Introduction to Econometrics

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.
Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

In this lesson you will learn ...

- the main tenets of basic econometric theory
- hands-on experience with empirical problem sets
- methods to estimate causal effects using observational data
- methods to forecast time series data
- how to evaluate regression results
- how to interpret results from some influential empirical economics papers
- how to get started with some popular econometric software, including **R** and **Python**

What Is Econometrics?

- Econometrics is a set of quantitative methods used for a variety of purposes.
  - To estimate relationships among economic variables.
  - To test economic theories.
  - To evaluate business practice. For instance, to identify the components of demand and costs, including labor costs.
  - To evaluate government policy, assess the effectiveness of government programs, subsidies, understand the sources of revenue.
  - To forecast future trends and fluctuations.
  - To estimate responses to future changes in policy and environmental circumstances.
- Econometrics started as an applied branch of economics, but has since been applied to many other fields. In recent years, techniques developed in the context of machine learning are gradually being used in econometrics.
- Econometrics is concerned with causal inference — Machine learning may or may not be.

Econometricians

- Econometrics is specialized branch of statistics. Statistics originates in the seventeenth century. Econometrics is more recent, built on the foundations of modern statistics, in particular the work on inference of Ronald Fisher, Jerzy Neyman, Egon Pearson, and Abraham Wald.
- Early pioneers include Jan Tinbergen, Ragnar Frisch, who won the first Nobel prize in economics in 1969. The other Nobel prize winners in the fields are: Simon Kuznets (1971), known for his applied work on national income, economic growth, and inequality; Tjalling Koopmans (1975); Trygve Haavelmo (1989); James Heckman, Daniel McFadden (2000); Robert Engle, Clive Granger (2003); Thomas Sargent, Christopher Sims (2011); Lars Peter Hansen (2013); Joshua Angrist, David Card, Guido Imbens (2021).
- Other famous economists associated with the field of econometrics include (non-exhaustive list): Phillip and Sewell Wright; Halbert White; Peter Phillips; Bruce Hansen; Hashem Pesaran; Jeffrey Wooldridge; James Stock; Mark Watson; Whitney Newey; David Hendry; Jerry Hausman; James Poterba; Tim Bollerslev; George Tauchen; James Hamilton; Charles Manski; Pierre Perron; Francis Diebold; Zvi Griliches; Soren Johansen; Donald Rubin; Alan Krueger; Stephen Pischke; David Autor; Lawrence Katz; Richard Blundell; Stephen Bond; Manuel Arellano; John List; Christian Hansen; Victor Chernozhukov; Susan Athey ...

Why Stochastics?

- Economic theory is built around models designed to capture relationships among variables of interest.
- Economic relationships can be complex. Models are imperfect and leave out important factors. Relevant variables may not be observable or may be mismeasured. Theories may miss out relevant variables or may misrepresent the nature of the interactions among variables.
- Theories are progressively refined as economists learn more about the phenomenon, but imperfect they will remain. Many of these imperfections in modeling and measurement are unpredictable — if they were predictable, a model could be designed to explain them!
- Econometric models assume some degree of randomness in the relationship — even if the underlying economic model is deterministic.
- Stochastic processes are used to capture random departures from economic theory — These departures are caused by errors in measurement, errors made by economic actors, errors caused by imperfect theories.

Course Overview

- Economic theory suggests causal relationships and broad policy implications, but without measurement and quantification, economic theory lacks precision and real-world relevance. Quantitative methods — statistics — are needed. Examples:
- Returns to education:
  - What is the quantitative effect on student achievement of reducing class size by one student?
  - By how much does another year of education increase earnings?
  - How are these distributed among medical students, law students, economics students?
  - By how much does a Master's degree increase earnings above a Bachelor's degree?
- Taxing tobacco:
  - What is the price elasticity of cigarettes?
  - How does a one-dollar increase in cigarette taxes affect health outcomes?
  - By how much does a one-dollar increase in cigarette taxes reduce health expenditures?
- Monetary policy:
  - What is the effect on output growth of a one percentage point increase in the Fed's funds rate?
- House prices:
  - What is the effect on house prices of changes in zoning laws?

Course Overview

- This course is about using data to measure causal effects. For instance,
  - Returns to education.
  - Cigarette prices.
  - Monetary policy.
- Ideally, we would like an experiment.
  - what would be an experiment to estimate the effect of class size on standardized test scores?
- But most of the data we have is not derived from designed experiments.
- This course deals with challenges arising from using observational data to estimate causal effects:
  - Confounding effects (omitted factors).
  - Simultaneous causality.
  - Correlation does not imply causation!

