Section 2

Econ 152

Spring 2020

GSI: Andrew Tai

Introduction

In this section, we'll cover concepts about regression and causal inference.

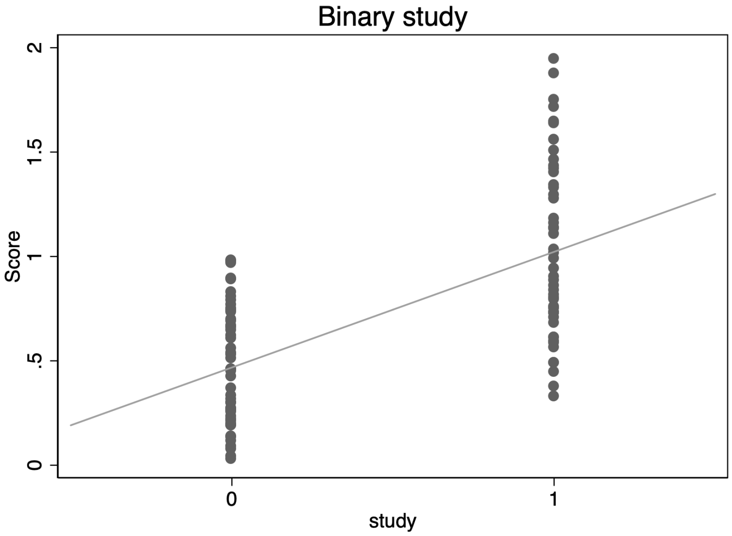
First, some terminology:

- **Regression** is a statistical technique for quantifying relationships. Most of you will be familiar with OLS regression; in this course, we will also use difference-in-differences and instrumental variables regression.
- **Causal inference** is the process of drawing conclusions about causal connections. Regression is one way to make causal inference – it's the most popular way for social scientists.

Exercise 1. We want to know whether studying for an exam increases scores on the exam. I asked 100 students whether they studied (1 if they did, 0 if they didn't), then estimated a regression. Here's the scatterplot and regression table of

$$\text{score}_i = \hat{\alpha} + \hat{\beta} \times \text{studied}_i + e_i$$

(i denotes an individual, here from 1 to 100; e_i is the regression error.)



	(1)
	score
study	0.554
	(0.0712)
constant	0.468
	(0.0394)
N	100

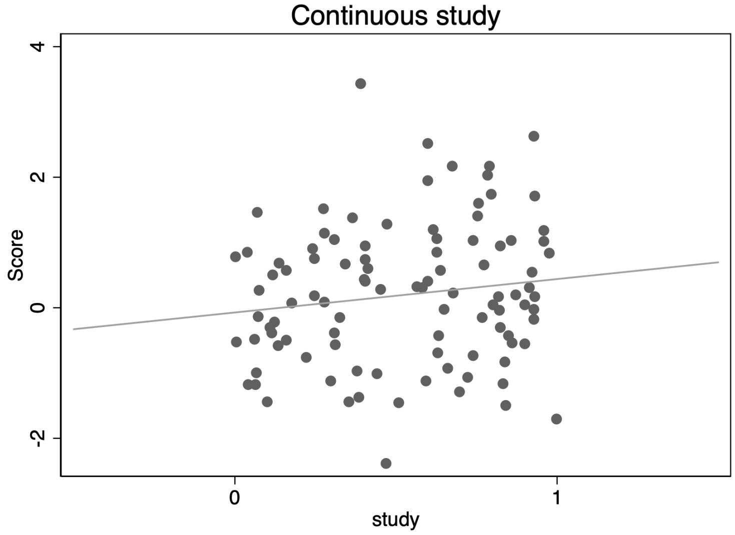
Standard errors in parentheses

1. What was the average score of students who didn't study? Of students who did?
2. Can you conclude that studying causes higher exam scores? Why or why not?

Exercise 2. Now suppose I asked the students how much they studied on a scale from 0 to 1 (rather than whether they did). Here's the scatterplot and regression table of

$$\text{score}_i = \hat{\alpha} + \hat{\beta} \times \text{StudyIntensity}_i + e_i$$

What was the average score of students who studied with 0 intensity? (In statistical notation, $\mathbb{E}\left[\text{score}_i \mid \text{StudyIntensity}_i = 0\right]$.) Of students who studied a 0.5 ($\mathbb{E}\left[\text{score}_i \mid \text{StudyIntensity}_i = 0.5\right]$)?



