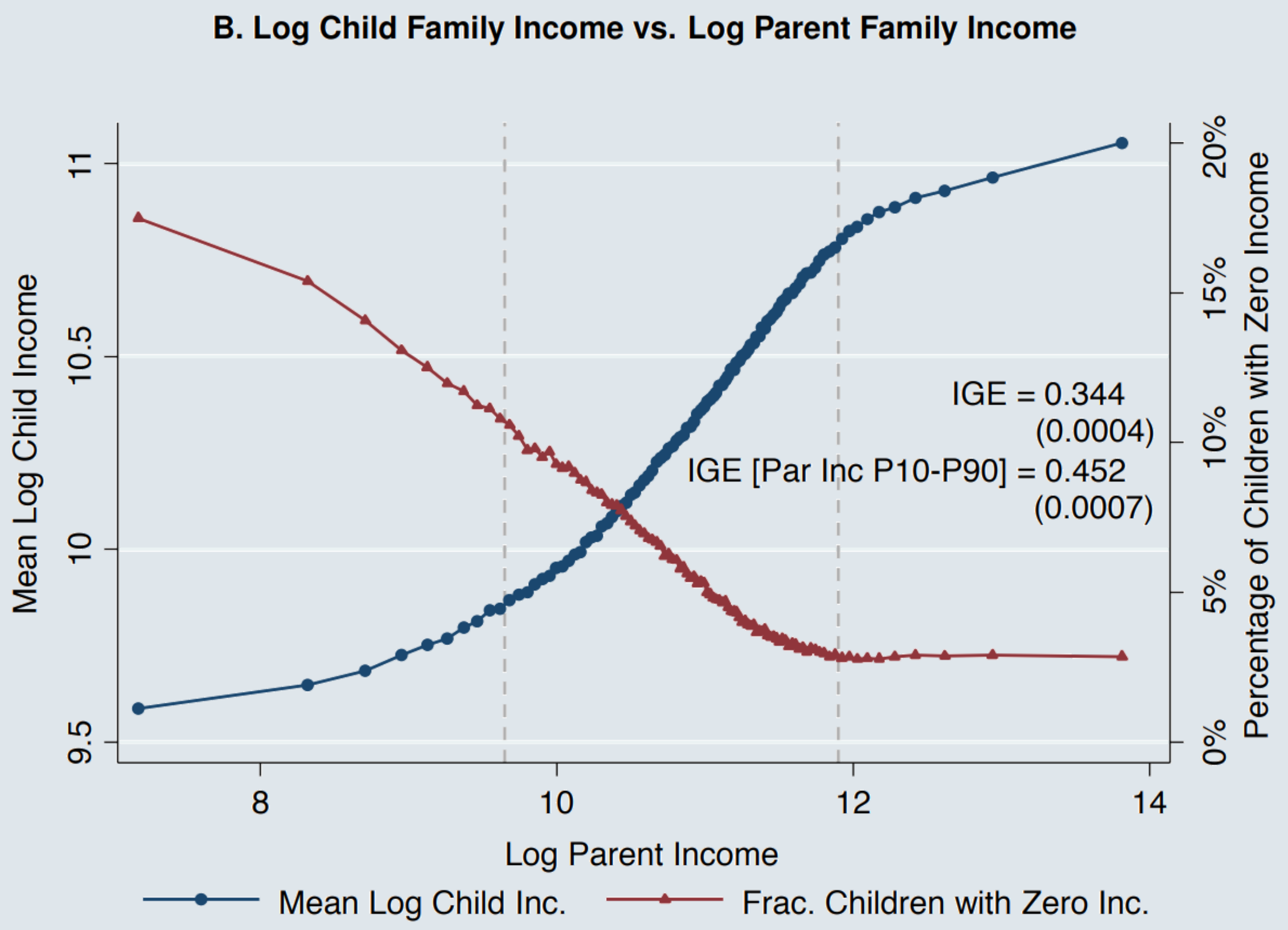
$$= \frac{\text{Cov}(x,y)}{\text{SD}(x)\times\text{SD}(y)} \times \frac{\text{SD}(y)}{\text{SD}(x)}$$

$$= \text{Cor}(x,y) \times \frac{\text{SD}(y)}{\text{SD}(x)}$$

- $\text{SD}(\log(y_i^c)) \lesseqqgtr \text{SD}(\log(y_i^p))$
  - The standard deviation of log income can be viewed as a measure of inequality
  - The IGE is sensitive to relative inequality across generations
- $\text{SD}(\text{percentile}(y_i^c)) = \text{SD}(\text{percentile}(y_i^p))$
  - The RRC is *not* sensitive to relative inequality across generations
  - And the regression coefficient indeed equals the correlation coefficient