Case Study: California Test Scores

- **Policy question:**

  What is the effect on test scores of reducing class size by one student per class?
- Variables:
  - fifth grade test scores:

    Stanford-9 achievement test, combined math and reading, district average.
  - Student-teacher ratio:

    No. of students in the district divided by No. of full-time equivalent teachers.

Case Study: California Test Scores

- Do districts with low STRs have higher test scores?
  1. **Estimate Relation:**
     Compare average test scores in districts with low STRs to districts with high STRs.
  2. **Test Hypothesis:**
     The null hypothesis that the mean test scores in the two types of districts are the same versus the alternative hypothesis that they differ.
  3. **confidence Interval:**
     Estimate an interval for the difference in the mean test scores, high-STR vs. low-STR.

Case Study: California Test Scores

Observation (District) Number	District Average Test Score (fifth grade)	Student–Teacher Ratio	Expenditure per Pupil ($)	Percentage of Students Learning English
1	690.8	17.89	$6385	0.0%
2	661.2	21.52	5099	4.6
3	643.6	18.70	5502	30.0
4	647.7	17.36	7102	0.0
5	640.8	18.67	5236	13.9
⋮	⋮	⋮	⋮	⋮
418	645.0	21.89	4403	24.3
419	672.2	20.20	4776	3.0
420	655.8	19.04	5993	5.0

Case Study: California Test Scores

Student-Teacher Ratios and Fifth-Grade Test Scores for 420 K-8 California Districts, 1999:

Distribution			Percentile						
	Average	Standard Deviation	10%	25%	40%	median 50%	60%	75%	90%
ST Ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test Score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

Case Study: California Test Scores

- **Preliminary data analysis:**

  Compare districts with low STR (< 20) and high STR (≥ 20):

Class Size	Average Score $\overline{Y}$	Standard Deviation s_Y	Sample Size n
< 20	657.4	19.4	238
≥ 20	650.0	17.9	182
all	654.2	19.1	420

- Steps of analysis:
  1. Estimation of $\Delta =$ difference between group means.
  2. Test the hypothesis that $\Delta = 0$.
  3. Construct a confidence interval for Δ.

Case Study: California Test Scores

1. **Estimate:**

$$\overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}} = \frac{1}{n_{\text{LOW}}}\sum_{i=1}^{n_{\text{LOW}}} Y_i - \frac{1}{n_{\text{HIGH}}}\sum_{i=1}^{n_{\text{HIGH}}} Y_i$$

- Is there a large difference? Should parents and school committees care?
- Standard deviation across districts = 19.1
- Difference between 60th and 75th percentiles of test score distribution:

$$667.6 - 659.4 = 8.2$$

- **This is a big enough difference to matter for school reform discussions.**

Case Study: California Test Scores

2. **Test Hypothesis:**

Apply a Difference-in-Means test

$$\mathbf{H_0}: \qquad \overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}} = 0$$

- The t-statistic:

$$t = \frac{\overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}}}{\sqrt{\frac{s^2_{\text{LOW}}}{n_{\text{LOW}}} + \frac{s^2_{\text{HIGH}}}{n_{\text{HIGH}}}}}$$

where s^2 stands for the sample standard devation,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{Y}_i - \overline{Y}\right)^2$$

and n stands for the sample size, for each group $_{\text{HIGH}}$/$_{\text{LOW}}$.

- The term in the denominator is the standard error.

Case Study: California Test Scores

2. Difference-in-Means test

$$\begin{aligned} t = \frac{\overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}}}{\sqrt{\frac{s^2_{\text{LOW}}}{n_{\text{LOW}}} + \frac{s^2_{\text{HIGH}}}{n_{\text{HIGH}}}}} &= \frac{657.4 - 650.0}{\sqrt{\frac{(19.4)^2}{238} + \frac{(17.9)^2}{182}}} \\ &= \frac{7.4}{1.83} \\ &= 4.05 \end{aligned}$$

- $|t| > 1.96$, so we reject the null hypothesis at the 0.05 significance level (two-sided test).
- **We reject the hypothesis that the two means are equal.**

Case Study: California Test Scores

3. **Confidence Interval:**

- A 95% **two-sided** confidence interval for the difference between the means:

$$\begin{aligned} & \overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}} && \pm && 1.96 && \times && \text{SE}(\overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}}) \\ = \; & 7.4 && \pm && 1.96 && \times && 1.83 \\ = \; & (3.8, 11.0) \end{aligned}$$

- **Interpretation:**
  - $\Delta = \overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}}$
  - The 95% confidence interval for Δ does not include 0.
  - The hypothesis that $\Delta = 0$ is rejected at significance level 0.05.

