Week 4: Threats to valid inferences 2

Selection and Omitted Variable Bias

POLS0007
Principles of Social Science Research

University College London

1. Last Week: Internal and External Validity
2. Random Selection
3. Selection on the Dependent Variable (Y)
4. Selection on the Independent Variable (X)
5. Self-Selection
6. Omitted Variable Bias

Last Week

- Association does not imply causation!
- Association only makes you cross the 3rd hurdle
- Each research design has a level of internal and external validity
- Internal validity: how confident we are about the causal inference derived from a design
- External validity: to what extent we can translate knowledge obtained from a study to other settings
- Randomization in experiments alleviates many threats to internal validity but not all
- Experiments may have high internal but low external validity

Selection

Ideal vs. In Practice

- Experiments are ideal if both internal and external validity are high
  - Random assignment of X: X is causing Y
  - Random selection of the sample: Results can be generalized
- In practice many experiments do not select at RANDOM
  - Not possible to have a random sample (e.g. too expensive)
  - Sometimes not what we want: study a specific population

Selection

- How do we select the cases that we want to study?
  - If we select the wrong ones we introduce a selection bias
  - Our results will be biased
- Example
  - We formulate a specific *hypothesis*, and knowing what we want the outcome of our research to be, we select only observations that support our hypothesis
- Different possible biases, due to different types of selection

1) Selection on Y

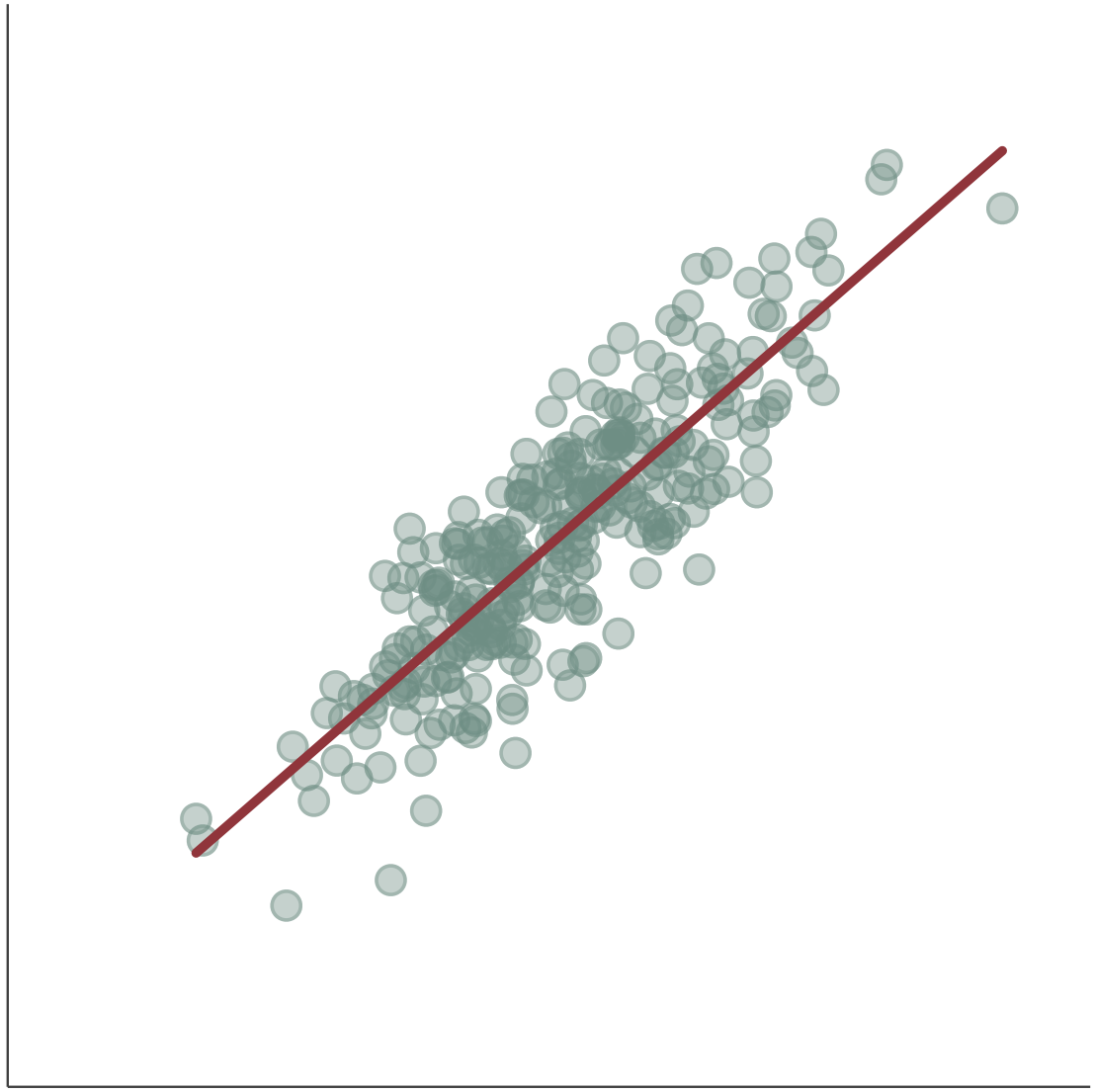
- Selection should allow for the possibility of at least some variation on the dependent variable Y
  - How do we study the causes of migration if we select only migrants?
  - How do we study if smoking causes cancer if we only select people with cancer?
- We can select values of Y when designing our research, but we need to be aware of the biases we introduce (and find ways to correct for them)

Selection on Y: Underestimation of Causal Effect

- We can still study the causal effect of X on Y
  - but our results are likely to be biased
- Any selection rule correlated with the dependent variable (Y) attenuates (weakens) the estimates of the causal effect
  - on average, the **true causal effect is larger** than what we find in our study
  - our estimates are a **lower bound** of the true causal effect
- Our results are biased, but in a predictable way that we can compensate for

