

# Causality

## Lecture 11

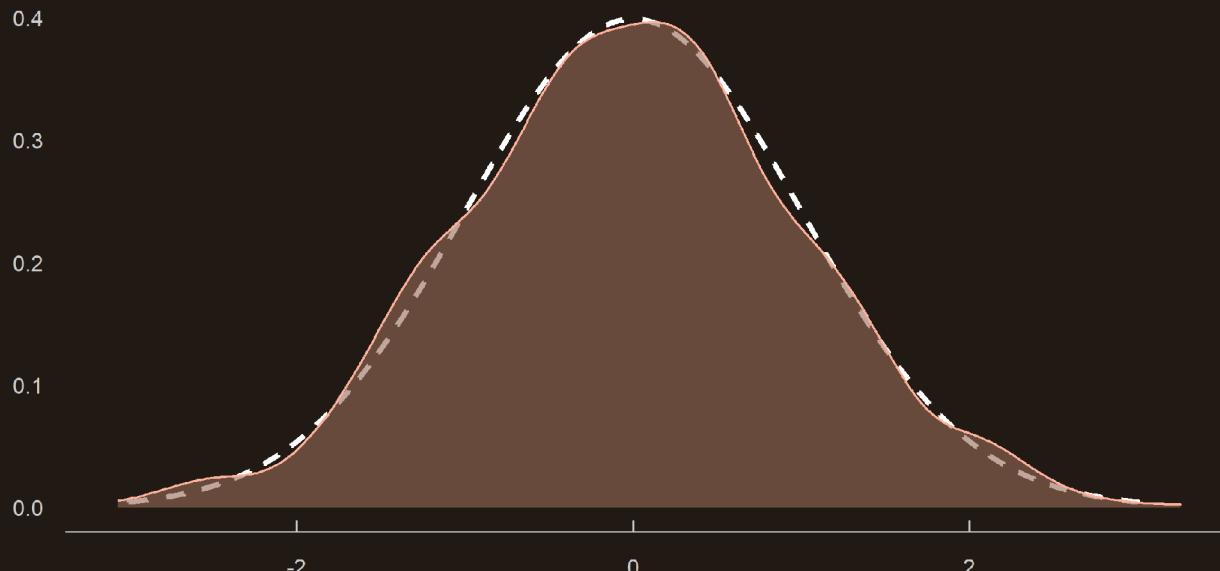
Louis SIRUGUE

CPES 2 - Fall 2022

# Quick reminder

## Data generating process

- In practice we estimate coefficients on a **given realization of a data generating process**
  - So the **true coefficient** is **unobserved**
  - But our **estimation** is **informative** on the values the true coefficient is likely to take



$$\frac{\hat{\beta} - \beta}{\text{SD}(\hat{\beta})} \sim \mathcal{N}(0, 1)$$



# Quick reminder

## Confidence interval

- This allows to infer a **confidence interval**:

$$\hat{\beta} \pm t(\text{df})_{1-\frac{\alpha}{2}} \times \text{se}(\hat{\beta})$$

- Where  $t(\text{df})_{1-\frac{\alpha}{2}}$  is the value from a **Student  $t$  distribution**
  - With the relevant number of **degrees of freedom**  $\text{df}(n - \#\text{parameters})$
  - And the desired **confidence level**  $1 - \alpha$
- And where  $\text{se}(\hat{\beta})$  denotes the **standard error** of  $\hat{\beta}$ :

$$\text{se}(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})} = \sqrt{\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{(n - \#\text{parameters}) \sum_{i=1}^n (x_i - \bar{x})^2}}$$



# Quick reminder

## P-value

- It also allows to **test** how likely is  $\beta$  to be **different from a given value**:
  - If the **p-value** < 5%, we can **reject** that  $\beta$  equals the **hypothesized value** at the 95% confidence level
  - This threshold, very common in Economics, implies that we have 1 chance out of 20 to be wrong

```
linearHypothesis(lm(ige ~ gini, ggcurve), "gini = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## gini = 0
##
## Model 1: restricted model
## Model 2: ige ~ gini
##
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     21 0.46733
## 2     20 0.26883  1     0.1985 14.767 0.001016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Today: Causality

## 1. Main sources of bias

- 1.1. Omitted variables
- 1.2. Functional form
- 1.3. Selection bias
- 1.4. Measurement error
- 1.5. Simultaneity

## 2. Randomized control trials

- 2.1. Introduction to RCTs
- 2.2. Types of randomization
- 2.3. Multiple testing

## 3. Wrap up!



# Today: Causality

## 1. Main sources of bias

- 1.1. Omitted variables
- 1.2. Functional form
- 1.3. Selection bias
- 1.4. Measurement error
- 1.5. Simultaneity



# 1. Main sources of bias

## 1.1. Omitted variable bias

- Consider the following regression:
  - Where  $\text{Earnings}_i$  denotes individuals' annual labor earnings
  - And  $\text{Education}_i$  stands for individuals' number of years of education

$$\text{Earnings}_i = \alpha + \beta \times \text{Education}_i + \varepsilon_i$$

