Interpreting regressions + Exam practice

https://pollev.com/jonathanold608

Roadmap - Regression, regression, regression

- Discussion PSet 1
- (Recap:) Control variables and bad controls
- How to approach an Econ 140R exam question
- Practice Exam questions

Problem Set 1:
some quick points

PSet 1 - some notes on interpretation

Whether there is a causal relationship or not **does not depend on**:

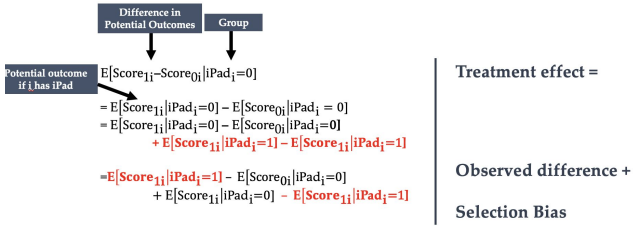
- Outliers
- Heterogeneous treatment effects
- Variance
- The size of a coefficient

But it **does** depend on:

- Selection Bias
- Omitted variable bias
- Reverse causality
- ...

PSet 1 - some notes on interpretation

Recall: Every observed effect in the data is a composite of:



Treatment effect =

Causal Effect

Observed difference +

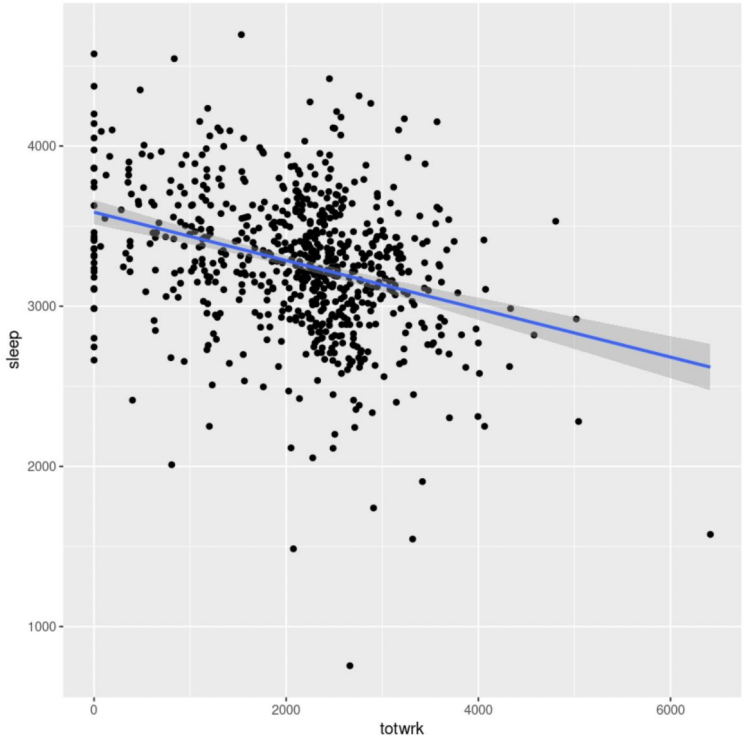
+

Selection Bias

Selection Bias

- We cannot observe either of these - so we can never be sure!
- Econometrics is all about uncertainty. You can state that there are different possibilities and you cannot know for sure!

PSet 1 - let's try together



- What does the data say?
- How do we interpret this?
- What would an RCT do?

Recap and big picture: Control variables

Why do we use control variables?

In decreasing order of importance:

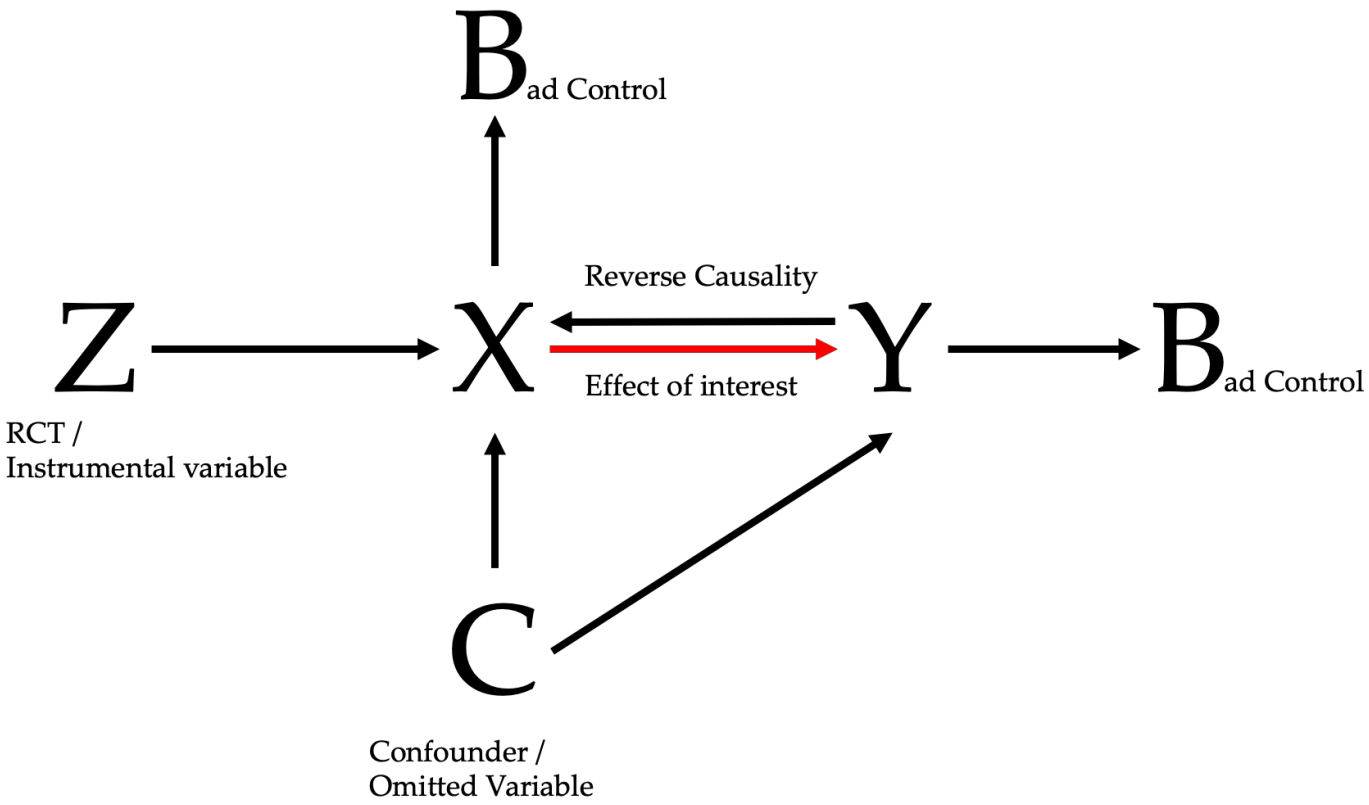
- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

How control variables work

- See here: https://nickchk.com/causalgraphs.html

Are more controls always good?

Are more controls always good?



Are more controls always good?

No, because:

- Some controls are **BAD controls**
- Some controls are unrelated to the outcome of interest - such variables are **unnecessary** and increase the risk of spurious correlation and overfitting

So what are “good” controls?

- Predetermined (before the “treatment” of interest)
- Immutable characteristics (not changed by treatment)

Interpreting regressions: A guide

[See separate slides]

Exam practice

A systematic way to approach an exam question

1. **Think**: About the question, about the real world
2. Start with the numbers **you see**
3. **Interpret** the numbers:
    a. Direction (positive or negative?)
    b. Magnitude (big or small?)
    c. Significance (significant or insignificant, at what level?)
4. Establish whether estimated relationship is **causal or not**
    a. Is X-variable randomized or as good as random? Do we have valid counterfactuals?
    b. If not: Do you expect **bias**? Of which sort (**OVB, reverse causality**, ...)?
    c. Find a plausible story around bias (using the OVB formula)
5. How would you find a causal effect?
    a. Can you run an RCT?

There are no traps!

Practice Exam: Question 1 (RCT, coefficient interpretation)

You are analyzing the results from a carefully constructed randomized evaluation of the effect of workplace incentives on worker productivity in a large moving company. You regress the log of worker output (measured in kg moved per hour) on a constant and an indicator for whether the individual was randomly selected to receive a financial incentive for high productivity.

A. **(4 marks)** The coefficient on the indicator variable is 0.20 with a standard error of 0.05. Explain, as if to a non-technical audience, what this estimate implies.

B. **(4 marks)** A colleague notices that because of the physical nature of the work, men are likely to be more productive at this job than women, all else equal. As such, your colleague claims, the worker's gender constitutes an omitted variable. You have data on each employee's gender so you rerun the regression including a gender dummy. How do you expect your estimate for the effect of incentives to change and why?

C. **(4 marks)** Another colleague claims that women may be more or less responsive to workplace financial incentives than men. Describe the regression you would run to test this claim and how you would interpret the results.

Practice Exam: Question 2 (Regression and conceptual)

Consider the following hypothetical example. The Department of Education wants to evaluate the effectiveness of a program from 16- and 17-year olds designed to increase the number of students studying science, technology, engineering and mathematics (STEM) at college and university. Some secondary schools offer the program and others do not.

A. **(10 marks)** You have data from a sample of 25-year olds. The data include each individual's current labour market income and whether or not the secondary school the individual attended offered the STEM program at the time the individual attended. Write down the equation for the bivariate regression using these two variables, with labour market income as the outcome variable. Define the variables precisely. How would you interpret the regression coefficients (describe them as if to a general audience)? Would you expect the slope coefficient to capture the causal effect of explanatory variable on the outcome? Why or why not?

B. **(7.5 marks)** A colleague makes the following claim: "You do not need to worry about establishing causality. The data were gathered using a random sample, so they are not biased." Critically assess this claim.