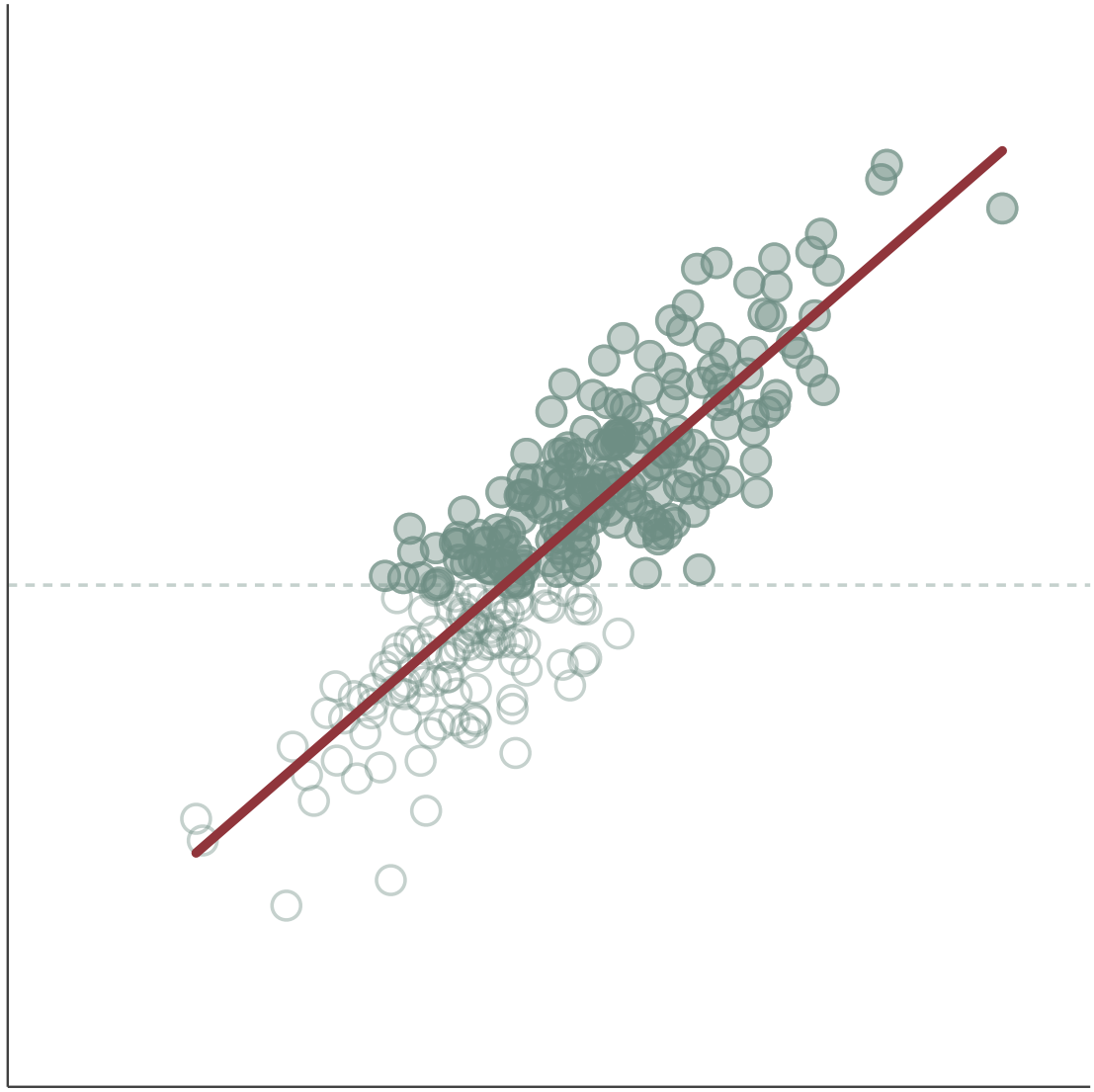
Selection on Y

- True population relationship



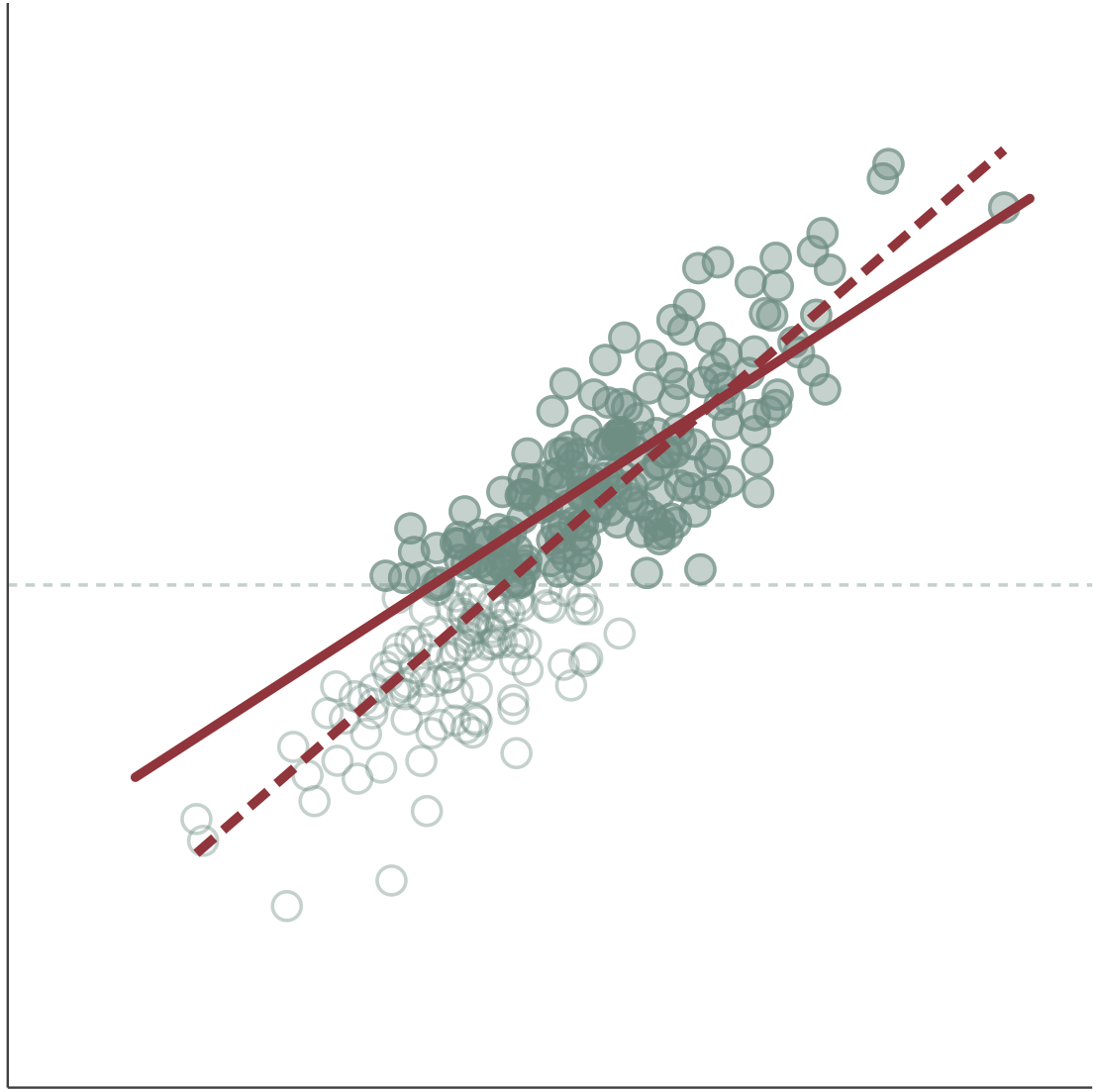
Selection on Y

- True population relationship
- Estimated only observing those with high Y



Selection on Y

- True population relationship
- Estimated only observing those with high Y