3. Applications in academic research

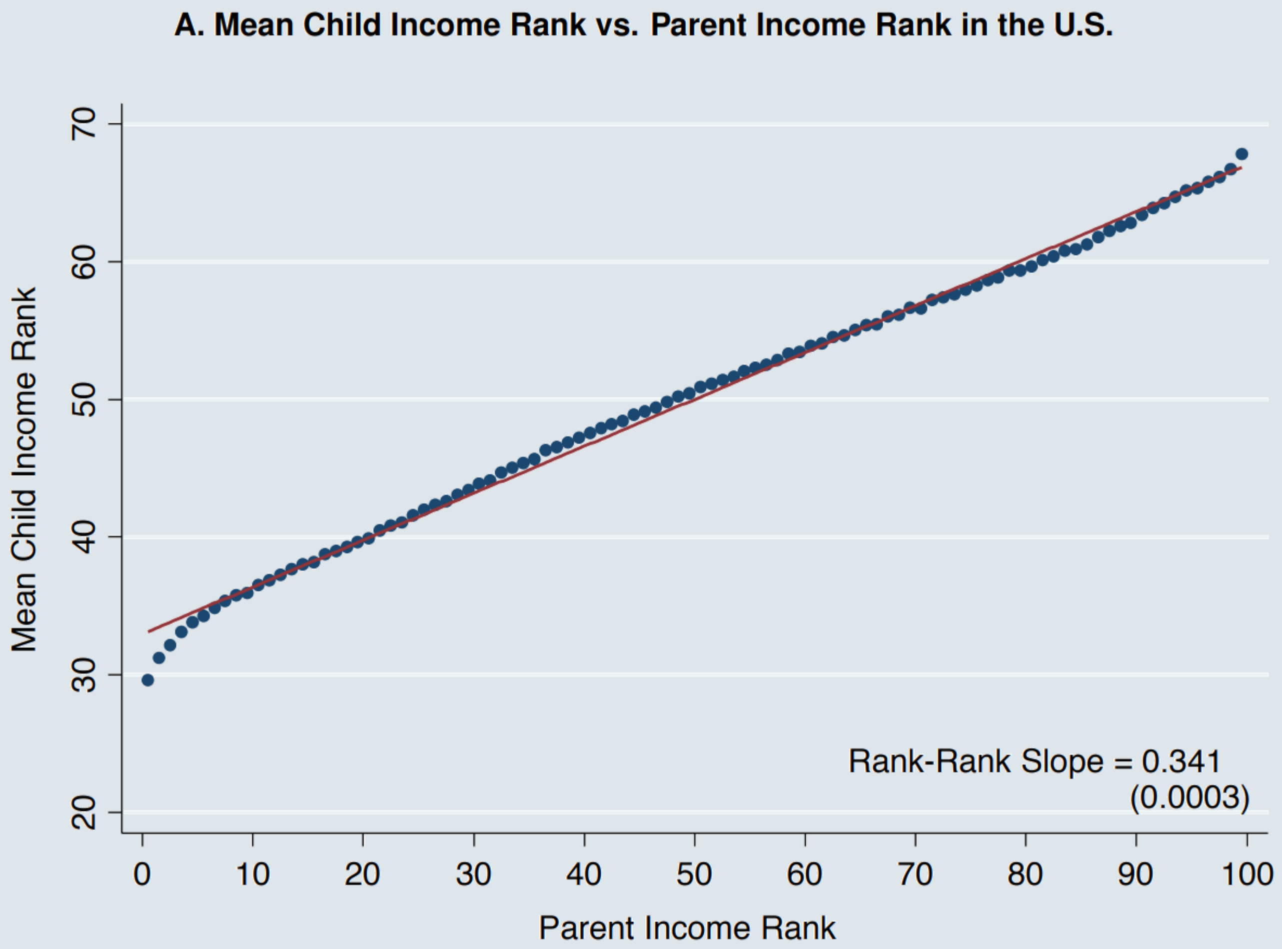
3.2. Intergenerational mobility (Chetty et al., 2014)