```
summary(lm(Earnings ~ Education, sim_dat))$coefficients
```

```
##             Estimate Std. Error    t value   Pr(>|t|)    
## (Intercept) 7514.800  2994.3060  2.509697 1.209949e-02
## Education    2643.312   205.2692 12.877294 1.220064e-37
```

- Taking  $\hat{\beta}$  at face value, the "**expected returns**" from an additional year of education amount to \$2,643/year
  - But if we were to enforce an additional year of education for randomly selected individuals, would they earn \$2,643 more than they would have earned otherwise?

→ *The answer is no, because the estimated effect is not causal!*



# 1. Main sources of bias

## 1.1. Omitted variable bias

- The estimated relationship could be partly driven by some **confounding factors**:
  - Maybe **more skilled** individuals both **study longer** and **earn more** because they are skilled
  - But with or without more education they would still earn more because they are skilled
- The ability variable acts as a **confounding factor** because it is correlated with both  $x$  and  $y$ 
  - This would also be the case of parental socio-economic status and many other variables
  - We need to put these variables in the regression as **control variables**

$$\text{Earnings}_i = \alpha + \beta_1 \times \text{Education}_i + \beta_2 \times \text{Skills}_i + \varepsilon_i$$

- According to you, would the estimated effect of education be higher or lower in this regression?

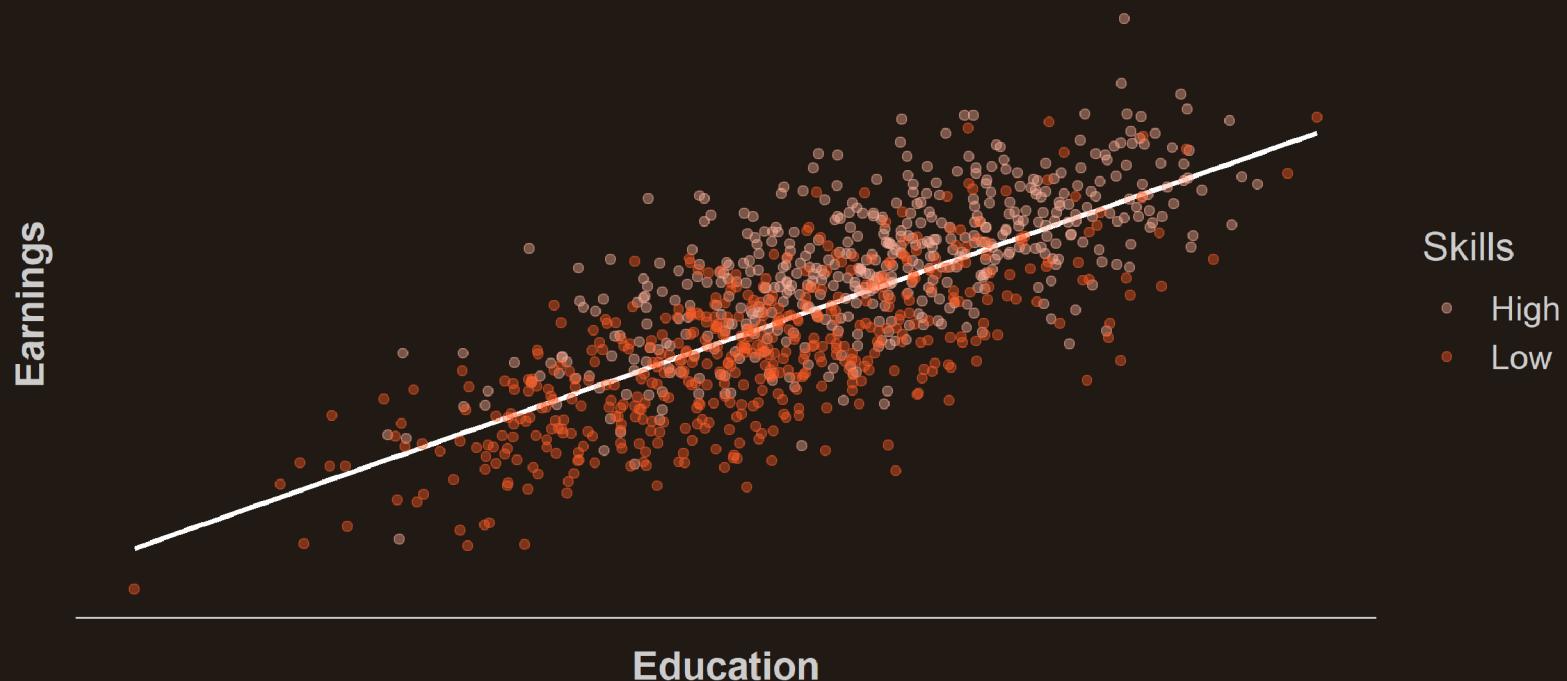
→ If skills is indeed **positively correlated with both** education and earnings, the new coefficient would be **lower**



# 1. Main sources of bias

## 1.1. Omitted variable bias

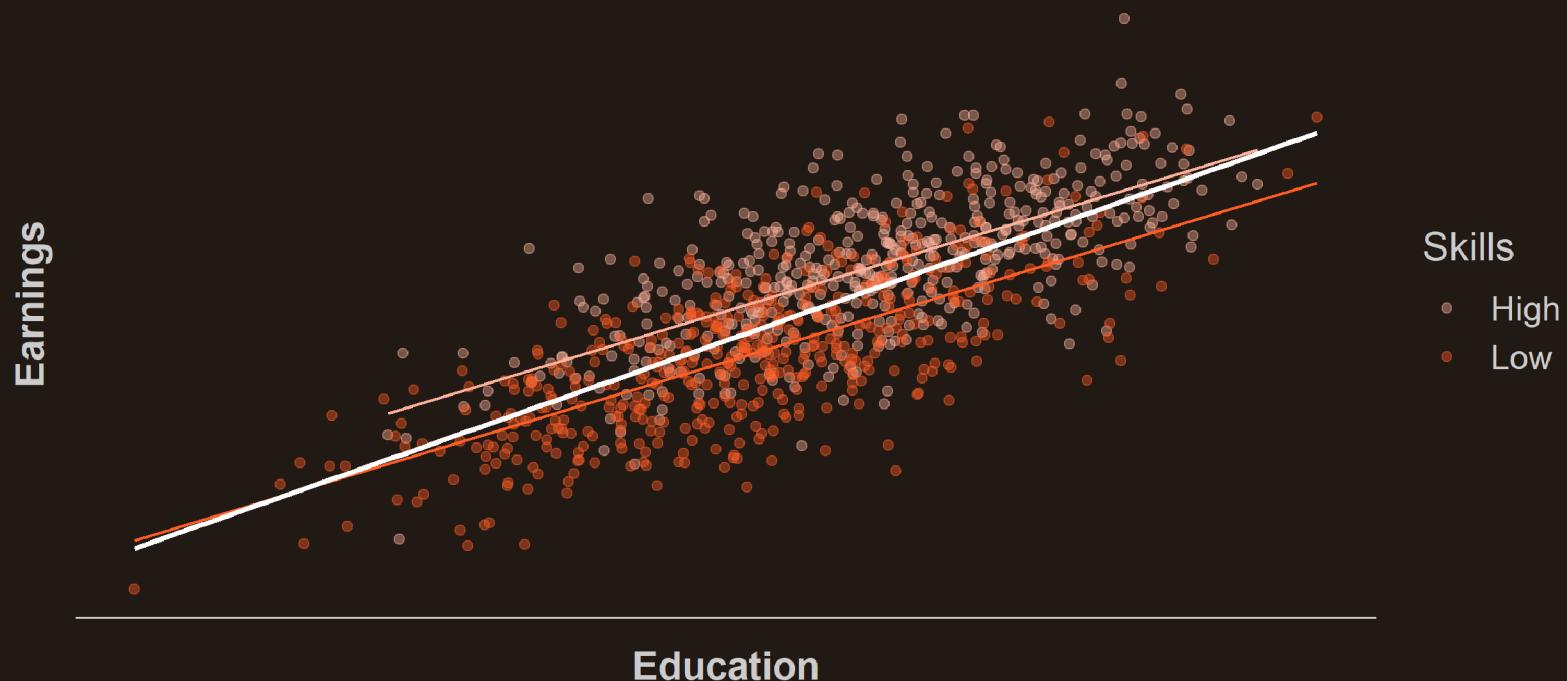
- Remember that **controlling** for a variable can be viewed as:
  - 
  -



# 1. Main sources of bias

## 1.1. Omitted variable bias

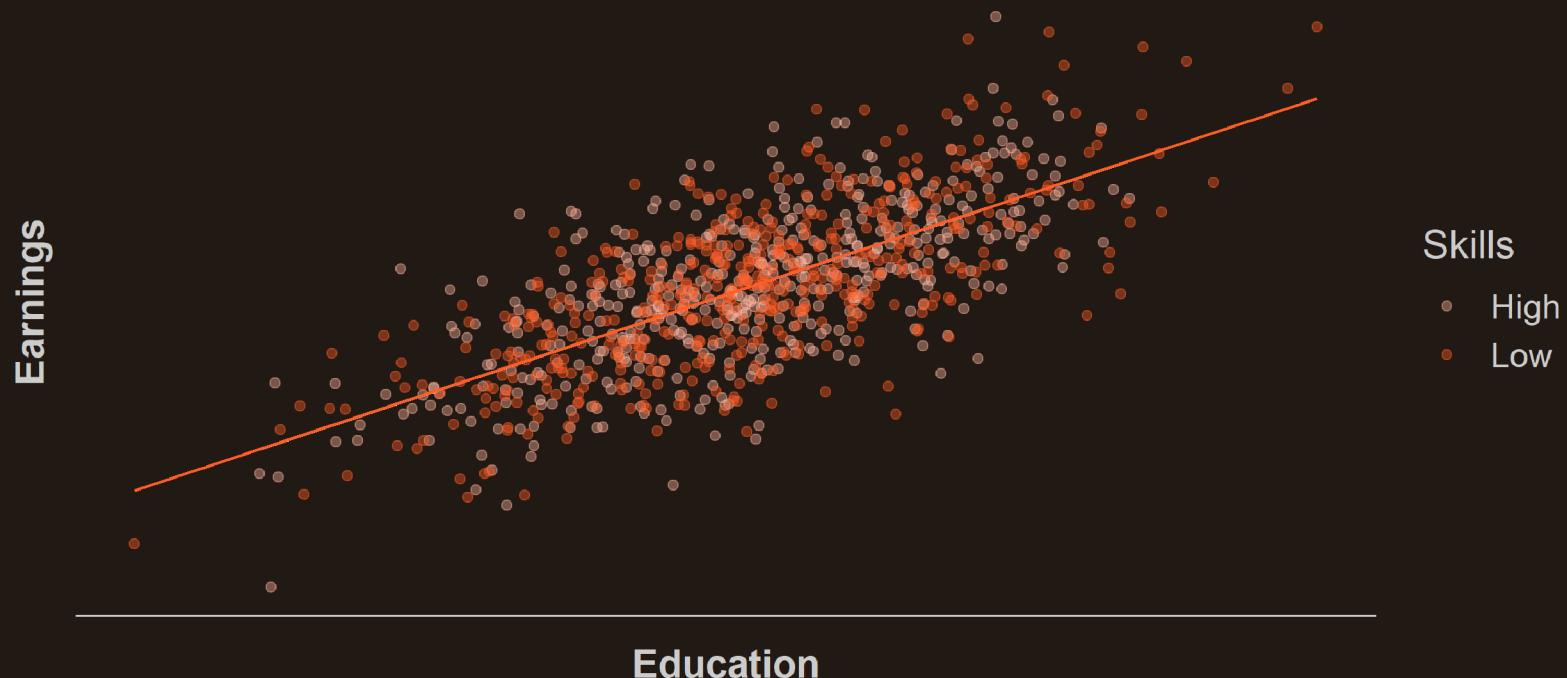
- Remember that **controlling** for a variable can be viewed as:
  - Allowing the **intercept** to **vary** with that variable
  -



# 1. Main sources of bias

## 1.1. Omitted variable bias

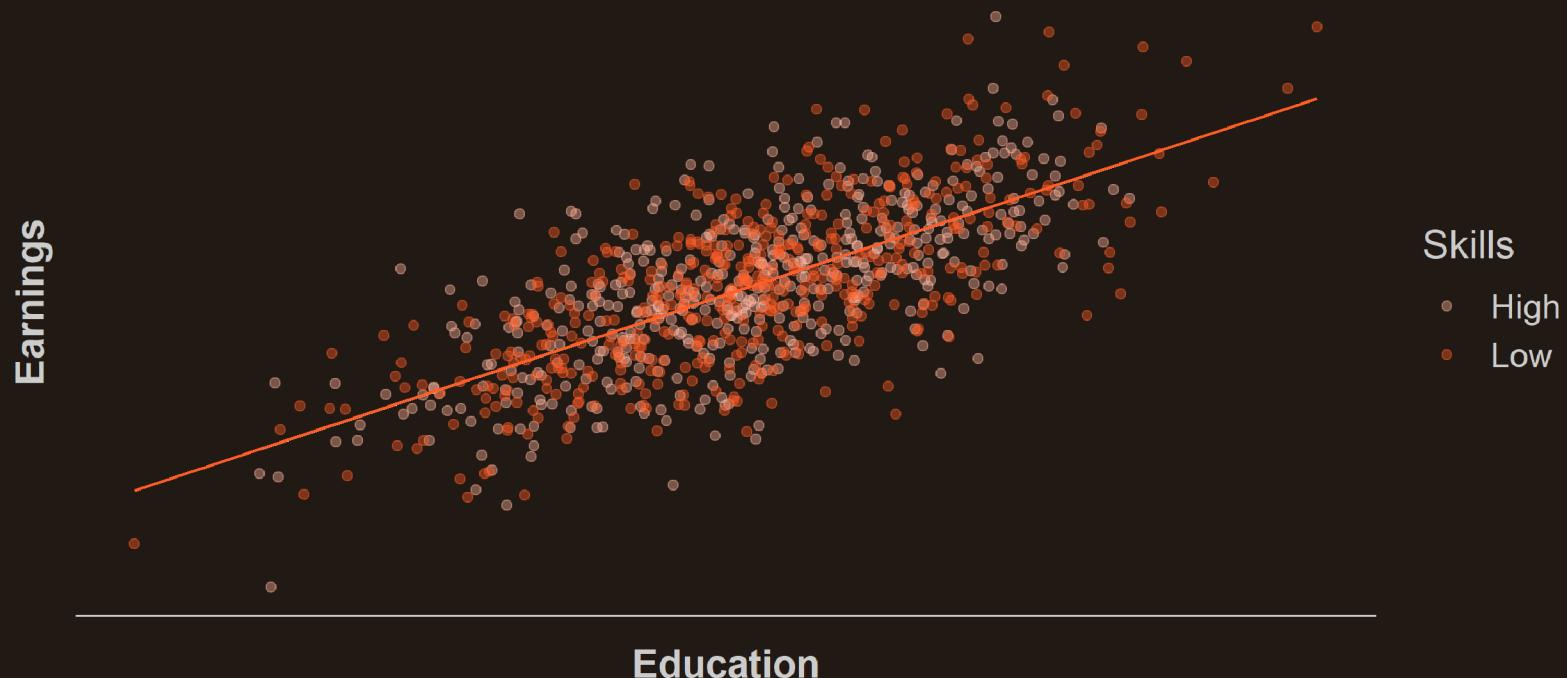
- Remember that **controlling** for a variable can be viewed as:
  - Allowing the **intercept** to **vary** with that variable
  - Keeping this **variable constant** as we move along the  $x$ -axis



# 1. Main sources of bias

## 1.1. Omitted variable bias

- In that case the **confounding** variable **no longer affects** our relationship of interest
  - It fixes the fact that more skilled individuals tend to have both higher education and earnings
  - Such that the **relationship** between education and earnings is **net of the effect of skills**





# 1. Main sources of bias

## 1.1. Omitted variable bias

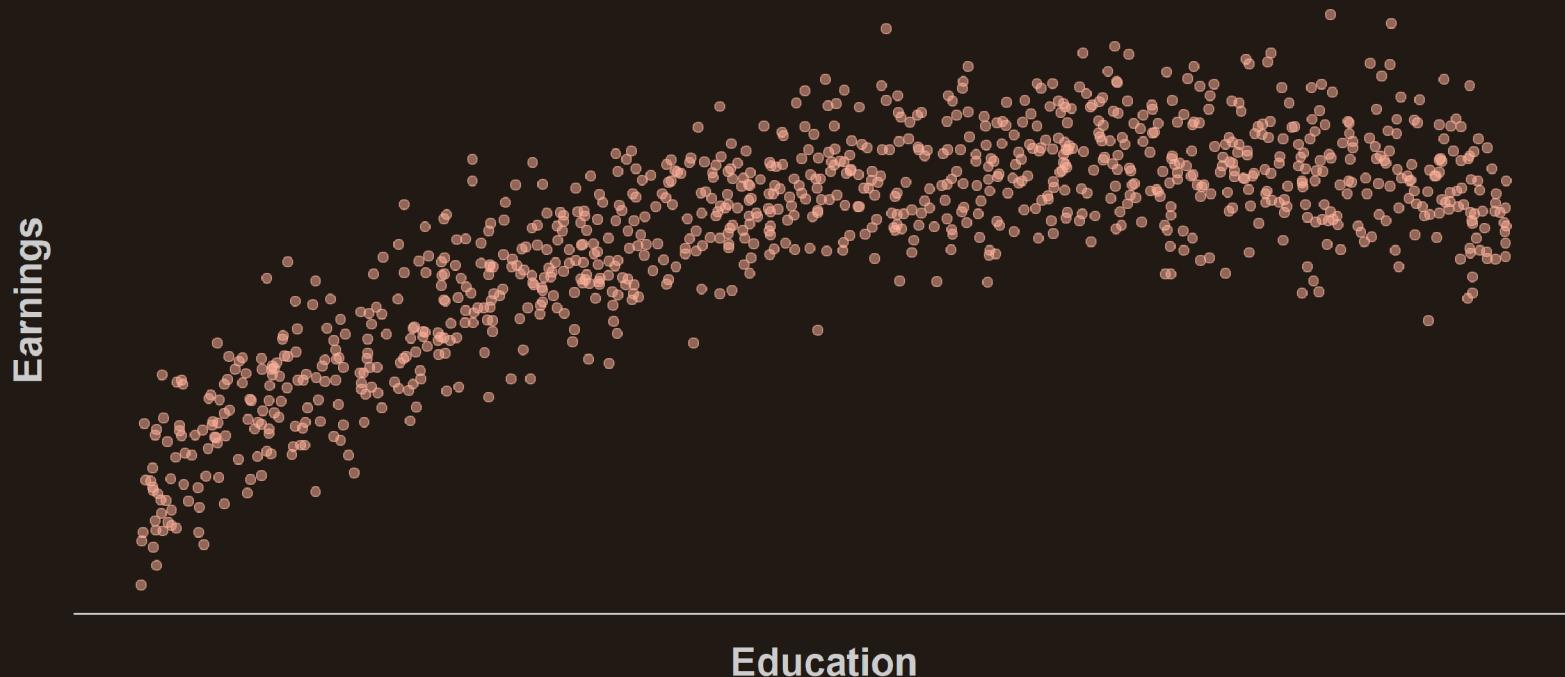
- But **we are never able to control for all** potential confounding factors
  - We can almost always think of variables that may affect both  $x$  and  $y$  but that are not in the data