- Estimated relationship is weaker



Selection on Y: Overestimation of Causal Effect

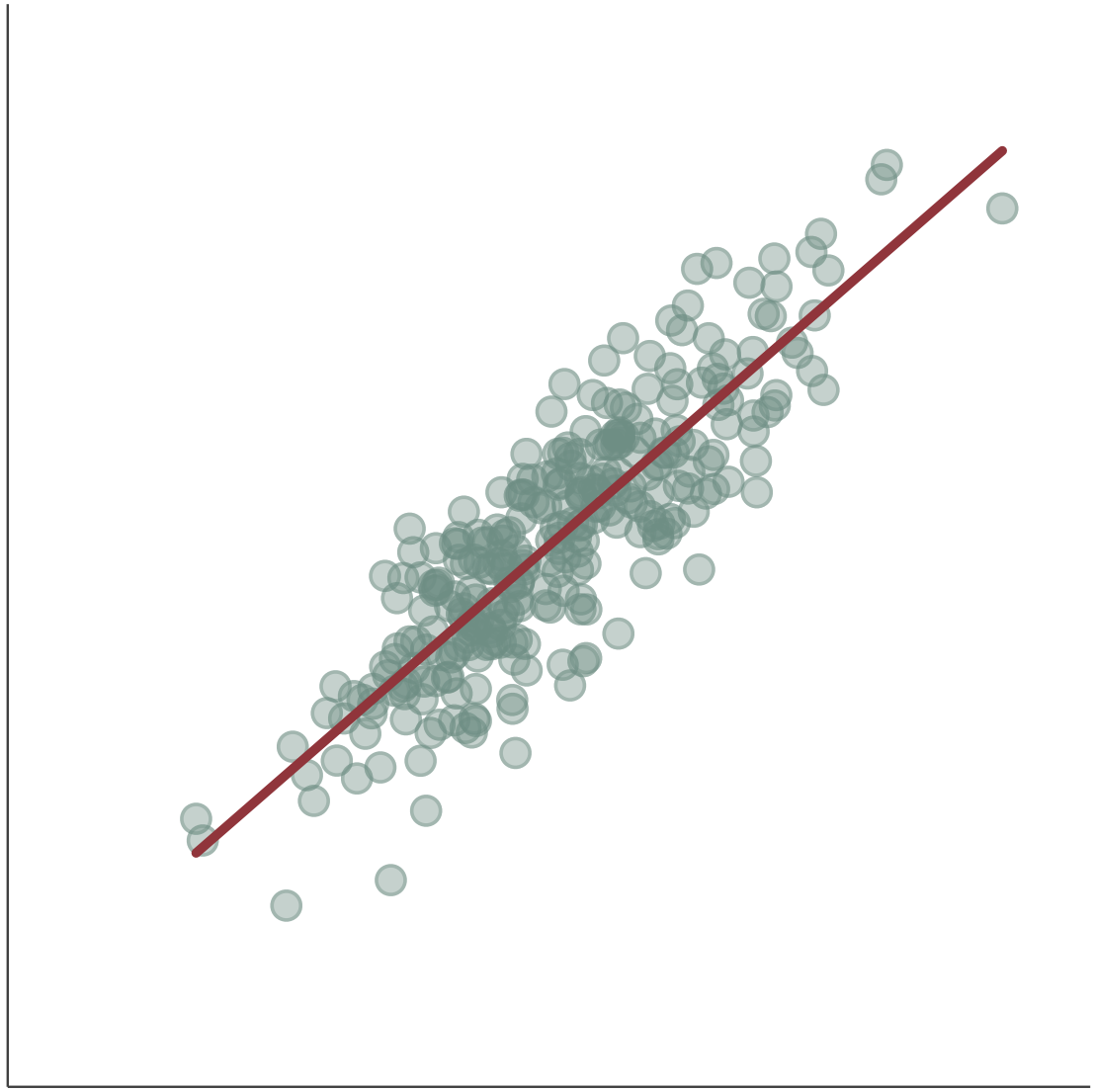
- Is it possible to *overestimate* a causal effect?
  - Yes! If the causal effect of X on Y varies across observations (NON LINEAR)
  - A selection rule correlated with the size of the causal effect would induce bias