	(1)
	score
StudyIntensity	0.512
	(0.322)
constant	-0.0723
	(0.186)
N	100

Standard errors in parentheses

Regression

Conditional average vs. causation

The first important point is that OLS doesn't give causal inference by itself. It's a statistical calculation that gives conditional averages.

Concept. Let D_i be a binary variable, i.e. $D_i = 0$ or $D_i = 1$. This can represent some binary treatment/non-treatment. Consider a regression equation $Y_i = \alpha + \hat{\beta} D_i + e_i$. The predicted values of Y_i, denoted $\hat{Y}_i$, are just the conditional linear averages: $\hat{Y}_i = \mathbb{E}\left[Y_i \mid D_i\right]$. Regression answers this question – for some value of X, what is the average Y among individuals with that value of D (0 or 1)?

More technical note. The interpretation for nonbinary X_i in $Y_i = \alpha + \hat{\beta} X_i + e_i$ is similar – regression gives the *linear* conditional average $\hat{Y}_i = \mathbb{E}^*\left[Y_i \mid X_i\right]$. (Linear here means that the conditional average has to change at a constant rate with X_i, given by the slope of the regression line). As long as the true relationship is linear, this isn't a problem. What's an example of a relationship that might not be linear?

Establishing causality with regression is really a question about where the regression came from.

Regression algebra

Definition. Correlation and **covariance** are related by

$$corr(X_i, Y_i) = \frac{Cov(X_i, Y_i)}{SD(X_i) SD(Y_i)}$$

Let's say we're interested in the causal impact of education on wages. How should we think about using regression to quantify this?

Concept. Suppose we have a model $Y_i = \alpha + \beta X_i + \varepsilon_i$, where we think that X_i *causes* Y_i. Think of ε_i as other potential determinants of Y_i. To measure β, we decide to fit a regression

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

Recall(?) that

$$\begin{aligned}
\hat{\beta} &= \frac{Cov(Y_i, X_i)}{Var(X_i)} \\
&= \frac{Cov(\alpha + \beta X_i + \varepsilon_i, X_i)}{Var(X_i)} \\
&= \frac{Cov(\alpha, X_i) + Cov(\beta X_i, X_i) + Cov(\varepsilon_i, X_i)}{Var(X_i)} \\
&= \frac{0 + \beta Var(x_i) + Cov(\varepsilon_i, X_i)}{Var(X_i)} \\
&= \beta + \frac{Cov(\varepsilon_i, X_i)}{Var(X_i)}
\end{aligned}$$

What's the issue here? $\frac{Cov(\varepsilon_i, X_i)}{Var(X_i)}$ might not be 0, so there's the potential that $\hat{\beta} \neq \beta$. This is called **omitted variable bias**. The idea is that ε_i, the unspecified determinants, might be correlated with X_i, and this will cause our estimate $\hat{\beta}$ to be biased.

Exercise 3. Back to our education and wages example, suppose we fit the regression

$$\text{wages}_i = \hat{\alpha} + \hat{\beta} \times \text{college}_i + e_i$$

where $\text{college}_i = 1$ if the individual has a college degree, and 0 otherwise. Remember that the causal model is

$$\text{wages}_i = \alpha + \beta \times \text{college}_i + \varepsilon_i$$

where ε_i is anything else affecting wages.

1. What is the correct interpretation of $\hat{\beta}$? Is it causal?
2. What are potential sources of $Cov(\varepsilon_i, X_i)$ that might bias $\hat{\beta}$ upward? (Think about what positive $Cov(\varepsilon_i, X_i)$ means.) How about negative?

Experiments or: How I Learned to Start Worrying and Love Randomization

This is the reason that randomized controlled trials (RCTs) are the "gold standard" for causal inference in science.

Concept. If the assignment of X_i is randomized, then $Cov(\varepsilon_i, X_i) = 0$. (Why?).

Randomized experiments, in which X_i is assigned randomly, can eliminate omitted variable bias even if you don't include the omitted variables in your regression. In economics, we usually can't run experiments, but there are methods that are **quasi-experimental**, like difference-in-differences and instrumental variables. These methods aren't literally experiments, but we hope that the data are *like* experiments.