- The IGE also implies to manage 0 and negative income
- Which is not the case with the RRC

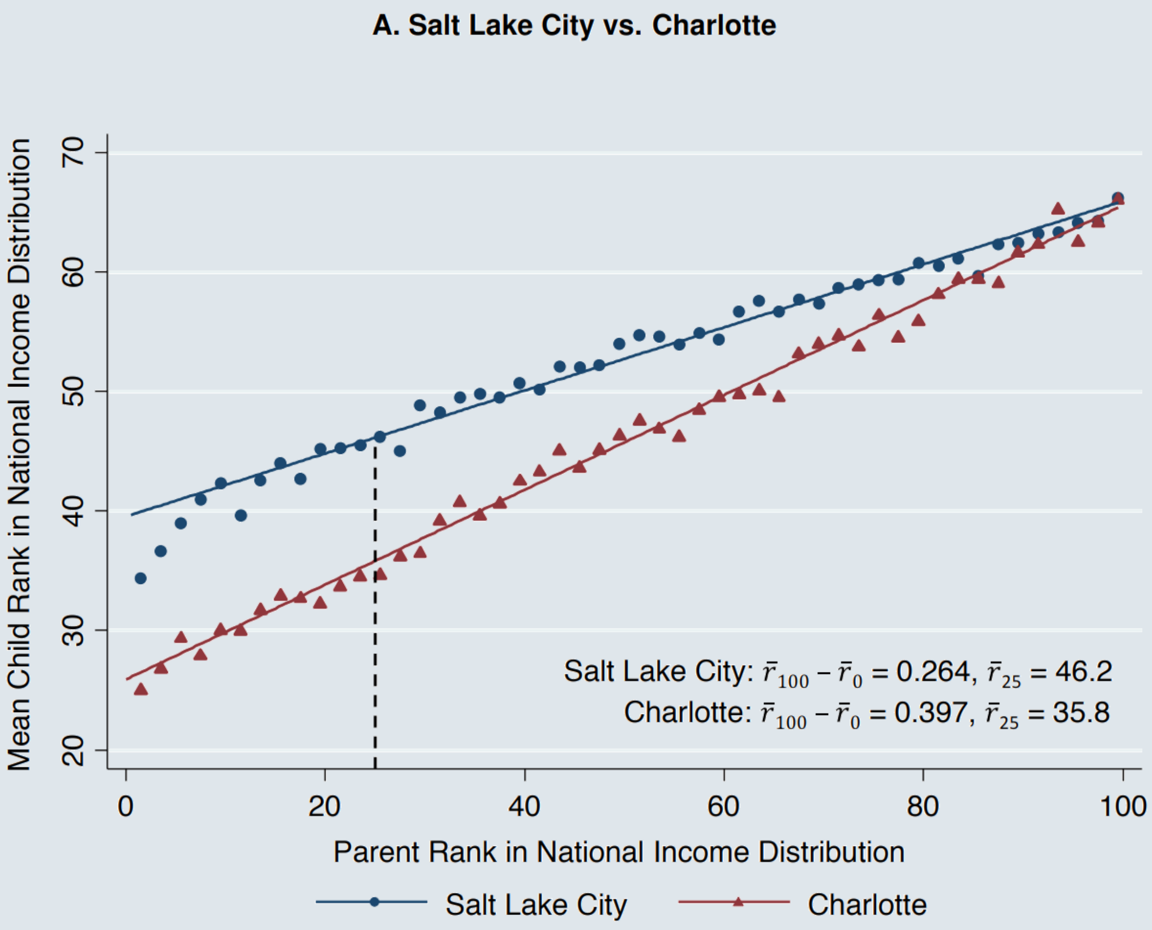
3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)



3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)



Recall:

$$\text{percentile}(y_i^c) = \alpha + \beta_{RRC}\text{percentile}(y_i^p) + \varepsilon_i$$

From this equation we can estimate:

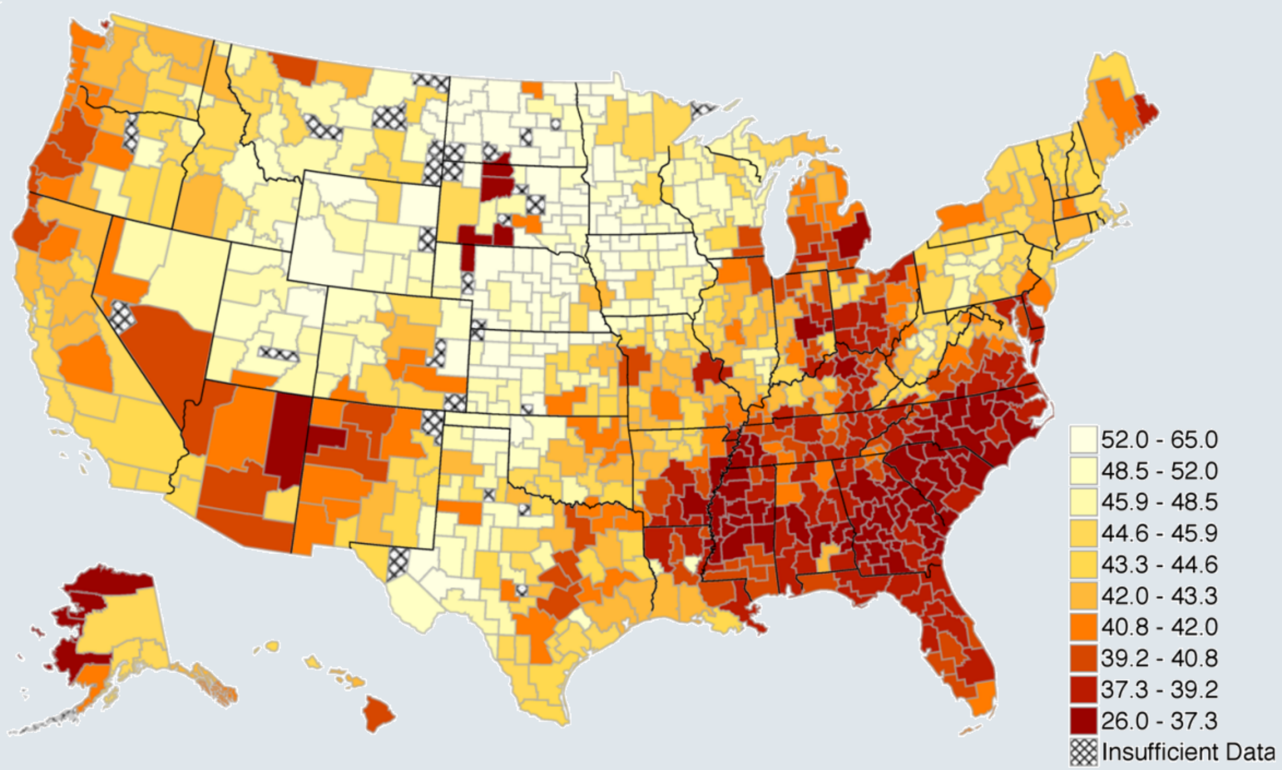
- Relative mobility: $\widehat{\beta_{RRC}}$
- Absolute mobility: $\widehat{\alpha} + 25 \times \widehat{\beta_{RRC}}$

And then estimate it separately for each commuting zone

3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

A. Absolute Upward Mobility: Mean Child Rank for Parents at 25th Percentile $(\bar{r}_{25})$ by CZ



3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

- Authors then investigate whether local characteristics of commuting zones are related to upward mobility
- But regressing directly upward mobility on different characteristics would give:
  - Lower coefficients for variables with bigger metrics (test scores)
  - Higher coefficients for variables with smaller metrics (frac. single moms)

- So authors standardize their variables for the comparability of their estimates

$$\beta = \frac{\mathrm{Cov}(\frac{x}{\mathrm{SD}(x)}, \frac{y}{\mathrm{SD}(y)})}{\mathrm{Var}(\frac{x}{\mathrm{SD}(x)})}$$

- To simplify this equation, you need to know that:
  - $\mathrm{Var}(kX) = k^2\mathrm{Var}(X)$
  - $\mathrm{Cov}(k_1X, k_2Y) = k_1k_2\mathrm{Cov}(X, Y)$