2) Selection on X

- Selection of observations to be included in a study based on the main independent variable X causes **no problems**
  - Selecting based on the values of X doesn't restrict the variation in Y
  - It may limit the generality of our conclusions

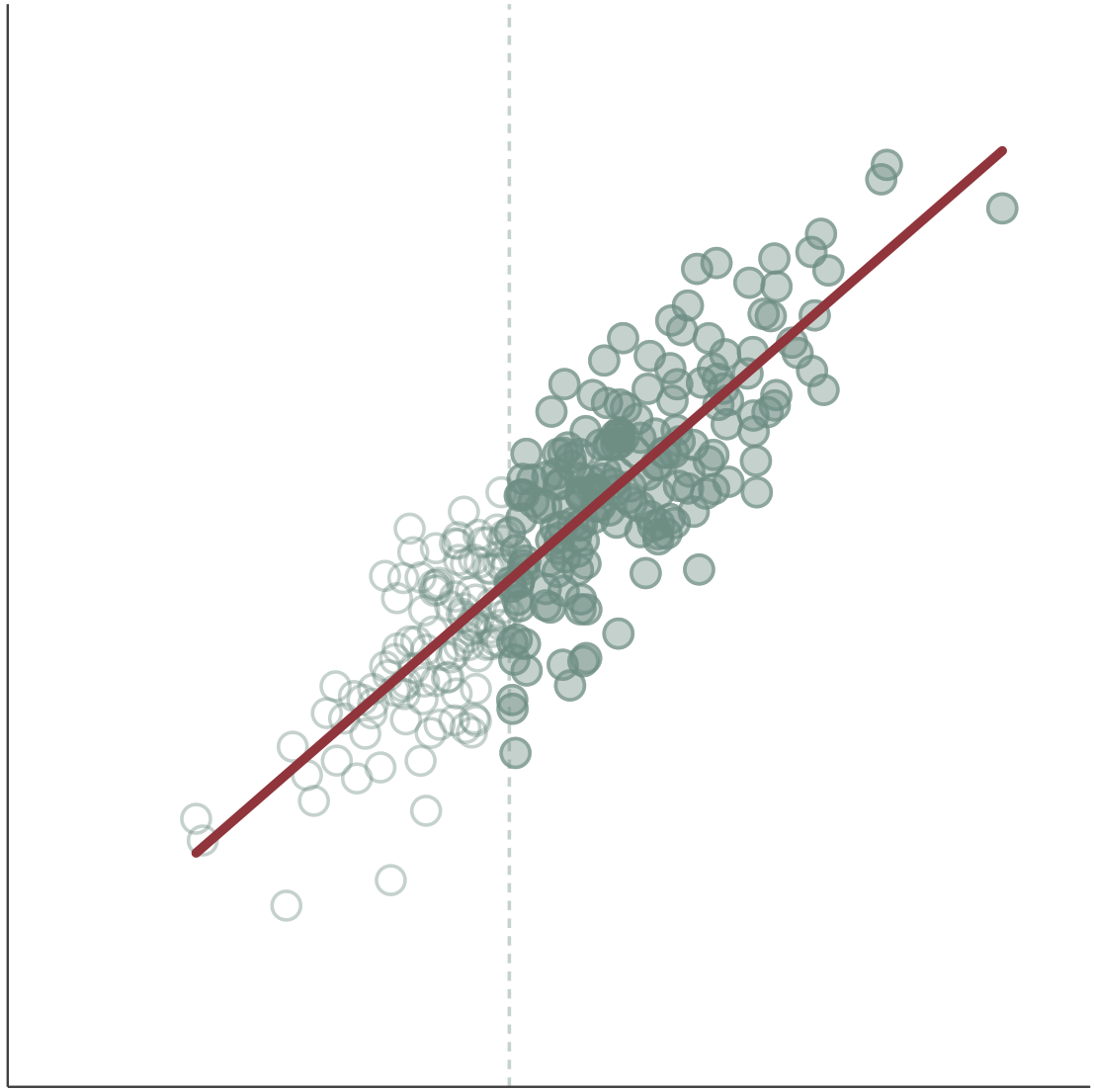
Selection on X

- True population relationship



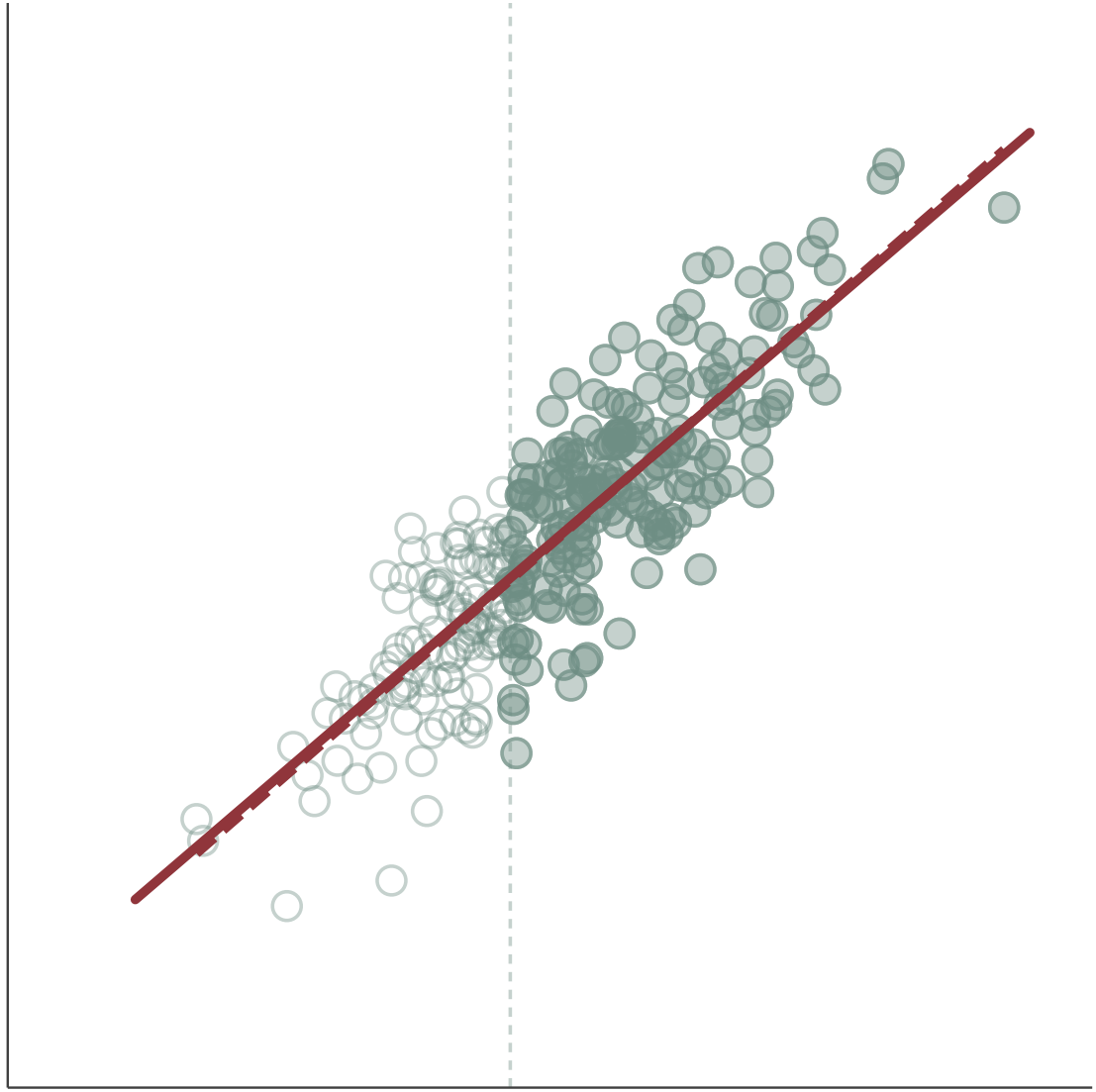
Selection on X

- True population relationship