Case Study: California Test Scores

3. **Confidence Interval:**

- A 95% **one-sided** confidence interval for the difference between the means:

$$
\begin{aligned}
& \overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}} && - && 1.64 && \times && \text{SE}(\overline{Y}_{\text{LOW}} - \overline{Y}_{\text{HIGH}}) \\
= & \ 7.4 && - && 1.64 && \times && 1.83 \\
= & \ (4.4, \infty)
\end{aligned}
$$

- **Interpretation:**
  - A greater student-teacher ratio cannot reduce students' test scores.
  - One-sided tests are more powerful.
  - The critical value for a one-sided 95% confidence interval is 1.64 instead of 1.96.

Data and causal effects

- Economists are interested in causal relations.
- Statistics establishes correlations.
- And correlation is not causation.
- In order to establish that one variable has a causal effect on another, other factors affecting the outcome must be held fixed.
- In natural sciences, scientists can use controlled experiments.
- Experiment are often impossible in economics (too costly and/or for ethical reasons)
- In chemistry, aerodynamics, computer science, experiences are the norm. In economics — however tempted some may be to do it — you cannot remove teachers from schools to test how that would affect students' test scores.
- Economists must rely on observational data — data that originates from the real world, with all its constraints and statistical "noise".

Example: Effect of Health Insurance On Health

- Question: what is the effect of health insurance coverage on health?
- Ideal experiment: randomly assign people so that some have health insurance and some don't — no matter their current health status and income. Monitor the situation, gather data, and compare their health status a few years later.
- Real world: Economists must rely on observational data.
- Observed data obtained from 2009 NHIS survey:

Group	Sample Size	Mean Health	Std.Dev.
Some insurance	8114	4.01	0.93
No insurance	1281	3.70	1.01

- Mean difference in health outcomes: $4.01 - 3.70 = 0.31$
- Is 0.31 the causal effect of health insurance on health?

Example: Effect of Health Insurance On Health

- Is the observed difference in mean health outcomes a measure of the causal effect of health insurance on health?
- No! People with insurance are very different from people without insurance, in ways that often will affect their health.
- They may differ in more than one way and, admittedly, in complicated and contradictory ways. For instance, people with health problems are more likely to want to be insured. But people with low incomes are more likely to have health problems, but also less likely to decide to pay for insurance.
- For instance, there are important differences in levels of income and education:

Group	Mean Education	Mean income
Some insurance	14.31	106,467
No insurance	11.56	45,656

Potential Outcomes

- **Potential outcomes:** Powerful way of thinking about causality — *aka* the Rubin causal model — named after Donald Rubin.
- Imagine two alternative worlds, each exhibiting a particular outcome, one where the "treatment" is applied and one where it isn't.
- In a controlled experiment, the treatment would be applied at random, in order to control for factors that would influence the choice of treatment. In the real world, the treatment is almost never, strictly speaking, applied at random. But sometimes the manner of the treatment is "quasi-random" — not completely random, but partly random, with the randomness identifiable in the data.
- To estimate the effect of the treatment on the treated, the scientist would compare the outcome for one individual in alternative futures. But because we can observe only one potential outcome, for a given individual, the other potential outcomes of interest are missing.
- **Fundamental problem of causal inference:** Potential outcomes of interest are not observed.

Potential Outcomes

- Let $D_i = 1$ if person i has health insurance, 0 otherwise
- Let Y_{i0} and Y_{i1} be the **potential outcomes**
  - Y_{i0} = health if person i does not have health insurance
  - Y_{i1} = health if person i has health insurance
- We observe one of two potential outcomes:

$$Y_i = \begin{cases} Y_{i0} & \text{if } D_i = 0 \\ Y_{i1} & \text{if } D_i = 1 \end{cases}$$

- We can write

$$\begin{aligned} Y_i &= (1 - D_i)Y_{i0} + D_i Y_{i1} \\ &= Y_{i0} + D_i \times (Y_{i1} - Y_{i0}) \end{aligned}$$

where $(Y_{i1} - Y_{i0})$ is the treatment effect of health insurance on individual i.

