Introduction: Economic Questions and Data (SW Ch. 1)

Ercan Karadas

Econometrics (ECON 3112)

Belk College of Business, UNCC

August 28, 2018

Why Do We Study Econometrics???

- Many decisions in economics, business, and government hinge on understanding relationship among variables in the world around us:
  - What is the quantitative effect of reducing class size on student achievement?
  - How does another year of education change earnings?
  - What is the price elasticity of cigarettes?
  - What is the effect on output growth of a 1 percentage point increase in interest rates by the Fed?
  - What is the effect on housing prices of environmental improvements?
- These questions require quantitative answers $\rightarrow$ **Econometrics!**

How Does Econometrics Provide Answers?

- We will answer these type of questions by measuring **causal effects.**
- In common usage, causality means that a specific action leads to a specific, measurable consequence.
- Ideally, we would like to conduct an experiment for measuring causal effects
- Question: Does putting fertilizer on your tomato plants cause them to produce more tomatoes?
- Experiment:
  - Plant many plots of tomatoes, each plot is tended identically except some plots get fertilizer, while the rest get none
  - Whether a plot is fertilized or not is determined randomly $\rightarrow$ any other differences between the plots are unrelated to whether they receive fertilizer
  - At the end of the growing season, weigh the harvest from each plot.
  - The difference between the average yield of the treated and untreated plots is the *causal* effect on tomato production of the fertilizer treatment.

How Does Econometrics Provide Answers?

- This is an example of a **randomized controlled experiment**, in which subjects are split into two groups randomly
  - **Control group:** receives no treatment (no fertilizer)
  - **Treatment group:** receives the treatment (fertilized)
- Randomization ensures that the only systematic difference between the control and treatment groups is the treatment.
- What would be an experiment to estimate the effect of class size on standardized test scores?
  - Randomly assign "treatments" of different class sizes to different groups of students
  - If the experiment is designed and conducted properly the only systematic difference between the groups of students is their class size, and this experiment would estimate the effect on test scores of reducing class size, holding all else constant, i.e. causal effect.

How Does Econometrics Provide Answers?

- The concept of an ideal randomized controlled experiment is useful because it gives a definition of a causal effect and serves as a benchmark for us.
- But almost always we only have observational (nonexperimental) data.
- For example, in our leading example we only observe classes of different sizes and their performances
- Most of this course deals with difficulties arising from using observational data to estimate causal effects
  - confounding effects (omitted factors)
  - simultaneous causality
  - 'correlation does not imply causation'

In this course you will:

- Learn methods for estimating causal effects using observational data
- Learn some tools that can be used for other purposes; for example, forecasting using time series data;
- Focus on applications - theory is used only as needed to understand the whys of the methods;
- Learn to evaluate the regression analysis of others - this means you will be able to read/understand empirical economics papers in other econ courses;
- Get some hands-on experience with regression analysis in your problem sets;
- Master R programming language!

The Leading Example: Class Size and Student Performance

- **Policy question:** What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? by 8 students/class?
- This is going to be our **leading example** throughout the semester and we will keep adding more features to the analysis to make more sophisticated arguments
- We must use data to find out (is there any way to answer this without data?)

Data: The California Test Score Data Set

- All K-6 and K-8 California school districts (n = 420)
- Variables:
  - District average of 5th grade test scores **(Test Score)**,
  - Student-teacher ratio **(STR)**:

$$\text{STR} = \frac{\#\text{ of students in the district}}{\#\text{ of full-time equivalent teachers}}$$

$\Longrightarrow$ Data Set: $\{\text{Test Score}_i, \text{STR}_i\}_{i=1}^{420}$

Data: Summary Statistics

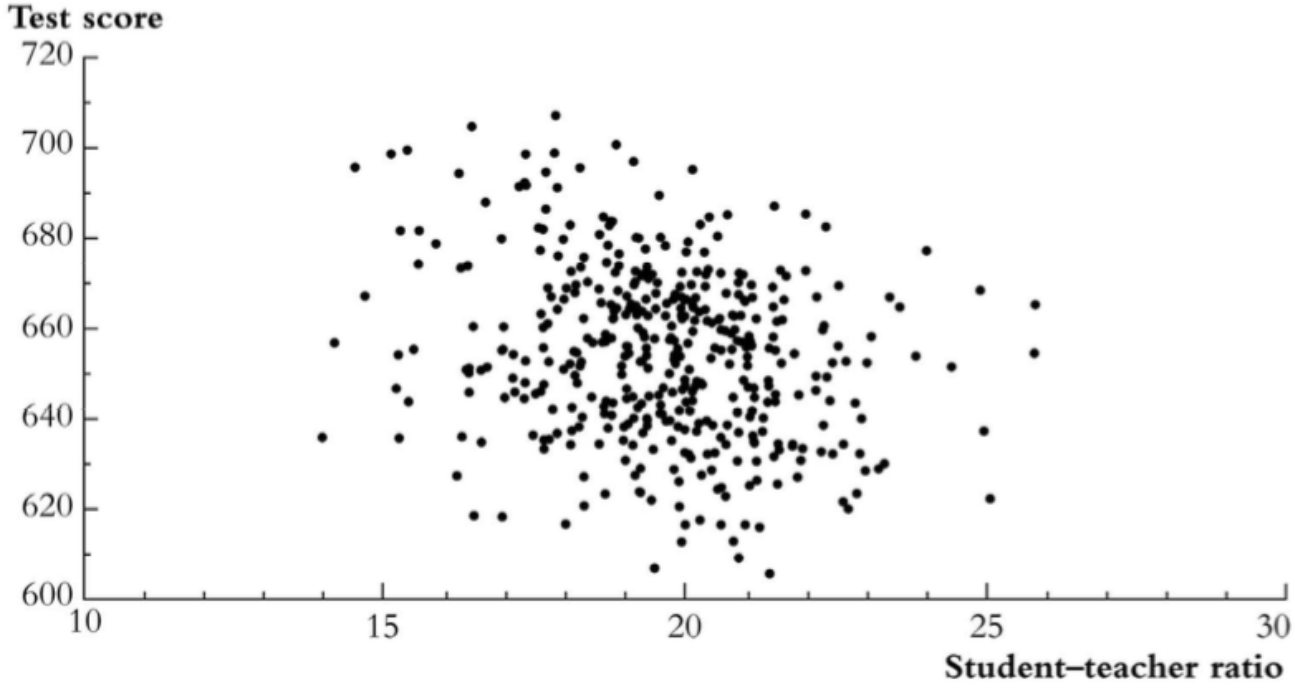
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