- Estimated only observing those with high X



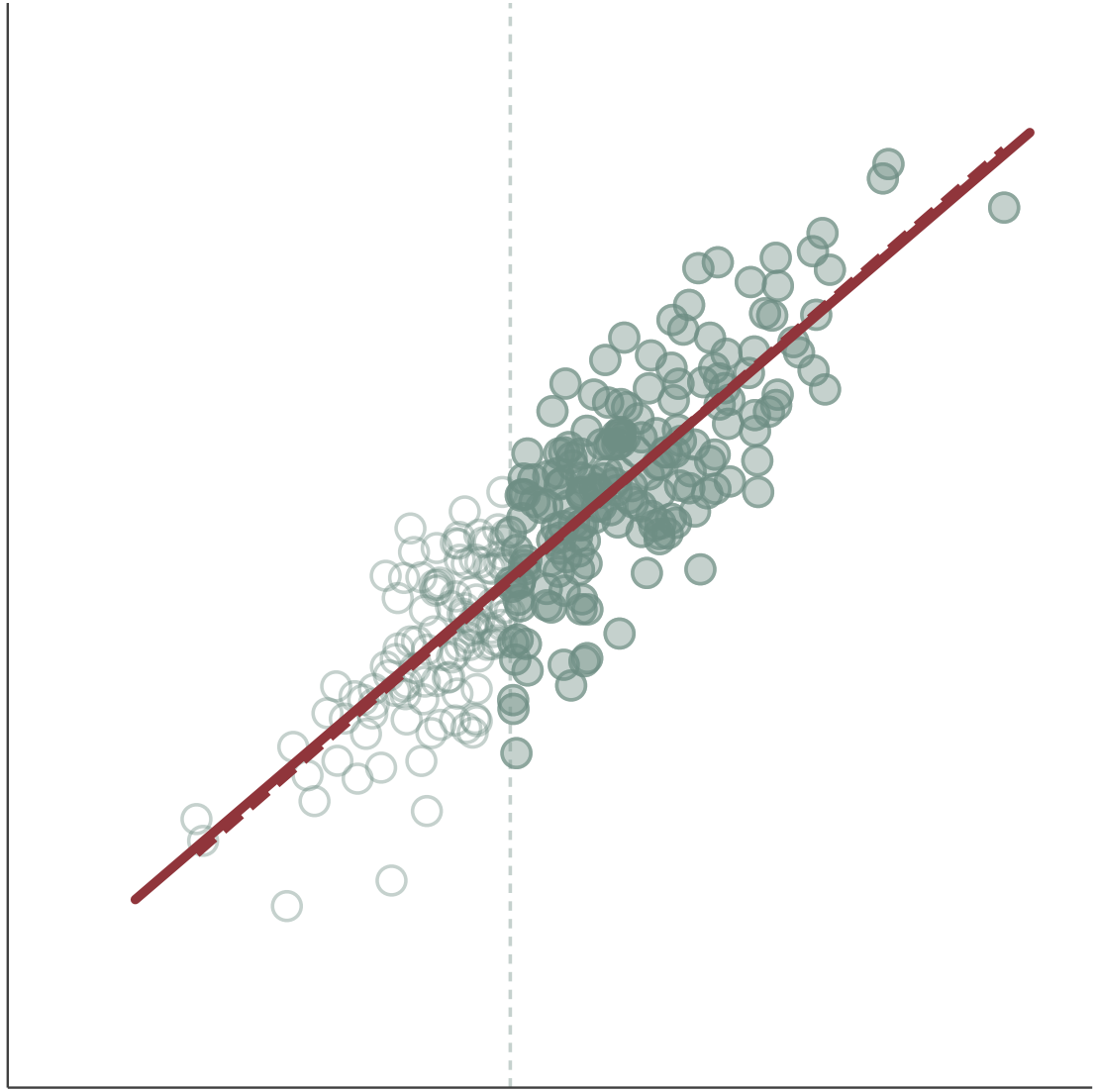
Selection on X

- True population relationship
- Estimated only observing those with high X
- Amazingly, no bias in this case



Selection on X

- True population relationship
- Estimated only observing those with high X
- Amazingly, no bias in this case
- But: real-world cases can be more complicated



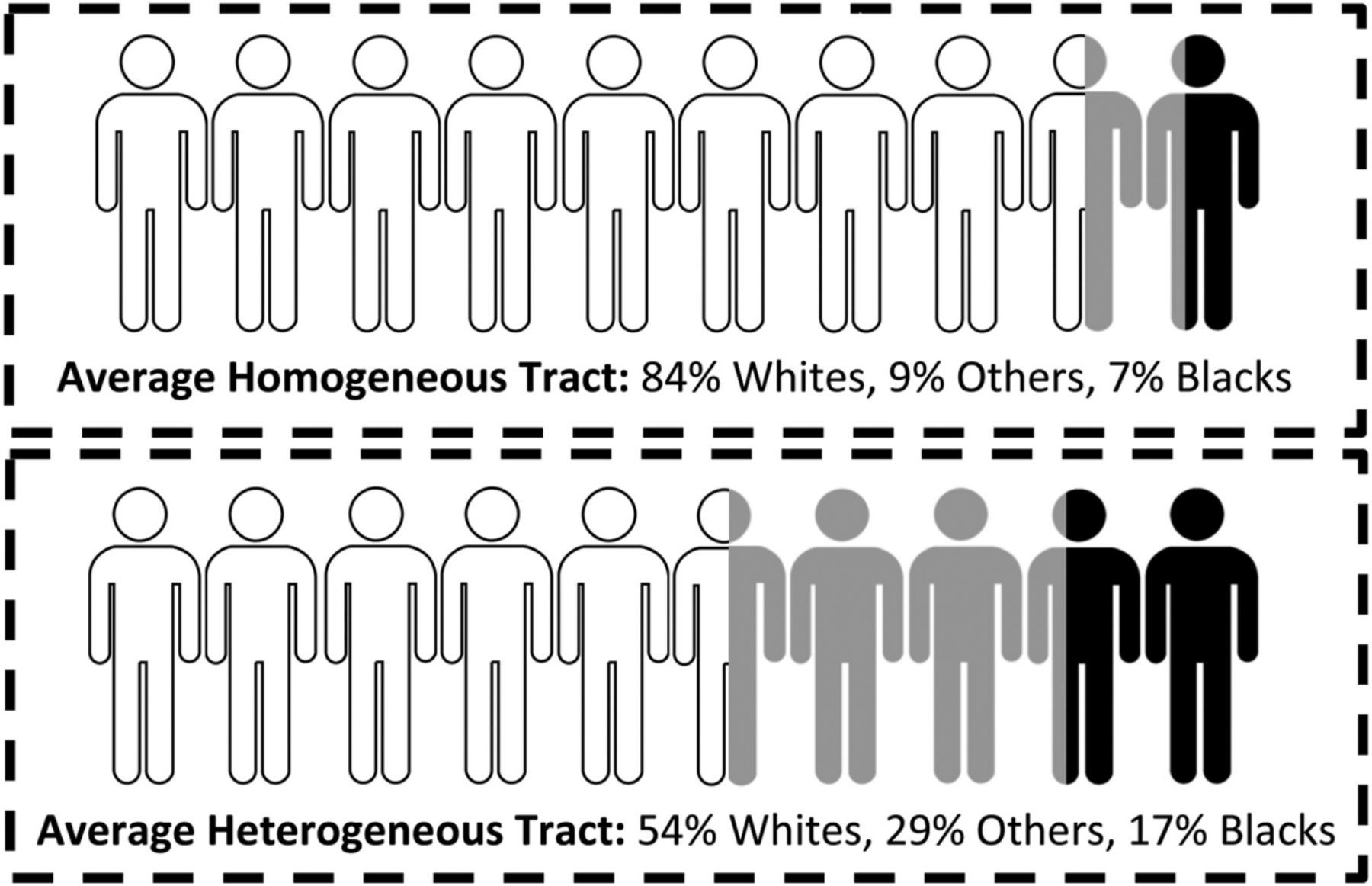
3) Self-selection

- Sometimes a bias is not introduced by the researcher or by the world
- Individuals select themselves into a group, causing a biased sample
- The characteristics of those in the sample are different from the characteristics of those not in the sample
- Example
  - **Healthy Immigrant Effect**: immigrants are on average healthier than native-born
  - Is migration (X) causing better health (Y)?