  - Resulting in what is called the **omitted variable bias**
- In that case you should either:
  - Use causal identification Econometrics techniques (not covered in this course, except RCT)
  - Acknowledge that your estimated effect is not causal with the phrase "**ceteris paribus**"
- *Ceteris paribus* means "**Everything else equal**"
  - We use these sentences to indicate that our **estimation is correct under the hypothesis that** when our  $x$  of interest moves, **no confounding factor** affecting  $y$  moves with it
  - Indeed, if there is no other variable varying with  $x$  and  $y$ , our regression doesn't need more controls
  - We know this assumption is **not correct**, but it is **important to be transparent and clear** about what the coefficient means



# 1. Main sources of bias

## 1.2. Functional form

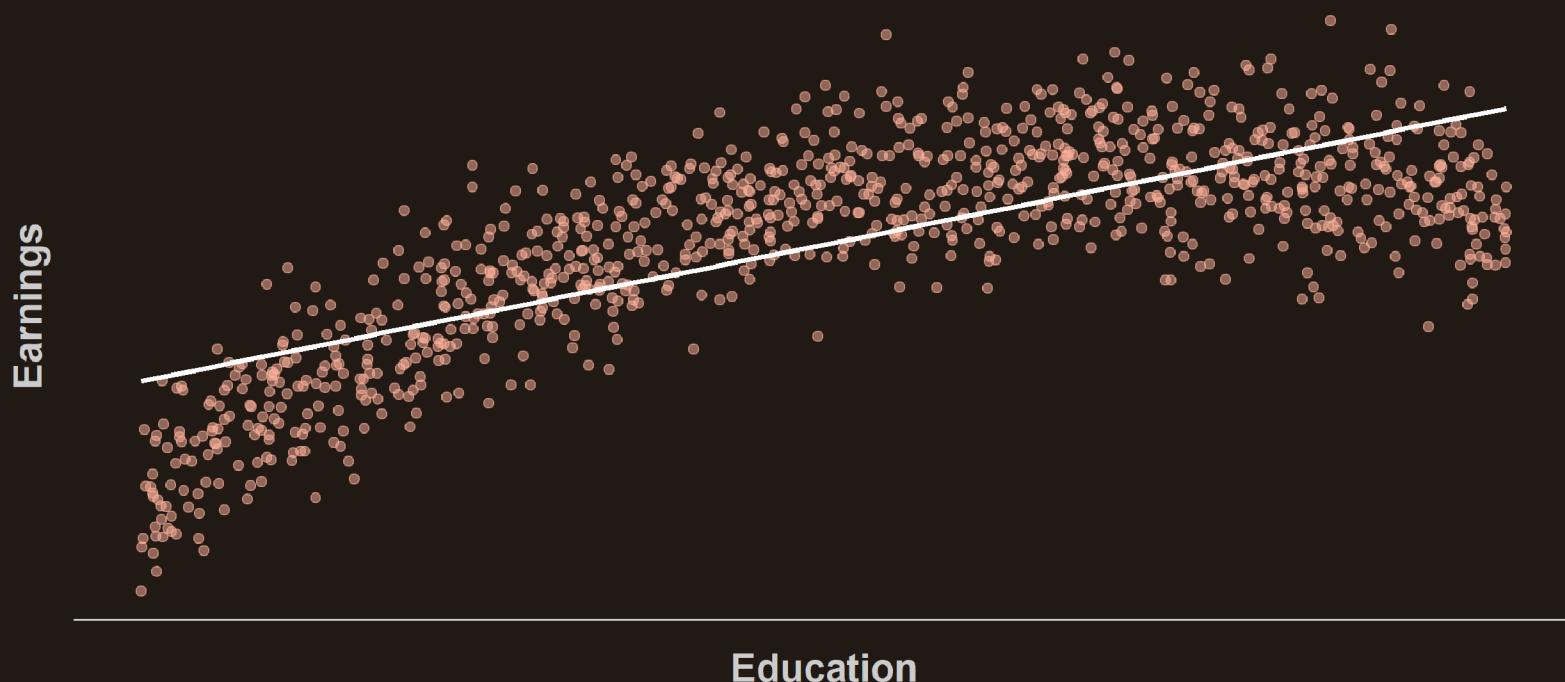
- Now consider the following relationship between years of education and earnings
  - 
  -



# 1. Main sources of bias

## 1.2. Functional form

- Now consider the following relationship between years of education and earnings
  - We can fit a regression line as we usually do
  - But would that be an appropriate estimation?





# 1. Main sources of bias

## 1.2. Functional form

- We must capture the **non-linearity**
  - The relationship cannot be correctly captured by a straight line
  -

$$\text{Earnings}_i = \alpha + \beta_1 \times \text{Education}_i + \varepsilon_i$$



# 1. Main sources of bias

## 1.2. Functional form

- We must capture the **non-linearity**
  - The relationship cannot be correctly captured by a straight line
  - It has the shape of a **polynomial of degree 2**

$$\text{Earnings}_i = \alpha + \beta_1 \times \text{Education}_i + \beta_2 \times \text{Education}_i^2 + \varepsilon_i$$

- Given the previous graph, what would be the signs of  $\hat{\beta}_1$  and  $\hat{\beta}_2$ ?



# 1. Main sources of bias

## 1.2. Functional form

- We must capture the **non-linearity**
  - The relationship cannot be correctly captured by a straight line
  - It has the shape of a **polynomial of degree 2**

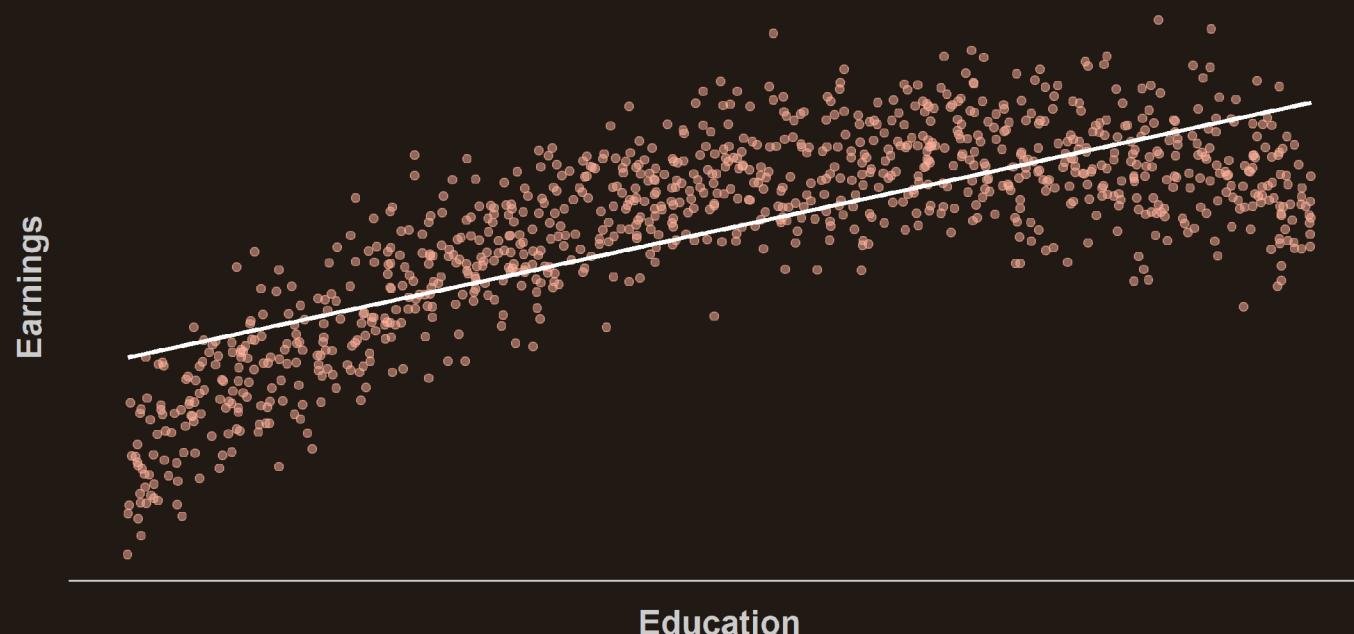
$$\text{Earnings}_i = \alpha + \beta_1 \times \text{Education}_i + \beta_2 \times \text{Education}_i^2 + \varepsilon_i$$

- Given the previous graph, what would be the signs of  $\hat{\beta}_1$  and  $\hat{\beta}_2$ ?
  - $\hat{\beta}_1$  would be positive because the relationship is increasing
  - $\hat{\beta}_2$  would be negative because the relationship is concave
- Polynomial functional forms are easy to handle in R
  - You can **square the dependent variable and add it** in lm()
  - geom\_smooth() also allows to plot a polynomial fit

# 1. Main sources of bias

## 1.2. Functional form

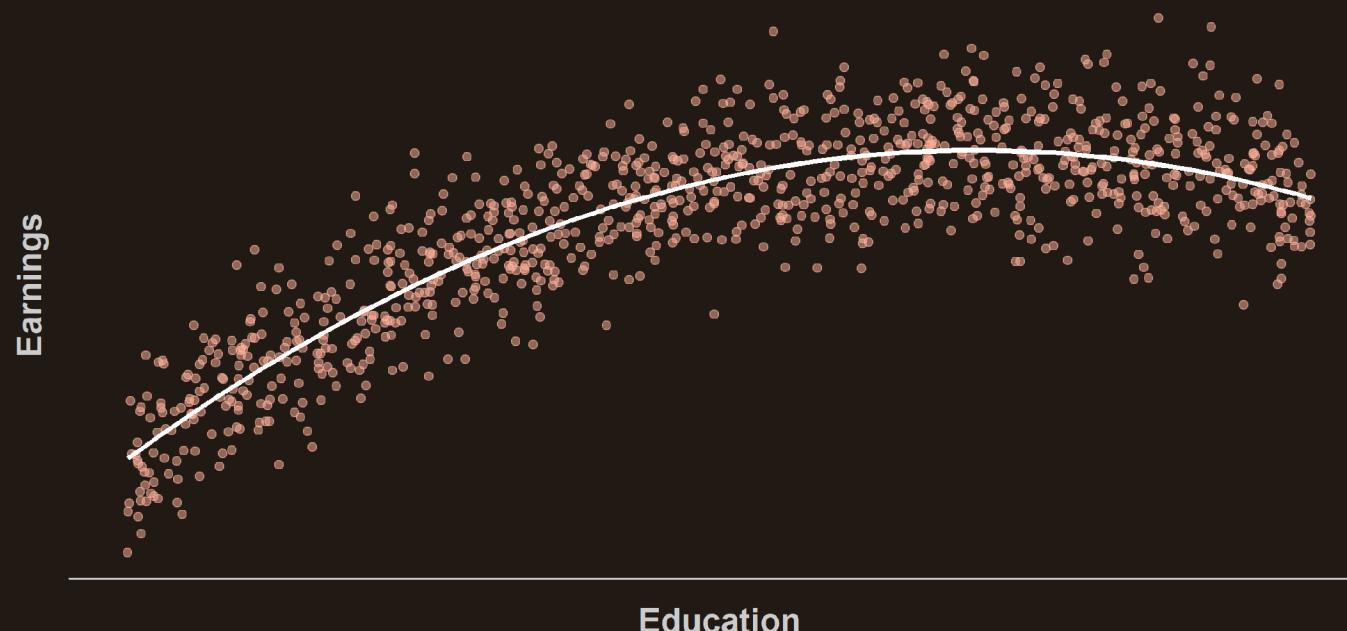
```
ggplot(quadratic, aes(x = Education, y = Earnings)) + geom_point() +  
  geom_smooth(method = "lm")
```



# 1. Main sources of bias

## 1.2. Functional form

