Week 2: How to read a regression table in a social science paper

Dr. Gabriel Geisler Mesevage

gabriel.mesevage@kcl.ac.uk
Office Hours: Wed. 10-11am & Thurs. 2-3pm

21 January 2020

What you will see

	(1) OLS
Log of initial income	−6.0603***
	[1.1125]
Growth in capital stock per capita	0.1971**
	[0.0760]
Growth in land per capita	0.3077
	[0.1855]
Log of 1 + agricultural tariff	−5.0697
	[4.8451]
Log of 1 + manufacturing tariff	14.7219**
	[6.1808]
Log of 1 + exotic tariff	−1.3786
	[0.8040]
Constant	8.1735***
	[1.3241]
Observations	70
Number of countries	10
R^2	0.51

The equation behind the table

$$\texttt{gdp\_growth\_pc}_i = \alpha + \beta_1 log(\texttt{initial\_income}_i) + \\ \beta_2 \texttt{growth\_capital\_stock}_i + \beta_3 \texttt{growth\_land\_pc}_i + \\ \beta_4 log(\texttt{agric\_tariff}_i + 1) + \beta_5 log(\texttt{manuf\_tariff}_i + 1) + \\ \beta_6 log(\texttt{exotic\_tariff}_i + 1) + \epsilon_i$$

- α and the β's are **parameters**: they are estimated from the data
- The numbers in the table are the **parameters** associated with each **variable**
  - E.g. $\beta_1 = -6.0603$ is the parameter for `initial_income`
- ϵ_i is the **error**:
  - The difference between what the model suggests and the data
  - We do not observe the true ϵ_i
  - We *do* observe how the model differs from the data: we call these observations $\hat{\epsilon}$ where the hat denotes an estimate

Let's start with a simple example

Simple example

Interpretation: Linear Conditional Expectation

- The line that makes the best guess at average y given x

A simple regression

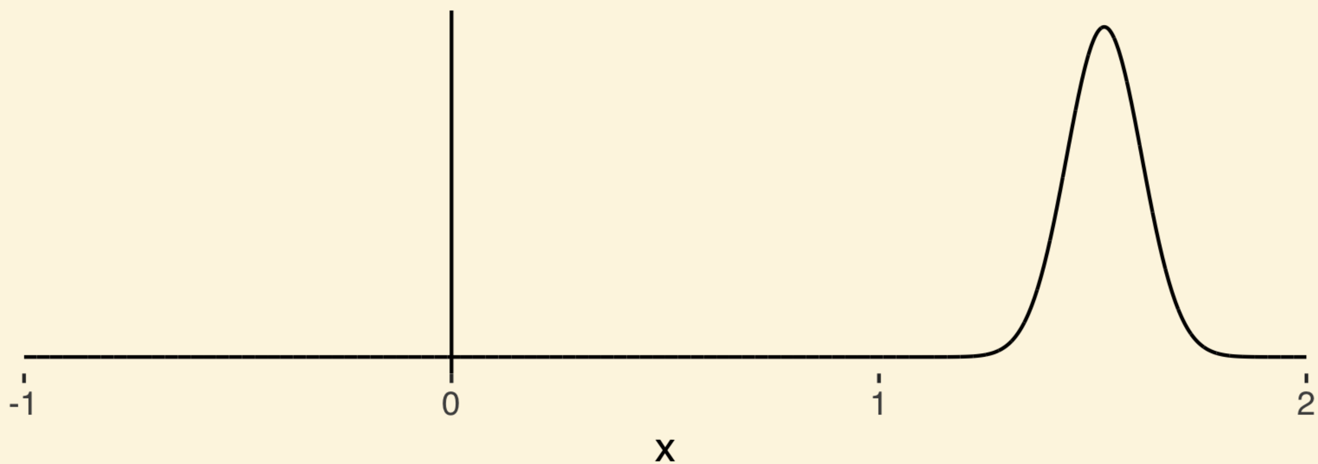
Example Regression Table

	Dependent variable:
	y
x	1.527^{***} (0.088)
Constant	39.912^{***} (0.091)
Observations	100
Adjusted R^2	0.753
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