3) Self-selection

- Sometimes a bias is not introduced by the researcher or by the world
- Individuals select themselves into a group, causing a biased sample
- The characteristics of those in the sample are different from the characteristics of those not in the sample
- Example
  - **Healthy Immigrant Effect**: immigrants are on average healthier than native-born
  - Is migration (X) causing better health (Y)?

  - **Immigrant self-selection**: Migrants tend to be positively selected on ambition, education, health, wealth, etc: they are better off than those who stayed behind



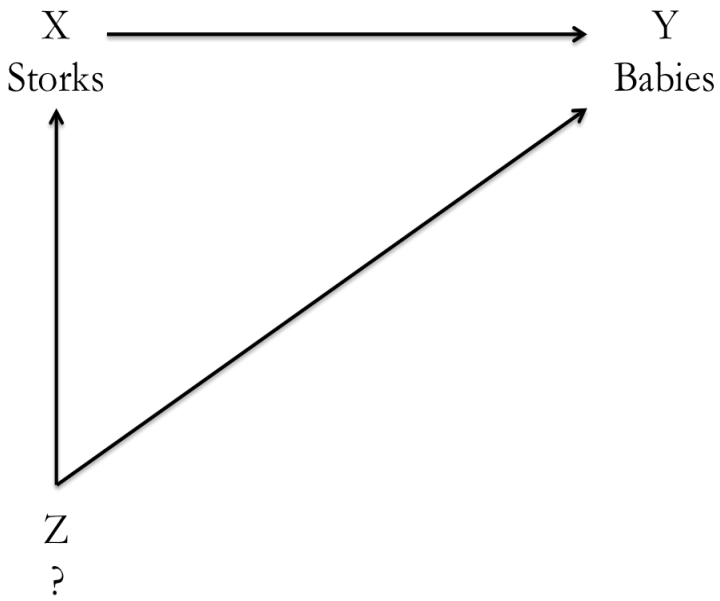
Abascal and Baldassarri (2015) AJS Volume 121 Number 3: 722–82.

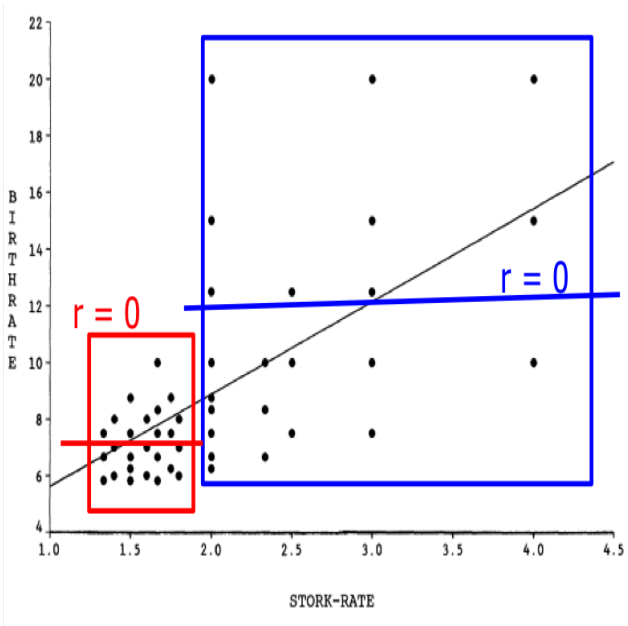
Omitted Variable Bias

What if we do not control for Z?

Storks and Babies

- The higher the **presence of storks** the higher the **birth rate**
- Do storks bring babies and *cause* the increase of fertility?
- Is there **an alternative explanation (Z)** to the observed relationship?





Omitted Variable Bias

- If we fail to take Z into account our estimates are biased, unless...
- (1) Z has NO effect on Y, i.e. Z is irrelevant

  OR
- (2) Z is NOT correlated with X

- If there is an omitted variable Z we should take it into account in the analysis. If we don't have data on Z, we should determine the direction of the bias

Example

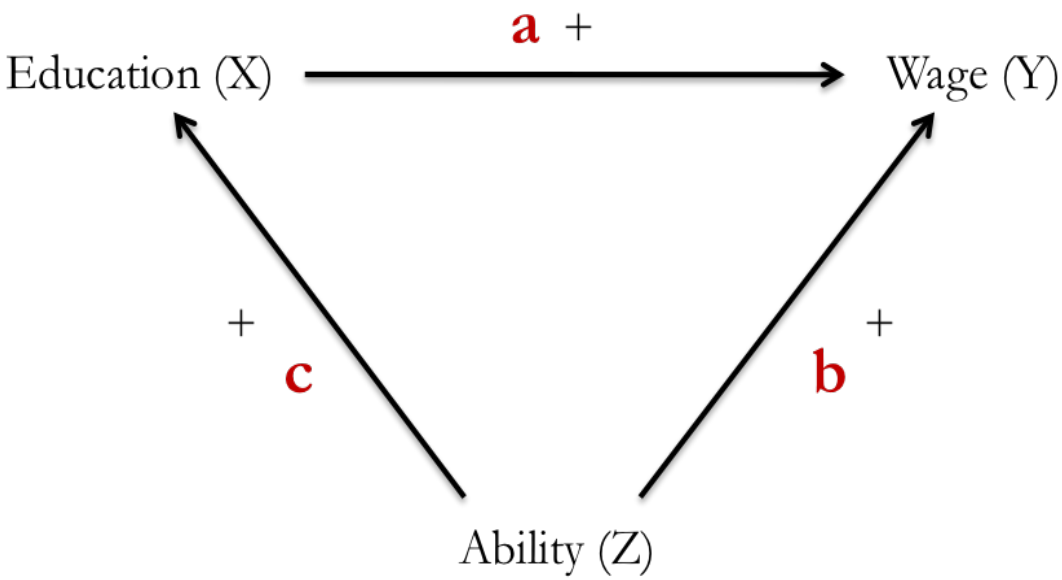
- Does more education (X) cause a higher wage (Y)?