```
ggplot(quadratic, aes(x = Education, y = Earnings)) + geom_point() +  
  geom_smooth(method = "lm", formula = y ~ poly(x, 2))
```

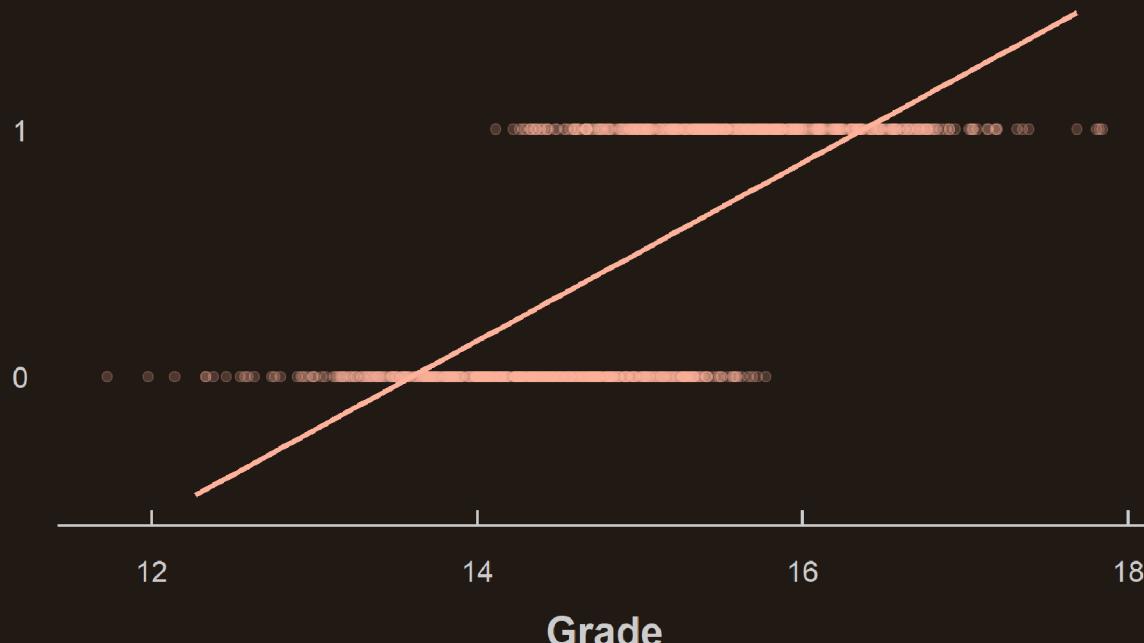


- But functional form is not only about polynomial degrees:
  - Interactions
  - Logs
  - Discretization
  - ...

# 1. Main sources of bias

## 1.3. Selection bias

- Now remember the example on high-school grades and job application acceptance
  - We plotted the **grades** of individuals on the  $x$ -axis
  - And **whether or not they got the job** on the  $y$ -axis

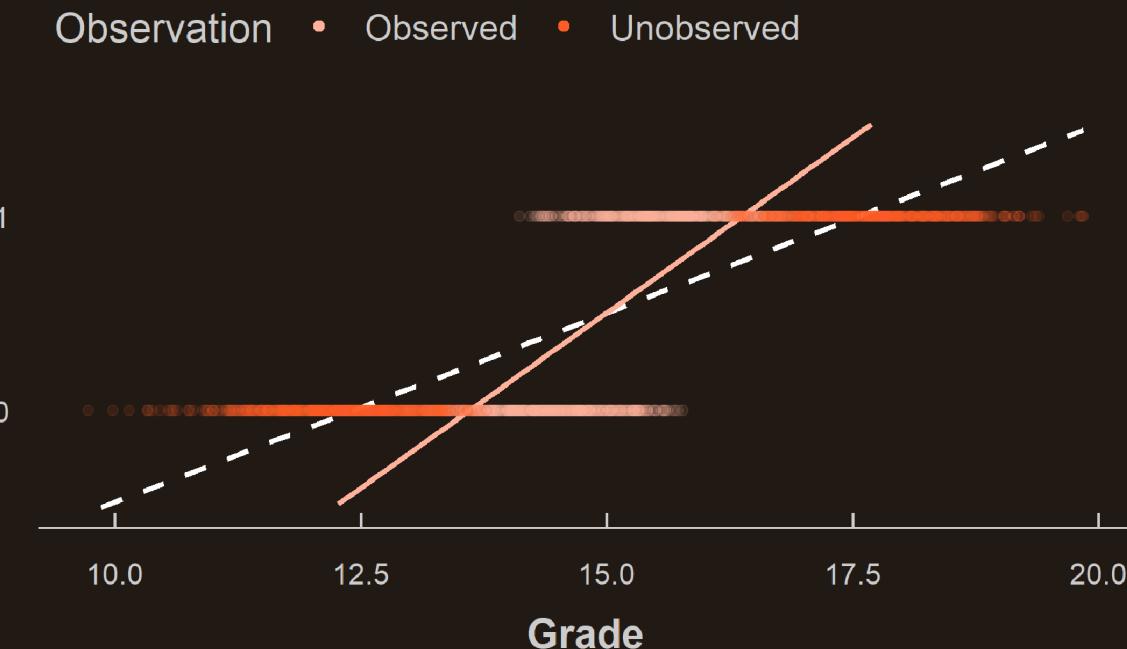


- We estimated that a **1 unit** increase in Grade (/20) would **increase the probability** to be accepted by about **a third** on expectation, **ceteris paribus**
- Is this estimation relevant?
  - Look at the support of  $x$

# 1. Main sources of bias

## 1.3. Selection bias

- The fact that almost all grades range between 13 and 17 hints at a **selection problem**:
  - Individuals with very **low grade won't apply** to the position because **they know they will be rejected**
  - Individuals with very **high grade won't apply** to the position because **they apply to better positions**



- Had these individuals applied, the estimated effect would be lower
- Our coefficient is specific to a non-representative sample
  - Issue of **external validity**
  - The interpretation only holds in our specific setting



# 1. Main sources of bias

## 1.3. Selection bias

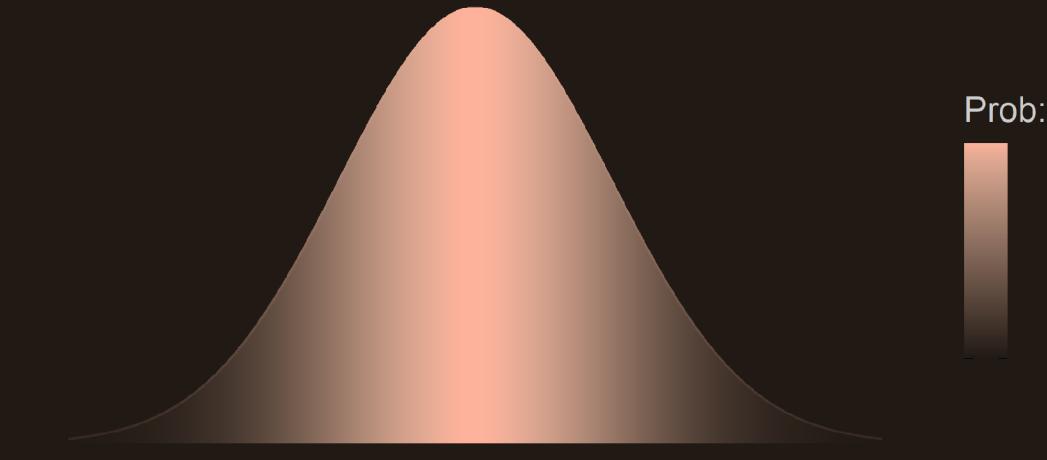
- Such **selection problems** are very common **threats to causality**
- What is the impact of going to a better neighborhood on your children outcomes?
  - Those who move may be different from those who stay: **self-selection issue**
  - Here it is not that the sample is not representative of the population, but that **the outcomes of those who stayed are different from the outcomes those who moved would have had, if they had stayed**
- This related to the notion of **counterfactual**
  - If those who moved were comparable to those who stayed, it would be valid to use the outcome of those who stayed as the counterfactual outcome of those who moved
  - But because of selection movers are not comparable to stayers so we don't have a credible counterfactual
- The notion of counterfactual is key to answer many questions:
  - What is the impact of an immigrant inflow on the labor market outcomes of locals?
  - We need to know how the labor market outcomes of locals would have evolved absent the immigrant inflow but we do not observe this situation



# 1. Main sources of bias

## 1.4. Measurement error

- Another way of obtaining **biased estimates** is to have a **dependent variable measured with errors**
  - For instance if you want to measure the effect of cognitive skills but you only have IQ scores
  - IQ is a noisy measure of cognitive skills as individuals' performances to such test are not always consistent
- It seems reasonable to assume that the measurement error follows a normal distribution:
  - 
  -

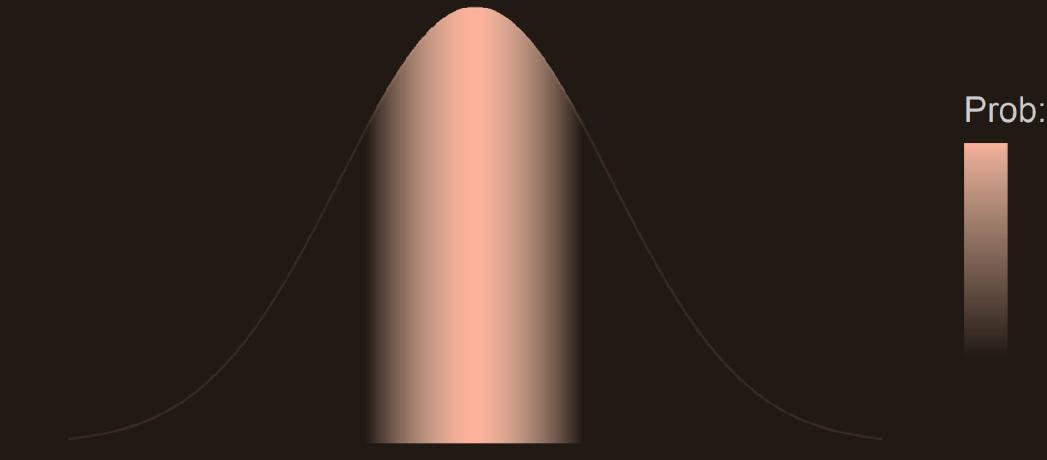




# 1. Main sources of bias

## 1.4. Measurement error

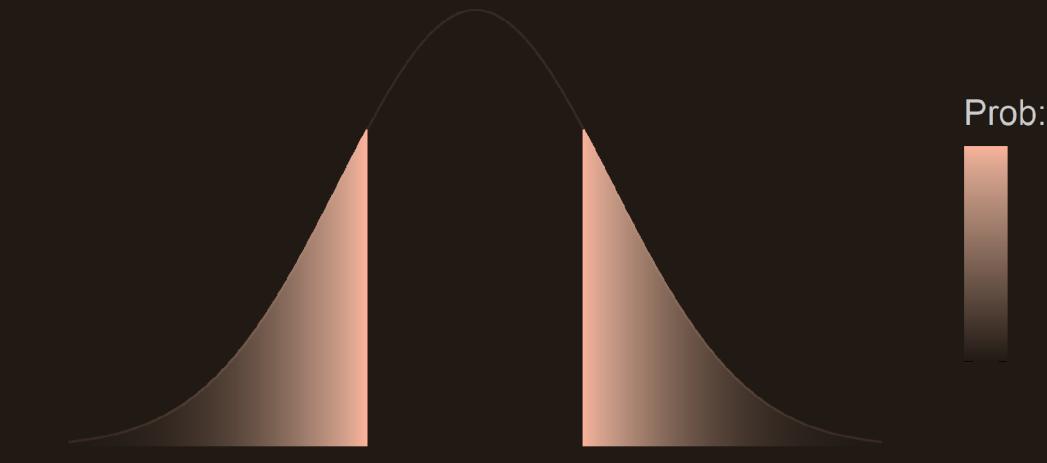
- Another way of obtaining **biased estimates** is to have a **dependent variable measured with errors**
  - For instance if you want to measure the effect of cognitive skills but you only have IQ scores
  - IQ is a noisy measure of cognitive skills as individuals' performances to such test are not always consistent
- It seems reasonable to assume that the measurement error follows a normal distribution:
  - Individuals **usually** perform **close to their average** performance
  -



# 1. Main sources of bias

## 1.4. Measurement error

- Another way of obtaining **biased estimates** is to have a **dependent variable measured with errors**
  - For instance if you want to measure the effect of cognitive skills but you only have IQ scores
  - IQ is a noisy measure of cognitive skills as individuals' performances to such test are not always consistent
- It seems reasonable to assume that the measurement error follows a normal distribution:
  - Individuals **usually** perform **close to their average** performance
  - And **larger deviations** are more **rare**





# 1. Main sources of bias

## 1.4. Measurement error

Denote  $x$  the IQ variable

$$x \sim \mathcal{N}(100, 15^2)$$

Denote  $\eta$  the measurement error

$$\eta \sim \mathcal{N}(0, 1)$$

- The true relationship is

$$y = \alpha + \beta x + \varepsilon$$

- But we only observe

$$\tilde{x} = x + \eta$$

- So we can only estimate:

$$y = \alpha + \beta \tilde{x} + \varepsilon \iff y = \alpha + \beta(x + \eta) + \varepsilon$$

→ Let's use simulations to see how it may affect our estimation



# 1. Main sources of bias

## 1.4. Measurement error

- We can start by **generating a relationship** without measurement error

$$y_i = 1 + 2x_i + \varepsilon_i, \text{ with } \varepsilon \sim \mathcal{N}(0, 1)$$