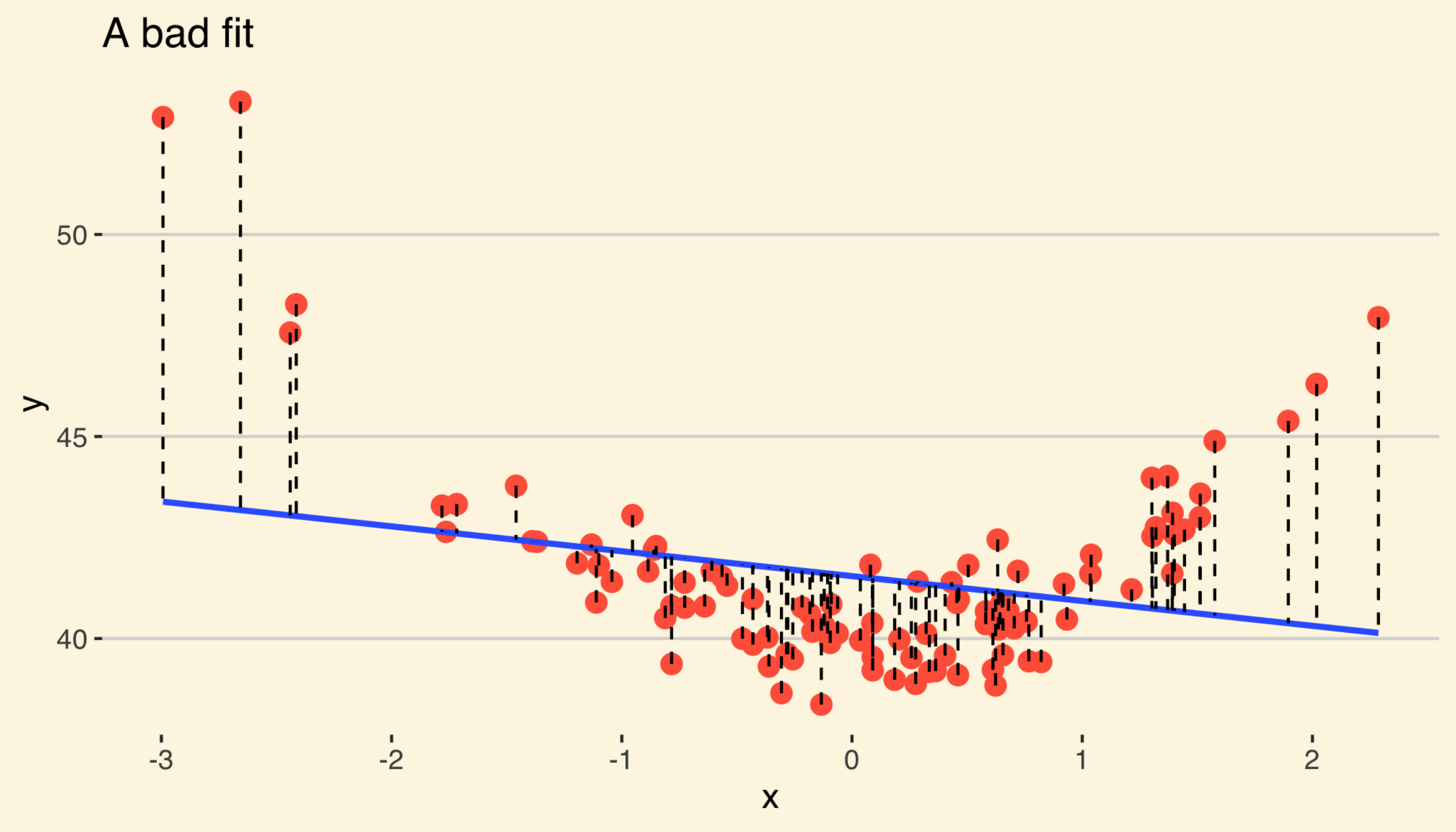
- **Standard errors** are in parantheses
- A **statistically significant** coefficient means the **slope is non-zero**
    - The stars by x indicate statistical significance
    - The stars by $Constant$ indicate that the intercept is far from zero

Statistical significance

- We estimated the coefficient on $\beta_1 = 1.527$
- We estimated its standard error to be $se_1 = .088$
  - This means we think on **average** the coefficient will be 1.527 with a variance of $se_1^2 = .088^2 = .007744$
- The mean and average imply a distribution for the coefficient -- is it far from zero?
  - Rule of thumb: multiply the standard-error by 2 and add/subtract from coefficient for 95% confidence interval