$$wage = a * education$$

- a is the TRUE causal effect of education on wage
- Is there any **omitted variable Z** that is correlated with wage and education?

$$wage = a * education + b * Z$$

Cognitive Ability



- $wage = a * education + b * ability$
- causal effect of education on wage $= a$
- causal effect of ability on wage $= b$
- causal effect of ability on education $= c$

Omitted Variable Bias

- If we do not consider ability (Z), what we think is the true causal effect (a) of education on wage is biased

$$estimated\ effect = true\ causal\ effect + bias$$

$$estimated\ effect = a + b * c$$

- If Z has no effect on Y, then $b = 0$ and $causal\ effect = a$
- If Z is not correlated to X, then $c = 0$ and $causal\ effect = a$
- If Z has a positive effect on Y ($b > 0$) and Z has a positive correlation with X ($c > 0$), the bias is positive and $estimated\ effect > a$ (overestimation)

Direction of the Bias

- Are we under- or overestimating the causal effect of X on Y?

	$c > 0$	$c < 0$
$b > 0$	positive bias	negative bias
$b < 0$	negative bias	positive bias

- Positive bias: overestimation
- Negative bias: underestimation

Example of Underestimation

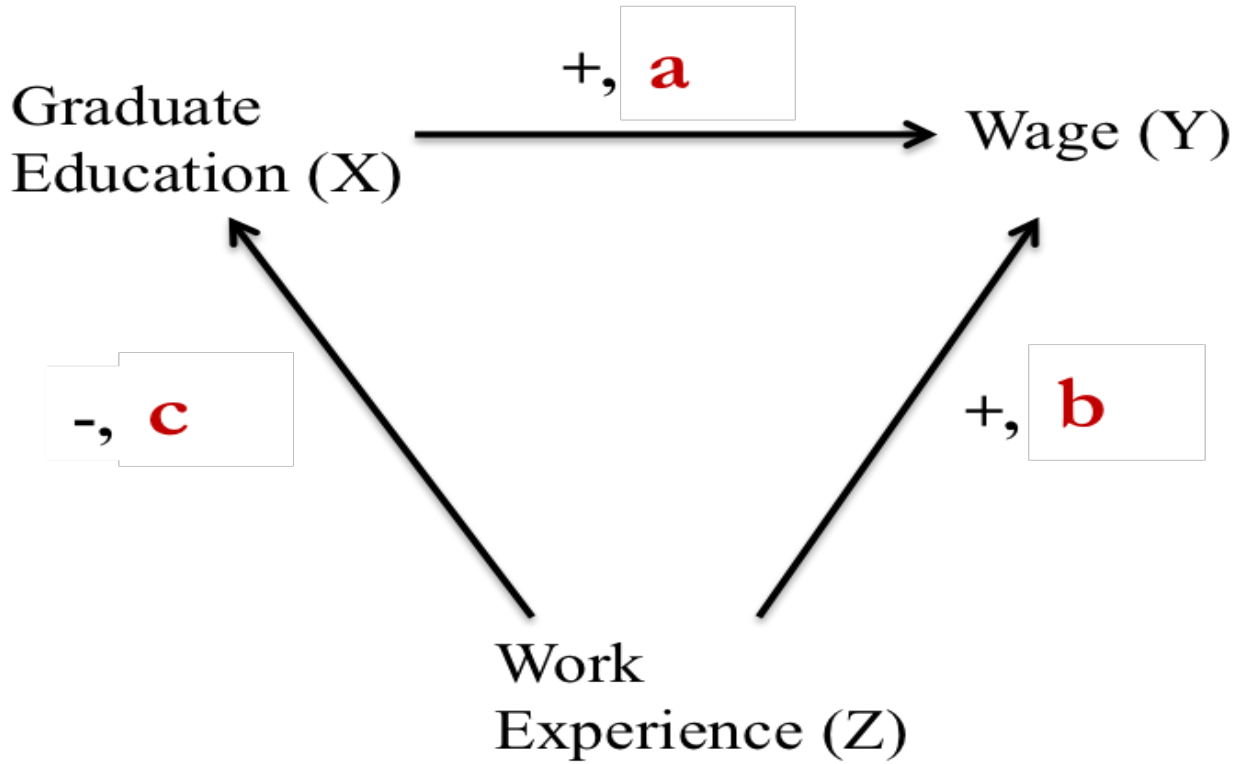
- Effect of graduate education (X) on wages (Y)?

$$wages = a * graduate\ education$$

- a is the TRUE causal effect of graduate education on wage
- is there any **omitted variable Z** that is correlated with wage and graduate education?

$$wage = a * graduate\ education + b*?$$

Example of Underestimation



Example of Underestimation

- Z = work experience

$$wage = a * graduate\ education + b * work\ experience$$

- $a > 0$, graduate education increases your wage
- $b > 0$, work experience increases your wage
- $c < 0$, graduate education and work experience are negatively correlated
- $estimated\ effect = a + b * c$
- $estimated\ effect = true\ causal\ effect + bias$
- $estimated\ effect < true\ causal\ effect : UNDERESTIMATION$

Conclusion (and Take Home Points)

- When trying to study causal relationships is important to:
  - **Select accurately** the cases/observations in the sample
  - Make sure we are considering all the **relevant variables**
  - Selection Bias (Sample, Individuals in the sample) $\neq$ Omitted Variable Bias (Variables in the analysis)
  - If we are aware of the problem (selection or omitted variable) but we cannot solve it, discuss the **direction and the magnitude of the bias**