- This table doesn't tell us anything about the relationship between Test Scores and the STR.

Data: Scatterplot

- Scatterplot of test score v. student-teacher ratio



- What does this figure show?
- Do districts with smaller classes have higher test scores?
- We need to get some numerical evidence on whether districts with low STRs have higher test scores - but how?

Statistician's Answer to $\downarrow$ STR $\Longrightarrow$ $\uparrow$ test scores?

- Compute the difference between average test scores in districts with low STRs and those with high STRs $\rightarrow$ **Estimation**
- Test the null hypothesis that the mean test scores in the two types of districts are the same, against the alternative hypothesis that they differ $\rightarrow$ **Hypothesis Testing**
- Estimate an interval for the difference in the mean test scores, high v.s low STR districts $\rightarrow$ **Confidence Interval**

Initial data analysis

Compare districts with 'Small' (STR $<$ 20) and 'Large' (STR $\geq$ 20) class sizes: Tabulation of group means

Class Size	Average Score ($\overline{Y}$)	Std. Dev. (s_Y)	**n**
Small ($STR < 20$)	657.4	19.4	238
Large ($SRT \geq 20$)	650.0	17.9	182

Define $\Delta =$ difference between group means

- **Estimation** of Δ?
- **Test the hypothesis** that $\Delta = 0$?
- Construct a **confidence interval** for Δ?

1. Estimation

- A natural estimator for Δ:

$$\begin{aligned}\hat{\Delta} &\equiv \bar{Y}_S - \bar{Y}_L \qquad (\text{ an estimator for } \Delta) \\ &= \frac{1}{n_S}\sum_{i=1}^{n_S} Y_i - \frac{1}{n_L}\sum_{i=1}^{n_L} Y_i \\ &= 657.4 - 650.0 \\ &= 7.4\end{aligned}$$

- Is this a large difference in a real-world sense?
  - Standard deviation across districts = 19.1
  - Difference between 60th and 75th percentiles of test score distribution is 667.6 - 659.4 = 8.2
  - This is a big enough difference to be important for school reform discussions, for parents, or for a school committee?

2. Hypothesis testing

- Is $\Delta \neq 0$ statistically?
- Compute t-statistic:

$$t = \frac{\hat{\Delta} - 0}{SE(\hat{\Delta})}$$
$$= \frac{\bar{Y}_S - \bar{Y}_L}{SE(\bar{Y}_S - \bar{Y}_L)}$$

where $SE(\bar{Y}_S - \bar{Y}_L)$ is the 'standard error ' of $\bar{Y}_S - \bar{Y}_L$:

$$SE(\bar{Y}_S - \bar{Y}_L) = \sqrt{\frac{s_S^2}{n_S} + \frac{s_L^2}{n_L}}$$

and

$$s_S^2 = \frac{1}{n_S - 1} \sum_{i=1}^{n_S} \left(Y_i - \bar{Y}_S\right)^2, \qquad s_L^2 = \frac{1}{n_L - 1} \sum_{i=1}^{n_L} \left(Y_i - \bar{Y}_L\right)^2$$

- Compute the difference-of-means t-statistic from the data:

Class Size	Average Score ($\overline{Y}$)	Std. Dev. (s_Y)	**n**
Small ($STR < 20$)	657.4	19.4	238
Large ($SRT \geq 20$)	650.0	17.9	182

$$\implies t = \frac{\bar{Y}_S - \bar{Y}_L}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_L^2}{n_L}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

- $|t| > 1.96$: so **reject** (at the 5% significance level) the null hypothesis that the two means are the same.

3. Confidence interval

- Point estimation does not reflect all the uncertainty that comes with working a single sample of data points $\rightarrow$ construct confidence interval
- A 95% confidence interval for the difference between the means (Δ):

$$\begin{aligned}\hat{\Delta} \pm 1.96 SE(\hat{\Delta}) &= (\bar{Y}_S - \bar{Y}_L) \pm SE(\bar{Y}_S - \bar{Y}_L) \\ &= 7.4 \pm 1.96 \times 1.83 \\ &= (3.8, \quad 11.0)\end{aligned}$$

- Two equivalent statements:
  - The 95% confidence interval for Δ doesn't include 0;
  - The hypothesis that $\Delta = 0$ is rejected at the 5% level.

Did the Statistician Answer the Policy Question?

- **Policy question:** What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? by 8 students/class?
- Not quite!

What comes next...

- The mechanics of estimation, hypothesis testing, and confidence intervals should be familiar
- These concepts extend directly to regression and its variants
- Before turning to regression, however, we will review some of the underlying theory of estimation, hypothesis testing, and confidence intervals:
  - Why do these procedures work, and why use these rather than others?
  - We will review the intellectual foundations of statistics and econometrics