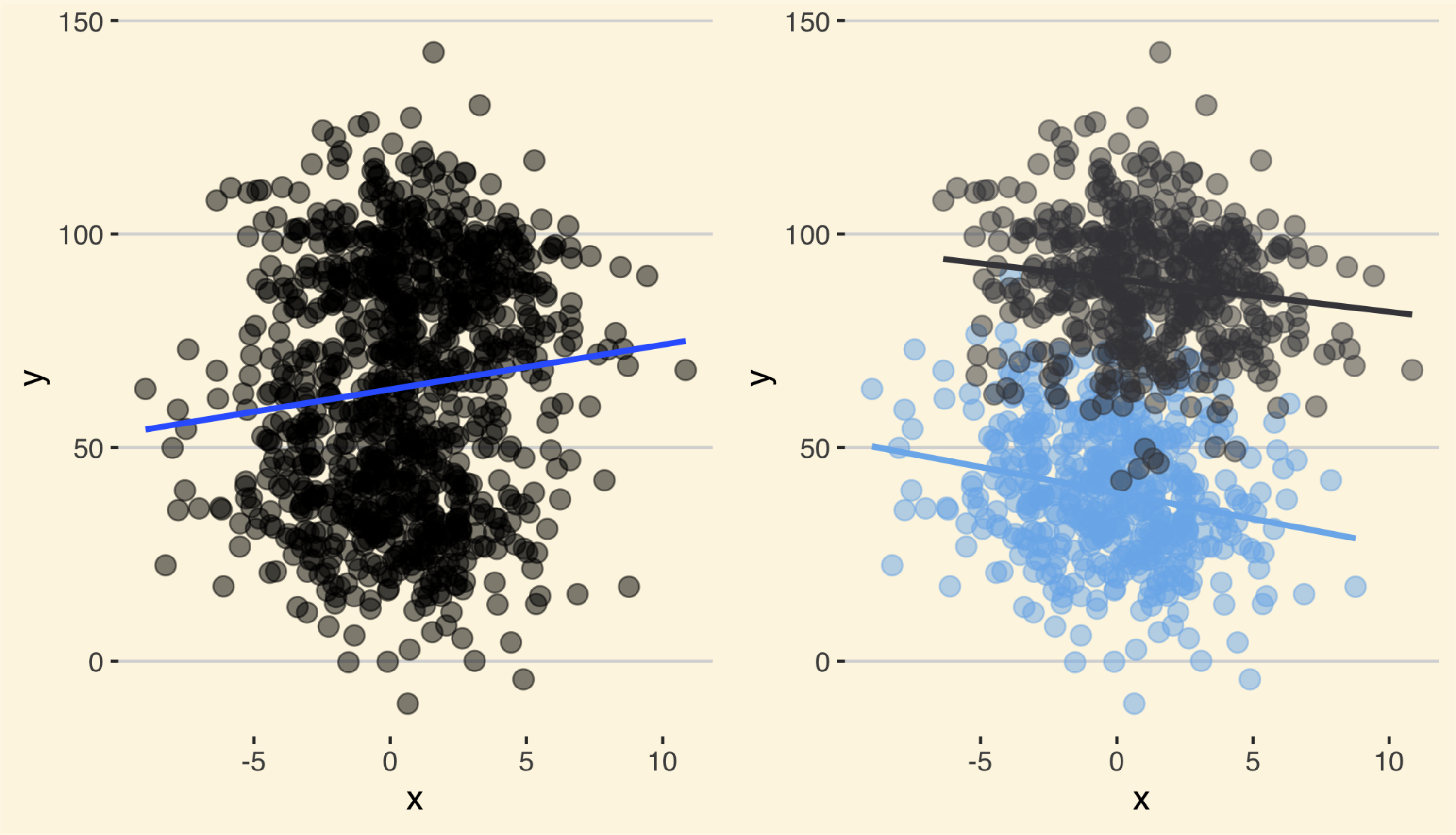


Assumptions: the relationship is linear



- Here the true model is $y = \alpha + \beta x^2 + \epsilon$
  - With two variables plotting your data should usually let you fix a problem like this

What happens with more variables?



Omitted Variable Bias in Action

	Dependent variable:	
	y	
	(1)	(2)
x	1.045*** (0.309)	-0.984*** (0.170)
z		50.174*** (1.002)
Constant	63.647*** (0.922)	39.452*** (0.689)
Observations	1,000	1,000
R^2	0.011	0.719
Note:		*p<0.1; **p<0.05; ***p<0.01

- The value a parameter takes depends on the other variables in the model
  - We call this **omitted variable bias**
  - Why you **should not** interpret regressions as **causal relations**
  - Intuition: z is large whenever x is large. If you look only at y and x the effect of x combines the effects of x and z
- Why does randomization produce **causal estimates**?
- R^2: a measure of how much of the variation in y the model explains

Questions to think about when reading a regression

- Did the authors omit variables that could change the relationship and should be included?
  - The constant concern of any statistical argument
- Is the model appropriate?
- Is the relationship of a meaningful magnitude?

Resources

Morgan, Stephen L., and Christopher Winship. Counterfactuals and Causal Inference : Methods and Principles for Social Research, Cambridge University Press, 2007: This is an introductory graduate level text frequently used in sociology and political science. It focuses on causal inference from data analysis.

Jeffrey Wooldridge, *Introductory Econometrics: A Modern Approach*, is what is often used with economics undergrads (no digital copy in the library unfortunately).

Joshua D. Angrist & Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, is a popular graduate-level treatment for economists that surveys common approaches.