```
dat <- tibble(x = rnorm(1000, 100, 15),
               y = 1 + (2 * x) + rnorm(1000, 0, 1))
```

- Estimate the **unbiased** relationship

```
lm(y ~ x, dat)$coefficient
```

```
## (Intercept)          x
##     0.824755    2.001394
```

Is it just random chance or is  $\hat{\beta}$  downward biased? →

- And **with measurement error**  $\eta \sim \mathcal{N}(0, 1)$

```
dat <- dat %>%
  mutate(noisy_x = x + rnorm(1000, 0, 1))

lm(y ~ noisy_x, dat)$coefficient
```

```
## (Intercept)      noisy_x
##     1.995596    1.990358
```



## 1. Main sources of bias

## 1.4. Measurement error

- Let's have a look at how  $\hat{\beta}$  behaves with an increasingly high  $SD(\eta)$



## 1. Main sources of bias

## 1.4. Measurement error

- Let's have a look at how  $\hat{\beta}$  behaves with an increasingly high  $SD(\eta)$



# 1. Main sources of bias

## 1.4. Measurement error

- Let's have a look at how  $\hat{\beta}$  behaves with an increasingly high  $SD(\eta)$

```
# Vector of standard deviations from 0 to 20
sd_noise <- 0:20

# Empty vector for beta...
beta <- c()

# ... to be filled in a loop
for (i in sd_noise) {

  #
  #

  #
  #

  #
  #

}

}
```



# 1. Main sources of bias

## 1.4. Measurement error

- Let's have a look at how  $\hat{\beta}$  behaves with an increasingly high  $SD(\eta)$

```
# Vector of standard deviations from 0 to 20
sd_noise <- 0:20

# Empty vector for beta...
beta <- c()

# ... to be filled in a loop
for (i in sd_noise) {

  # Generate noisy x with corresponding SD(eta)
  dat_i <- dat %>% mutate(noisy_x = x + rnorm(1000, 0, i))

  #
  #

}

}
```



# 1. Main sources of bias

## 1.4. Measurement error

- Let's have a look at how  $\hat{\beta}$  behaves with an increasingly high  $SD(\eta)$

```
# Vector of standard deviations from 0 to 20
sd_noise <- 0:20

# Empty vector for beta...
beta <- c()

# ... to be filled in a loop
for (i in sd_noise) {

  # Generate noisy x with corresponding SD(eta)
  dat_i <- dat %>% mutate(noisy_x = x + rnorm(1000, 0, i))

  # Estimate the regression
  beta_i <- lm(y ~ noisy_x, dat_i)$coefficient[2]

}

}
```



# 1. Main sources of bias

## 1.4. Measurement error

- Let's have a look at how  $\hat{\beta}$  behaves with an increasingly high  $SD(\eta)$