3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

$$\beta = \frac{\text{Cov}(\frac{x}{\text{SD}(x)}, \frac{y}{\text{SD}(y)})}{\text{Var}(\frac{x}{\text{SD}(x)})}$$

$$= \frac{\frac{1}{\text{SD}(x)\text{SD}(y)}\text{Cov}(x,y)}{\frac{1}{\text{SD}(x)^2}\text{Var}(x)}$$

$$= \frac{\text{Cov}(x,y)}{\text{SD}(x)\text{SD}(y)} \times \frac{\text{SD}(x)^2}{\text{Var}(x)}$$

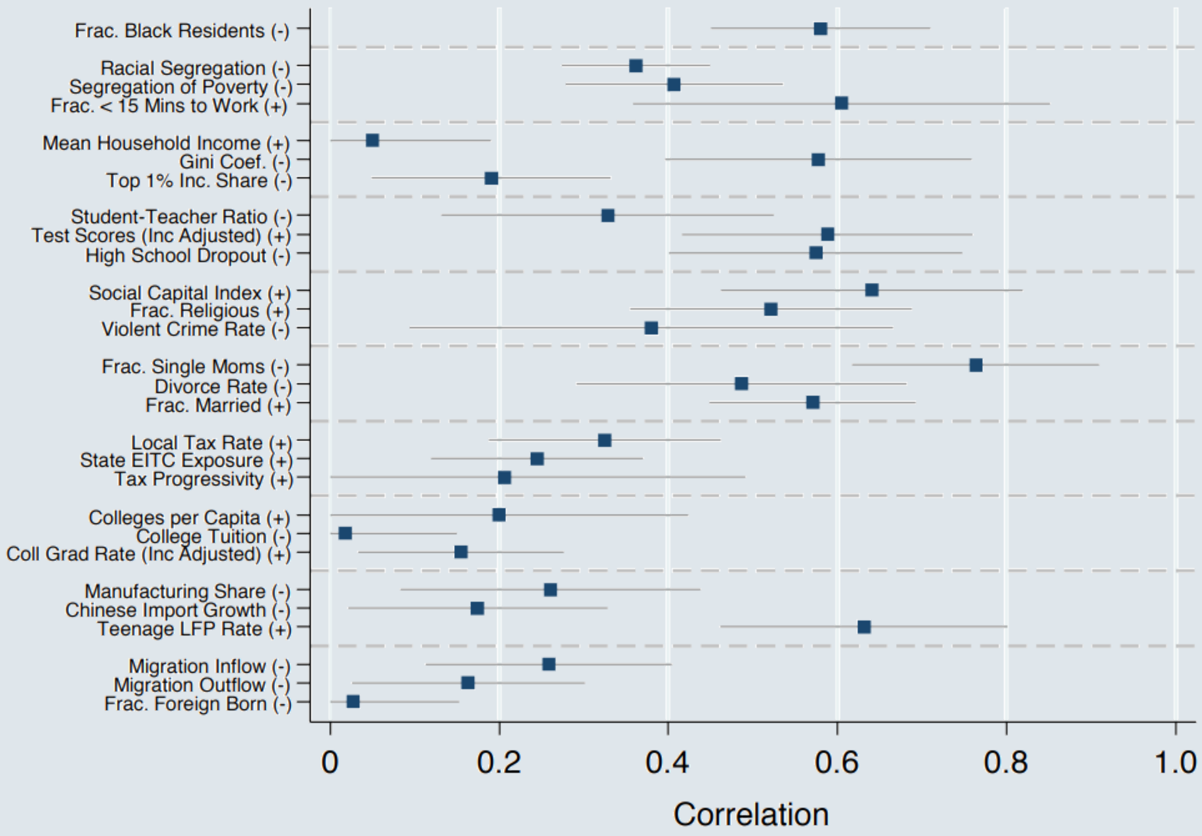
$$= \text{Corr}(x,y)$$

➜ ***Standardizing variables allows to obtain a correlation coefficient from a regression***

3. Applications in academic research

3.2. Intergenerational mobility (Chetty et al., 2014)

FIGURE VIII: Correlates of Spatial Variation in Upward Mobility



Note that these coefficients combine:

- A neighborhood effect
- A selection effect

Overview

1. Catch up: Randomness ✓

- 1.1. β vs. $\hat{\beta}$

2. Catch up: Causality from randomness ✓

- 2.1. Randomized Controlled Trials
- 2.2. Types of randomization
- 2.3. Multiple testing

3. Applications in academic research ✓

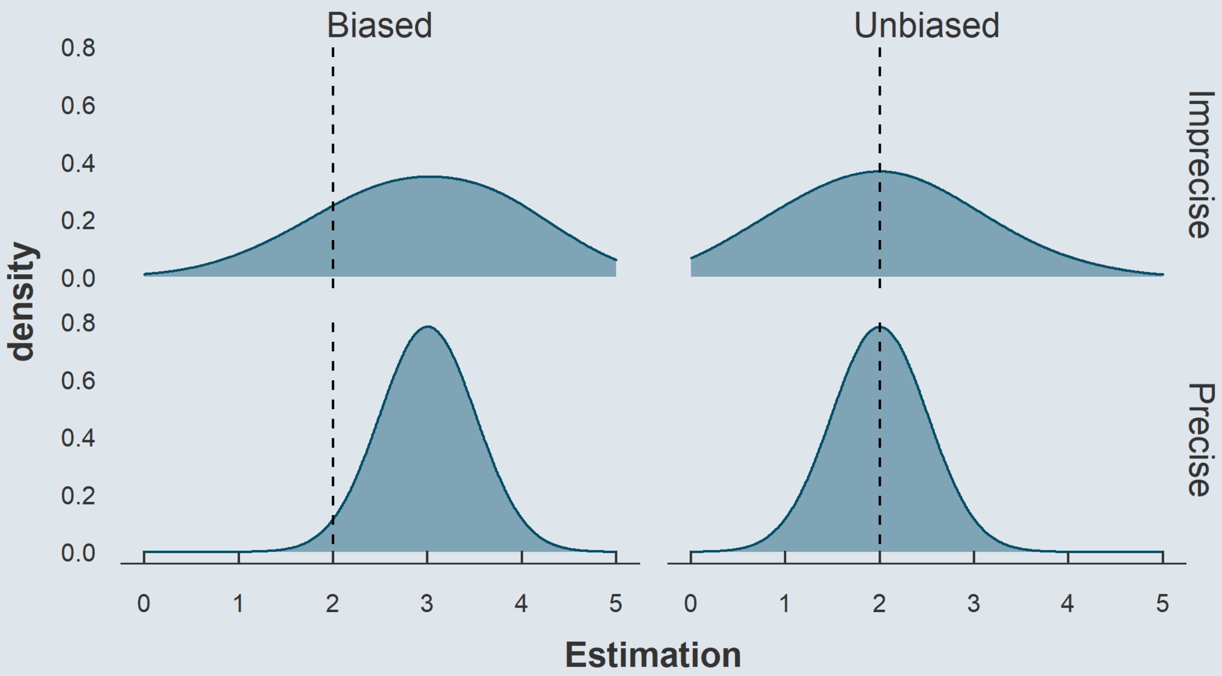
- 3.1. Labor market discrimination (Behaghel et al., 2015)
- 3.2. Intergenerational mobility (Chetty et al., 2014)

4. Wrap up!

4. Wrap up!

β vs. $\hat{\beta}$

- Keep in mind that consistency, unbiasedness, and precision, are very distinct concepts
    - Consider these 4 cases where we compare the distribution of estimations $\hat{\beta}$ from 1,000 randomly drawn samples to the true β



- An estimator is **unbiased** if on expectation it gives the true value we want to estimate
- An estimator is **precise** if the estimations it provides are close to each other (low variance)
- An estimator is **consistent** if the larger the sample size the higher the probability that we obtain the true value we want to estimate

4. Wrap up!

Randomized controlled trials

- A Randomized Controlled Trial (RCT) is a type of experiment in which the thing we want to know the impact of (called the treatment) is randomly allocated in the population
  - It is a way to obtain causality from randomness as on expectation two randomly drawn population have the same average observable and unobservable characteristics, which solves the omitted variable bias

- But RCTs are not immune to every problem:
  - Self-selection issues can arise
  - The population should be representative for external validity
  - Individuals should comply with the treatment allocation
  - The sample must be sufficiently large
  - ...