Practice Exam: Question 2 (Regression controls)

Consider the following hypothetical example. The Department of Education wants to evaluate the effectiveness of a program from 16- and 17-year olds designed to increase the number of students studying science, technology, engineering and mathematics (STEM) at college and university. Some secondary schools offer the program and others do not.

D. **(7.5 marks)** The program was funded by tax revenues raised by local districts. Therefore, you gather additional data and control for average district income in your regression. First, explain as if to a general audience what it means to control for a variable in a regression. Second, describe what you expect would happen to the slope coefficient you described above when you control for income. Clearly state any assumptions you make.

Practice Exam: Question 3 (Regression)

The Ministry of Truth is interested in a rumour that air pollution could impact mental health. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, gas, or grass fires. A research team is sent to investigate the rumour. The team randomly selects and surveys 19,920 people across 71 districts of the country. The key variable, $Exposure_i$, is a dummy variable equal to 1 if the individual i is exposed to a large amount of PM2.5 in the last two years, and 0 otherwise. The team also conducts a standardised questionnaire to record depressive symptoms in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, $Depression_i$, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual i. The variable has a sample average of 2.96. Running regressions with $Depression_i$ as the dependent variable, the analyst Winston Smith obtains the following results:

Practice Exam: Question 3 (Regression)

Dependent variable: $Depression_i$

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
$Exposure_i$	0.834	0.635	0.614	0.598	0.422	0.554
	(0.032)	(0.039)	(0.045)	(0.021)	(0.020)	(0.042)
$Exposure_i \times Female_i$		-0.834	0.065	0.054		
		(0.013)	(0.024)	(0.011)		
$Female_i$			-0.739	-0.732	-0.745	-0.825
			(0.036)	(0.018)	(0.019)	(0.066)
$Manual\ Job_i$				0.324	0.050	
				(0.122)	(0.008)	
Age_i					0.324	0.452
					(0.111)	(0.132)
Age_i^2					0.421	0.524
					(0.122)	(0.121)

Notes. All estimations contain a constant term. Robust standard errors are in the parentheses. $Manual\ Job_i = 1$ if i works in a manual job, and 0 otherwise. $Female_i = 1$ if i is a female, 0 otherwise. Age_i is the age (years old) of individual i, and Age^2 is the square of Age_i.

Practice Exam: Question 3

Dependent variable: $Depression_i$

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
$Exposure_i$	0.834	0.635	0.614	0.598	0.422	0.554
	(0.032)	(0.039)	(0.045)	(0.021)	(0.020)	(0.042)
$Exposure_i \times Female_i$		-0.834	0.065	0.054		
		(0.013)	(0.024)	(0.011)		
$Female_i$			-0.739	-0.732	-0.745	-0.825
			(0.036)	(0.018)	(0.019)	(0.066)
$Manual\ Job_i$				0.324	0.050	
				(0.122)	(0.008)	
Age_i					0.324	0.452
					(0.111)	(0.132)
Age_i^2					0.421	0.524
					(0.122)	(0.121)

Notes. All estimations contain a constant term. Robust standard errors are in the parentheses. $Manual\ Job_i = 1$ if i works in a manual job, and 0 otherwise. $Female_i = 1$ if i is a female, 0 otherwise. Age_i is the age (years old) of individual i, and Age^2 is the square of Age_i.

a) Interpreting the coefficient in Column (1), a journalist, Katherine, claims: "Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression."

i. Explain carefully why Katherine is wrong, specifying the direction of bias(es) if there is any. Which assumption(s) would she need to impose for the causality claim to hold?

[13 marks]

ii. What is the correct interpretation from Column (1) that Katherine should have made?

[2 marks]

Practice Exam: Question 3

Dependent variable: $Depression_i$

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
$Exposure_i$	0.834	0.635	0.614	0.598	0.422	0.554
	(0.032)	(0.039)	(0.045)	(0.021)	(0.020)	(0.042)
$Exposure_i \times Female_i$		-0.834	0.065	0.054		
		(0.013)	(0.024)	(0.011)		
$Female_i$			-0.739	-0.732	-0.745	-0.825
			(0.036)	(0.018)	(0.019)	(0.066)
$Manual\ Job_i$				0.324	0.050	
				(0.122)	(0.008)	
Age_i					0.324	0.452
					(0.111)	(0.132)
Age_i^2					0.421	0.524
					(0.122)	(0.121)

Notes. All estimations contain a constant term. Robust standard errors are in the parentheses. $Manual\ Job_i = 1$ if i works in a manual job, and 0 otherwise. $Female_i = 1$ if i is a female, 0 otherwise. Age_i is the age (years old) of individual i, and Age^2 is the square of Age_i.

c) Another analyst, Julia, suggests that the team should include into the regressions a variable, $Health\ Expenditure_i$, which captures individual i's total expenditures on health-related services in the last two years. Her rationale is that financial distress may cause depression and is related to air pollution in the area where people with financial distress work. Should the team follow her suggestion? Explain why or why not.

[10 marks]