```
# Vector of standard deviations from 0 to 20
sd_noise <- 0:20

# Empty vector for beta...
beta <- c()

# ... to be filled in a loop
for (i in sd_noise) {

  # Generate noisy x with corresponding SD(eta)
  dat_i <- dat %>% mutate(noisy_x = x + rnorm(1000, 0, i))

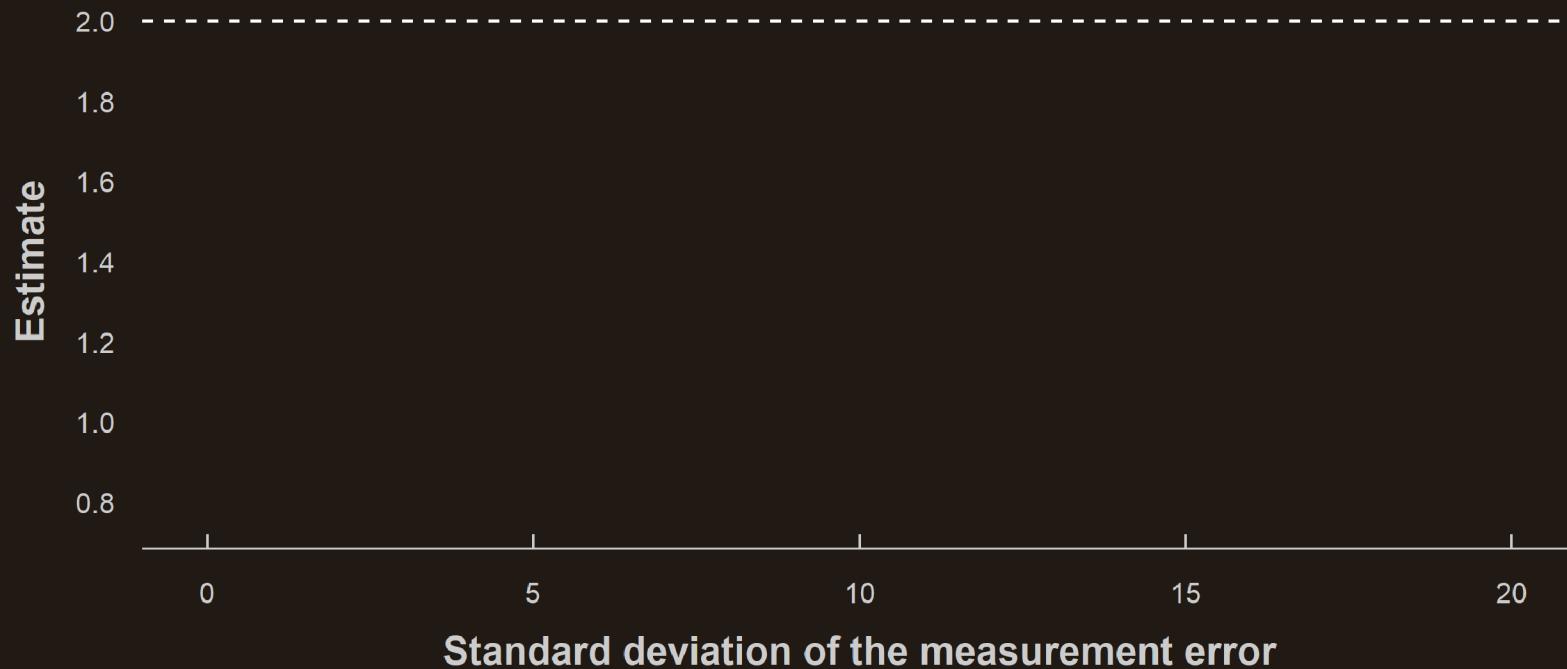
  # Estimate the regression
  beta_i <- lm(y ~ noisy_x, dat_i)$coefficient[2]

  # Store the coefficient
  beta <- c(beta, beta_i)
}
```

# 1. Main sources of bias

## 1.4. Measurement error

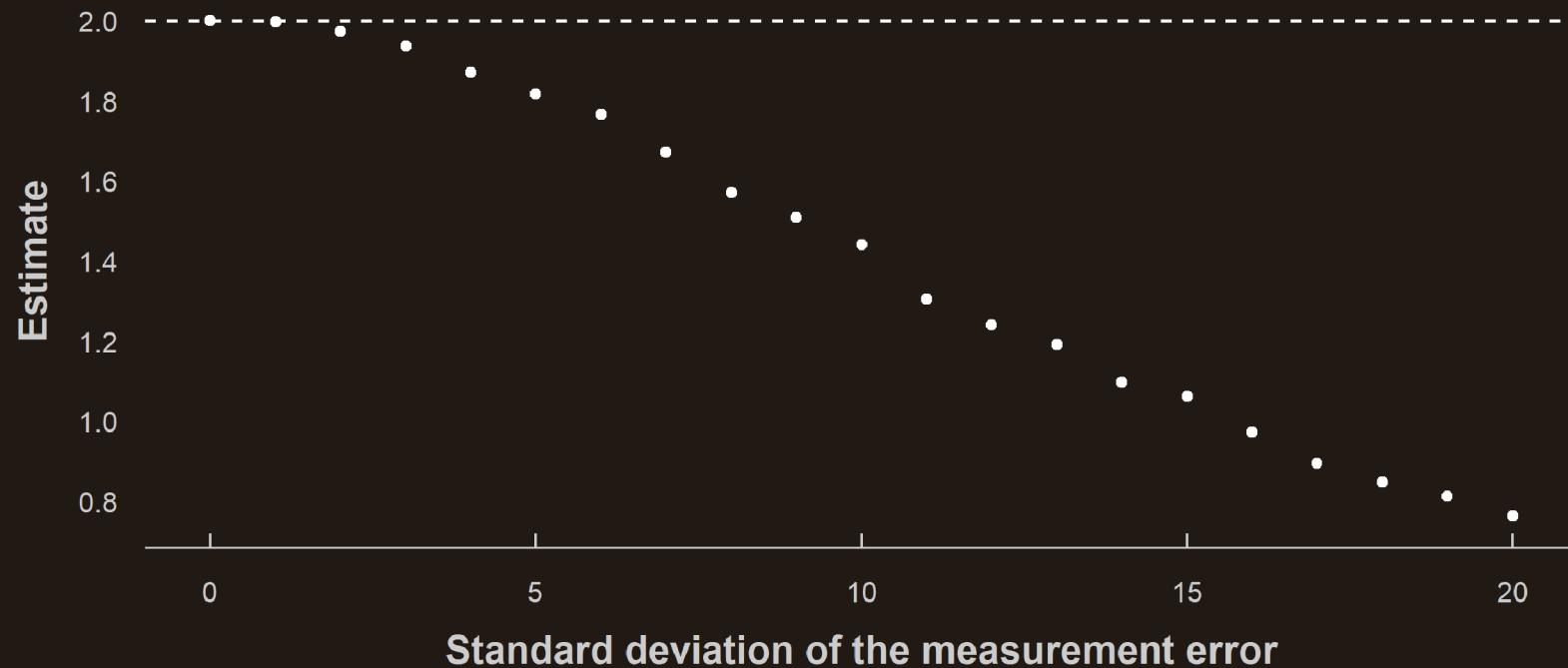
- We can then plot the  $\hat{\beta}$  for each value of  $SD(\eta)$ 
  - 
  -



# 1. Main sources of bias

## 1.4. Measurement error

- We can then plot the  $\hat{\beta}$  for each value of  $SD(\eta)$ 
  - It is clear that the **measurement error** puts a **downward pressure** on our estimate
  - And that the **noisier** the measure of  $x$  the **larger** the **bias**





# 1. Main sources of bias

## 1.4. Measurement error

- And this phenomenon can easily be shown **mathematically**:
  - 
  -

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

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

$$\hat{\beta} = \frac{\text{Cov}(x, y) + \text{Cov}(\eta, y)}{\text{Var}(x) + \text{Var}(\eta) + 2\text{Cov}(x, \eta)}$$

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



# 1. Main sources of bias

## 1.4. Measurement error

- And this phenomenon can easily be shown **mathematically**:
  - The extra term in the denominator puts a **downward pressure** on our estimate
  - And the bias is **increasing in the amplitude of the measurement error**

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

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

$$\hat{\beta} = \frac{\text{Cov}(x, y) + \text{Cov}(\eta, y)}{\text{Var}(x) + \text{Var}(\eta) + 2\text{Cov}(x, \eta)}$$

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



# 1. Main sources of bias

## 1.5. Simultaneity

- **So far** we considered relationships whose **directions** were quite **unambiguous**
  - Education → Earnings, and not the opposite
  - High-school grades → Job acceptance, and not the opposite

*But now consider the relationship between **crime rate** and **police coverage intensity***

- **What is the direction** of the relationship?
  - It's likely that more crime would cause a positive response in police activity
  - But also that police activity would deter crime
- There is no easily solution to that problem apart from:
  - Working out a **theoretical model** sorting this issue beforehand
  - Or **designing an RCT** that cuts one of the two channels



# Overview: Causality

## 1. Main sources of bias ✓

- 1.1. Omitted variables
- 1.2. Functional form
- 1.3. Selection bias
- 1.4. Measurement error
- 1.5. Simultaneity

## 2. Randomized control trials

- 2.1. Introduction to RCTs
- 2.2. Types of randomization
- 2.3. Multiple testing

## 3. Wrap up!



# Overview: Causality

## 1. Main sources of bias ✓

- 1.1. Omitted variables
- 1.2. Functional form
- 1.3. Selection bias
- 1.4. Measurement error
- 1.5. Simultaneity

## 2. Randomized control trials

- 2.1. Introduction to RCTs
- 2.2. Types of randomization
- 2.3. Multiple testing



## 2. Randomized control trials

### 2.1. Introduction to RCTs

- 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
- RCTs are very powerful tools to **sort out issues of:**
  - Omitted variables
  - Selection bias
  - Simultaneity
- This method is particularly used to **identify causal relationships** in:
  - Medicine
  - Psychology
  - Economics
  - ...

*But how does randomness help obtaining causality?*



## 2. Randomized control trials

### 2.1. Introduction to RCTs

- Consider estimating the **effect of vitamin** supplements intake **on health**
  - Comparing health outcomes of vitamin **consumers vs. non-consumers**, the effect **won't be causal**
  - Vitamins consumers might be **richer** and **more healthy in general** for other reasons than vitamin intake
- **Randomization** allows to **solve** this selection **bias**
  - If you take two groups randomly, they would have the **same characteristics** on expectation
  - And thus they would be perfectly **comparable**

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

```
asec_2020 %>%
  summarise(Earnings = mean(Earnings), Hours = mean(Hours),
            Black = mean(Race == "Black"), Asian = mean(Race == "Asian"),
            Other = mean(Race == "Other"), Female = mean(Sex == "Female"))
```

```
##   Earnings     Hours      Black      Asian      Other     Female
## 1  62132.37 39.54742 0.1062391 0.0703805 0.03764611 0.4809749
```



## 2. Randomized control trials

### 2.1. Introduction to RCTs

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

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