- There are different types of randomization to help dealing with such problems
  - **Randomization by block for small samples:** Randomly assign the treatment within groups of individuals whose characteristic should be balanced
  - **Randomization by cluster for spillovers:** If spillovers may occur within given units, consider these units as the observational level for the treatment allocation
  - ...

4. Wrap up!

Labor market discrimination (Behaghel et al., 2015)

- Applicants resumes randomly anonymized or not before being sent to employers

$$Y_i = \alpha + \beta D_i + \gamma An_i + \delta D_i \times An_i + \varepsilon_i$$

- $\hat{\delta}$ captures how the difference in interview rates between the minority and the majority differs between the treated and the control employers

```
summary(lm(interview ~ minority + treatment + minority*treatment,
           data_rct, weights = weight))$coefficients[, c(1, 4)]
```

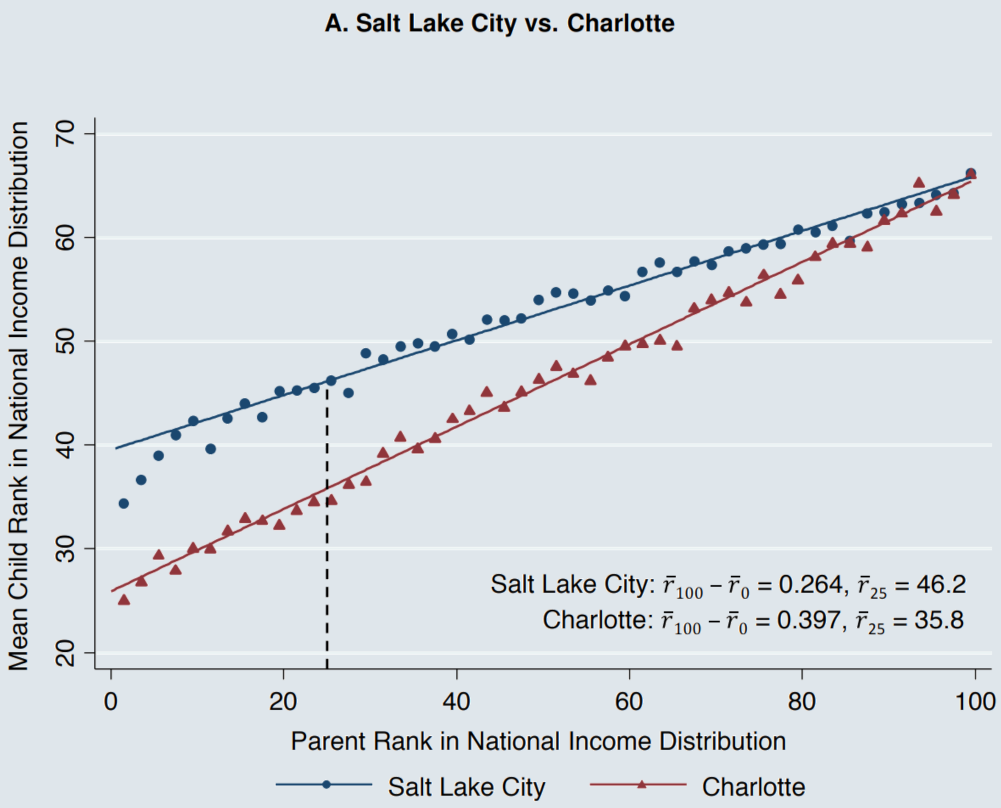
```
##                                       Estimate     Pr(>|t|)
## (Intercept)                         0.11638530 1.575140e-12
## minorityMinority                   -0.02365790 3.180243e-01
## treatmentTreatment                  0.06101349 1.181630e-02
## minorityMinority:treatmentTreatment -0.10670915 2.210982e-03
```

→ Self-selection issue: discriminatory employers did not enter the program

4. Wrap up!

Intergenerational mobility (Chetty et al., 2014)

$$\text{percentile}(y_i^c) = \alpha + \beta_{RRC}\text{percentile}(y_i^p) + \varepsilon_i$$



Relative mobility: $\widehat{\beta_{RRC}}$

Absolute mobility: $\widehat{\alpha} + 25 \times \widehat{\beta_{RRC}}$

- Strong persitence in the United-States
- Large variations across commuting zones
- Intergenerational mobility correlated with characteristics of childhood environment