```
## # A tibble: 2 x 8
##   Group      n Earnings Female Black Asian Other Hours
##   <chr>     <int>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1 Control    32195    62234.    48.2   10.7   7.02  3.80  39.5
## 2 Treatment  32141    62030.    48.0   10.5   7.05  3.73  39.6
```



## 2. Randomized control trials

### 2.1. Introduction to RCTs

- 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 there 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. Randomized control trials

### 2.2. Types of randomization

- To some extent there are ways to deal with these problems
  - Notably we can **adjust the way the treatment is randomized**
- 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**
  - We don't give the treatment randomly hoping that we'll 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. Randomized control trials

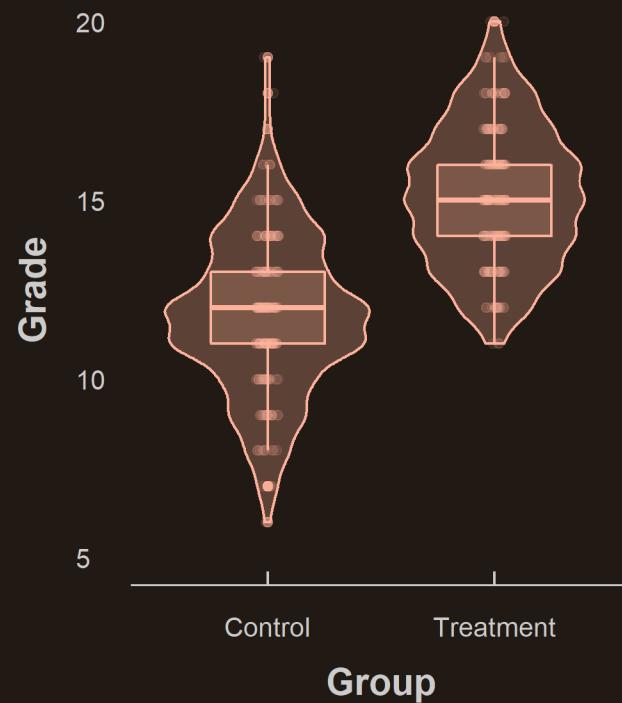
### 2.2. Types of randomization

- What if you want to estimate the impact of **calorie intake** at the **10am break** on **pupils grades**
  1. Find a school to run your experiment
  2. Take the list of pupils and randomly allocate them to treatment and control group
  3. Provide families with treated pupils a snack for the 10am break every school day
  4. Do that for a few month and collect the data on the grades of both groups
  5. Compute the difference in average grade for the treated and the control group
- If the 10am snack has a **positive effect**:
  - This causal identification framework should ensure the correct estimation of that effect
  - Right?
- But what about **non-compliance**?
  - It is likely that during the 10am break, treated children share their snack with their untreated friends
  - How would that **affect our estimation**?

## 2. Randomized control trials

### 2.2. Types of randomization

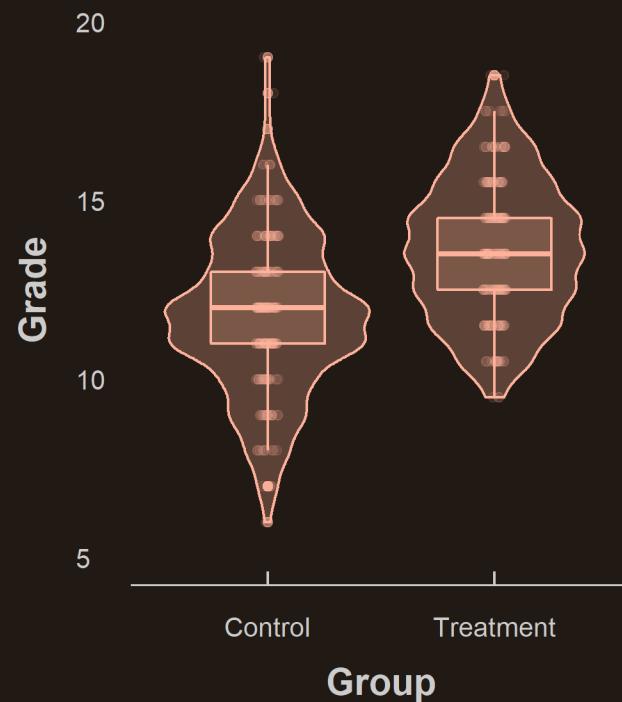
- While the observed effect would be positive under full compliance, **under treatment sharing:**
  - 
  -



## 2. Randomized control trials

### 2.2. Types of randomization

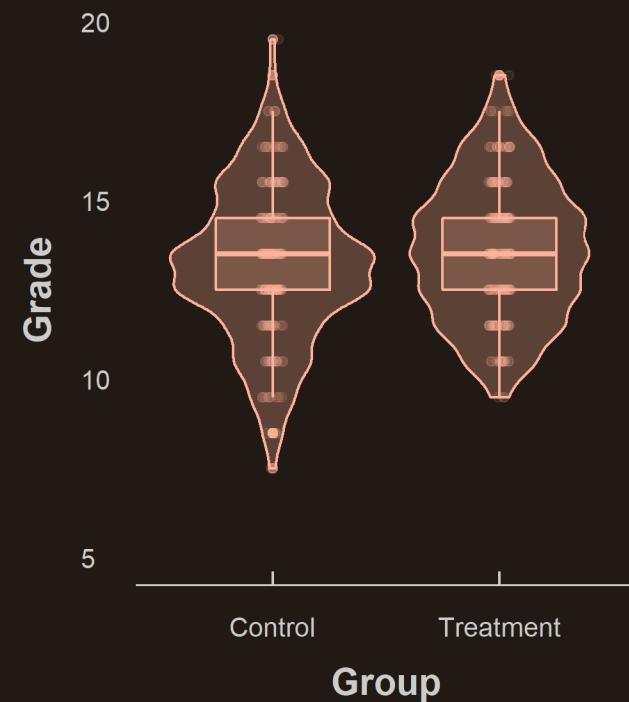
- While the observed effect would be positive under full compliance, **under treatment sharing:**
  - **Treated children** would have **lower grades** because they would benefit from less calories
  -



## 2. Randomized control trials

### 2.2. Types of randomization

- While the observed effect would be positive under full compliance, **under treatment sharing:**
  - **Treated children** would have **lower grades** because they would benefit from less calories
  - **Untreated children** would have **higher grades** because they would benefit from less calories





## 2. Randomized control trials

### 2.2. Types of randomization

- Thus **non-compliance** can bias our estimation
  - There would be a **downward bias**
  - And our estimation **wouldn't be causal**
- One solution to that problem is to **randomize by cluster**
  - Children cannot share their snack with children from other schools
- We must **treat at the school level** instead of the child 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. Randomized control trials

### 2.3. Multiple testing

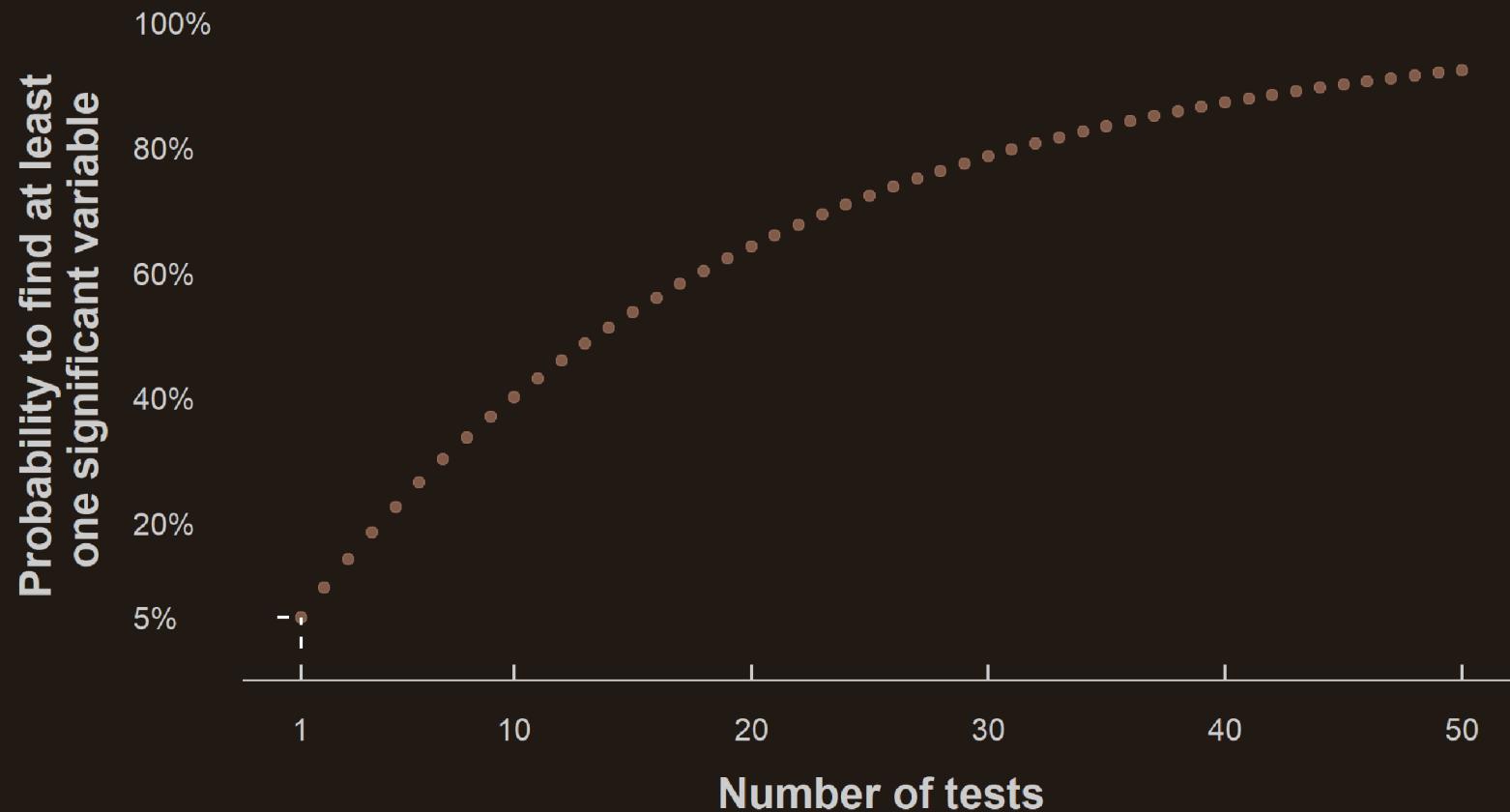
- Another inference issue that RCTs can be subject to is **multiple testing**
  - If you conduct an RCT you might be tempted to exploit the causal framework to test a myriad of effects
- You randomize your treatment and you compare the averages of many outcomes between treated and untreated individuals
  - You would be tempted to **conclude** that there is a **significant effect** for **every variable** whose corresponding **p-value < .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** for a null true effect among these multiple tests is **greater than 5%**
  - The greater the number of tests, the higher the probability to get a significant result just by chance

This is what we call *multiple testing*



## 2. Randomized control trials

### 2.3. Multiple testing





## 2. Randomized control trials

### 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: Causality

## 1. Main sources of bias ✓

- 1.1. Omitted variables
- 1.2. Functional form
- 1.3. Selection bias
- 1.4. Measurement error
- 1.5. Simultaneity

## 2. Randomized control trials ✓

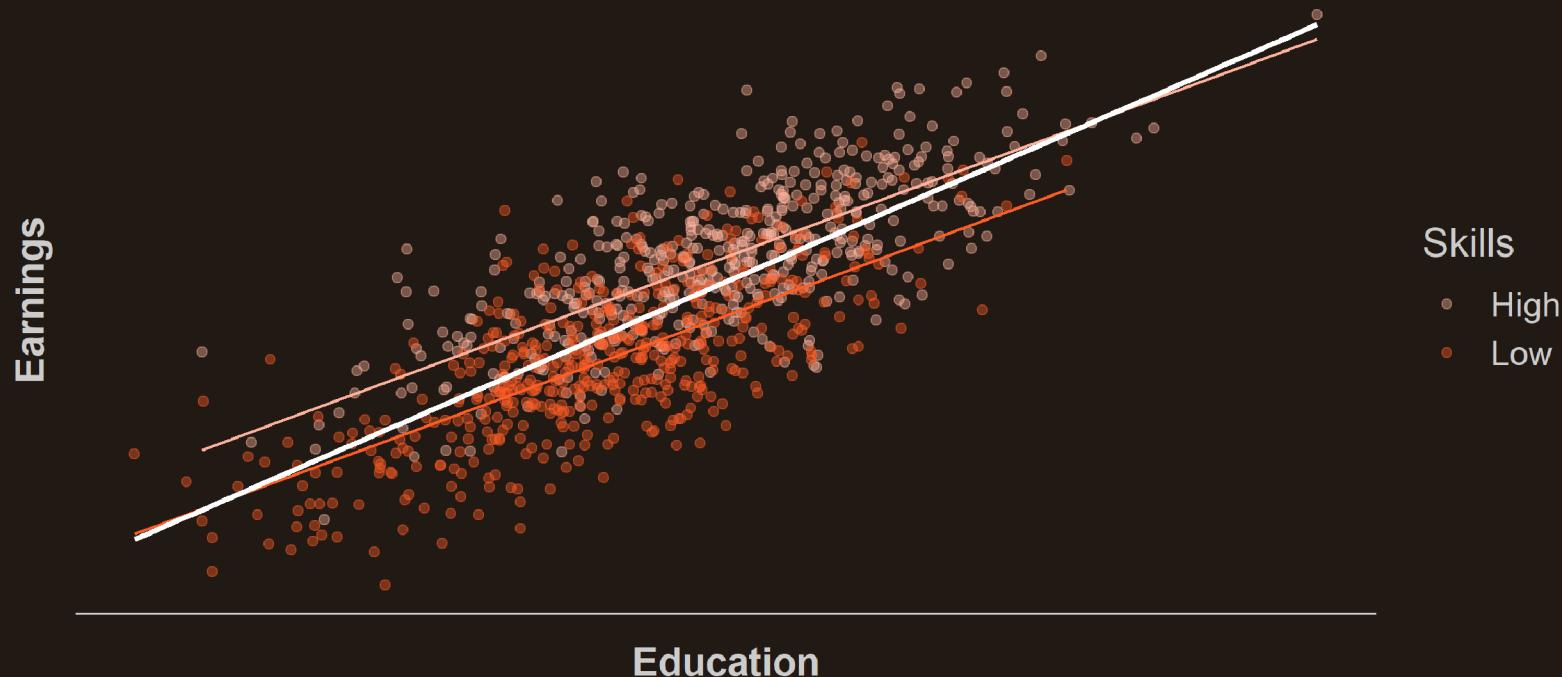
- 2.1. Introduction to RCTs
- 2.2. Types of randomization
- 2.3. Multiple testing

## 3. Wrap up!

### 3. Wrap up!

#### Omitted variable bias

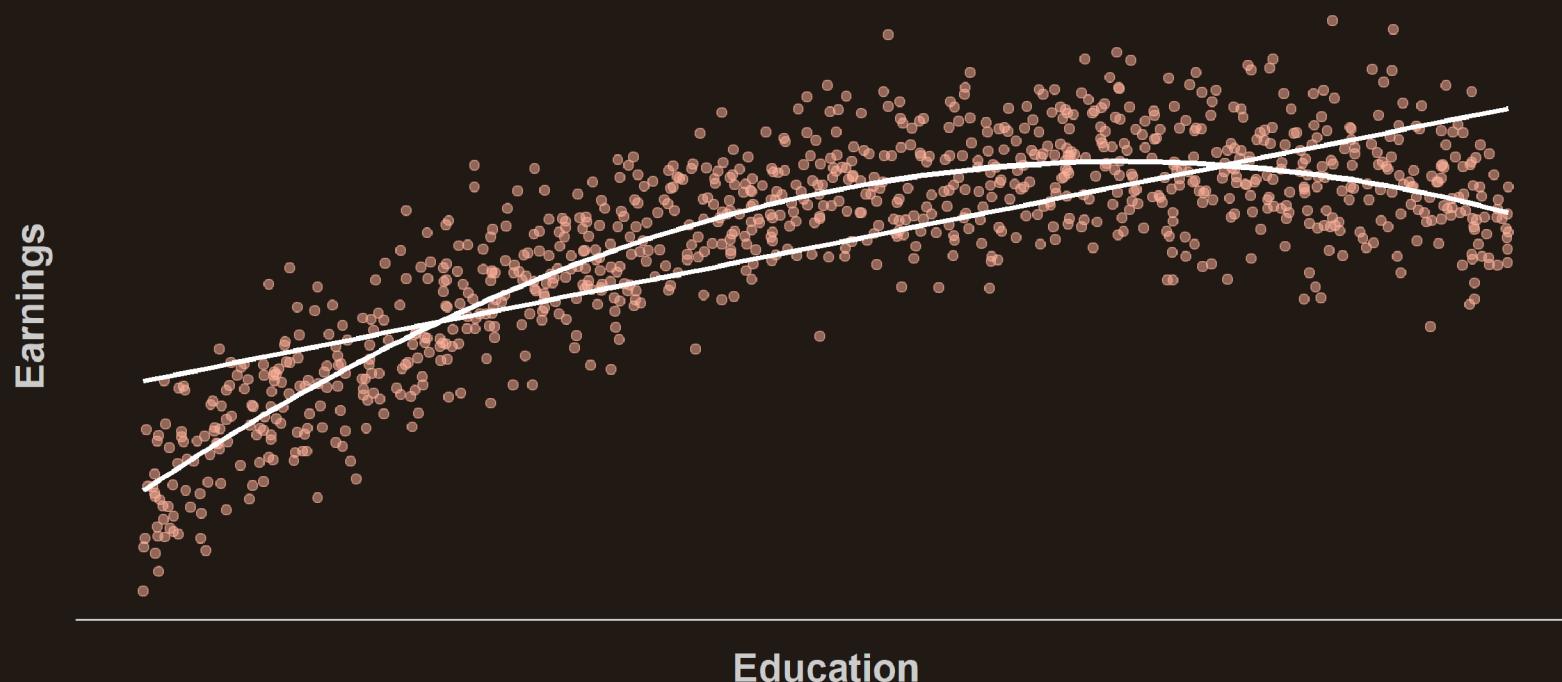
- If a third **variable** is correlated with both  $x$  and  $y$ , it would **bias the relationship**
  - We must then **control** for such variables
  - And if we can't we must acknowledge that our estimate is not causal with '*ceteris paribus*'



### 3. Wrap up!

#### Functional form

- Not capture the **right functional** form correctly might also lead to biased estimations:
  - Polynomial order, interactions, logs, discretization matter
  - **Visualizing the relationship** is key





### 3. Wrap up!

#### Selection bias

- **Self-selection** is also a common threat to causality
- What is the impact of going to a better neighborhood on your children outcomes?
  - We cannot just regress children outcomes on a mobility dummy
  - Individuals who move may be different from those who stay: **self-selection issue**
  - Here it is not that the sample is not representative of the population, but that **the outcomes of those who stayed are different from the outcomes those who moved would have had, if they had stayed**

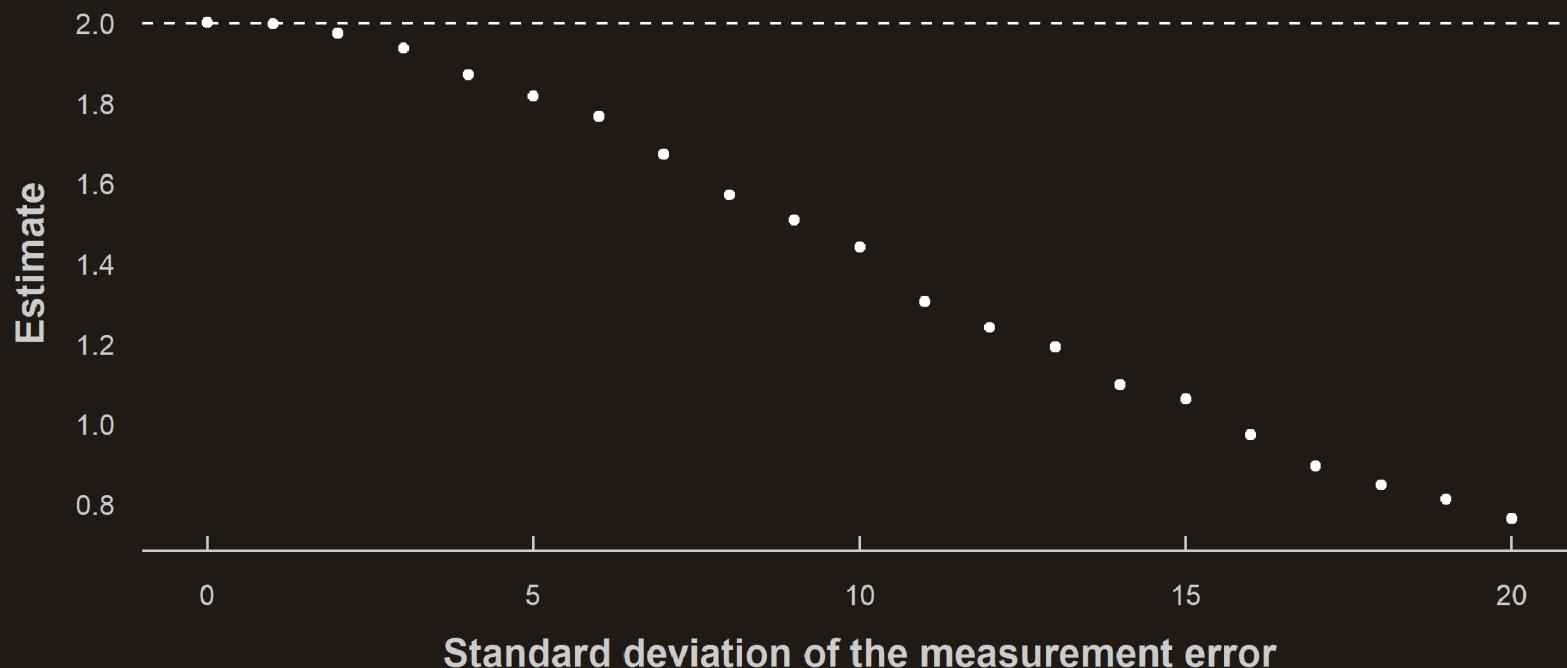
#### Simultaneity

- Consider the relationship between **crime** rate and **police coverage** intensity
- What is the **direction of the relationship?**
  - We cannot just regress crime rate on police intensity
  - It's likely that more crime would cause a positive response in police activity
  - And also that police activity would deter crime

### 3. Wrap up!

#### Measurement error

- **Measurement error** in the independent variable also induces a bias
  - The resulting estimation would mechanically be **downward biased**
  - The **noisier** the measure, the **larger the bias**





### 3. 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
  - The two **groups** would then have the same characteristics on expectation, and would be **comparable**
  - It is a way to obtain **causality** from randomness
- RCTs are very **powerful tools** to sort out issues of:
  - Omitted variables
  - Selection bias
  - Simultaneity
- But RCTs are **not immune** to every problem:
  - The sample must be representative and large enough
  - Participants should comply with their treatment status
  - Independent variables must not be noisy measures of the variable of interest
  - ...