Potential Outcomes

- We do not observe individual i's treatment effect. But, we do observe the average taken over a group of individuals — The expected difference:
- **Observed difference in average outcomes:**

$$\begin{aligned} &\mathrm{E}[Y_{i1}|D_i = 1] - \mathrm{E}[Y_{i0}|D_i = 0] \\ =&\mathrm{E}[Y_{i1} - Y_{i0}|D_i = 1] + \mathrm{E}[Y_{i0}|D_i = 1] - \mathrm{E}[Y_{i0}|D_i = 0] \end{aligned}$$

- **Average treatment effect on the treated:**

$$\mathrm{E}[Y_{i1} - Y_{i0}|D_i = 1]$$

- **Selection bias:**

$$\mathrm{E}[Y_{i0}|D_i = 1] - \mathrm{E}[Y_{i0}|D_i = 0]$$

- Under random assignment of the treatment, D_i (the treatment) and Y_i (the outcome) are independent, implying

$$\mathrm{E}[Y_{i0}|D_i = 1] = \mathrm{E}[Y_{i0}|D_i = 0]$$

- **Random assignment eliminates selection bias!**

Potential Outcomes

- In practice, selection bias is a real problem. Higher income individuals are more likely to be healthier and are also more likely to buy health insurance. On the other hand, unhealthy individuals would be more willing to purchase health insurance, if they could afford it.
- The "treatment" is clearly not random.
- Empirical economists seek to identify situations where observation data can be interpreted as experimental data — a situation called a "quasi-experiment".
- Sometimes economic theory can be used for inference. Since hospitalization is costly — both in terms of time and money — individuals who choose hospitalization do so because the expected improvement in their health outcome is greater than the cost.
- Since the benefit of the treatment decreases in Y_{i0}, it follows that:

$$\mathrm{E}[Y_{i0}|D_i = 0] \geq \mathrm{E}[Y_{i0}|D_i = 1]$$

- The observed difference in average health outcomes is a lower bound for the average treatment on the treated (ATT). And that is a valuable lower bound to estimate!

Famous Empirical Studies

- **Return to education:**

$$\log(\text{Wage}) = \alpha + \beta \cdot \text{Years of Schooling} + U$$

The main challenge with this class of regression model is the “**omitted variable bias**”: Ignoring the effect of systematic factors such as ability can cause the regression model to overestimate the effect of education on wages. This model is often called a “Mincer regression”, named after Jacob Mincer.

- **Effect of minimum wage and unemployment:**

$$\text{Unemployment} = \alpha + \beta \cdot \text{Minium Wage} + U$$

The main challenge with this class of regression model is “**reverse causality**“: High employment may lead to political pressure to raise the minimum wage. In other words, there is a two-way causality. This model is associated with the work of David Card and Alan Krueger.

- In these regression models, U is a random catch-all term. The more U is distributed like a mean-zero normal variate, the easier it is to make inferences.

Famous Empirical Studies

- **Effect of policing on crime:**

$$\text{Number of Crimes} = \alpha + \beta \cdot \text{Size of Police Force} + U$$

The challenge with this class of regression model is “**spurious correlation**”: Cities with a lot of criminal activity have a bigger police force. The correlation can spuriously indicate that the size of the police force has a positive effect on the crime rate.

- **Impact of MTV show “16 and Pregnant” on teen pregnancy:**

$$\log(\text{Teen Pregnancy}) = \alpha + \beta \cdot \text{Show's Rating Among Teens} + U$$

The challenge with this class of regression model is “**self selection**”: Teens who would be adverse to getting pregnant could be more likely to watch the show. See the 2015 article “Media Influences on Social Outcomes: The Impact of MTV’s 16 and Pregnant on Teen Childbearing” by Melissa S. Kearney and Phillip B. Levine.

Summary

- Many decisions in business and economics require quantitative estimates of how a change in one variable affects another variable.
- The ideal way to estimate a causal effect is a randomized controlled experiment, but performing experiments in economic applications can be unethical, impractical, or too expensive.
- Econometrics provides tools for estimating causal effects using observational data and exploiting accidental quasi-experiments.
- Econometrics also provides tools for predicting the value of a variable of interest using information contained in other, related variables.
- **Cross-sectional data:** generated by observing multiple units at a single point in time.
- **Time series data:** generated by observing a single unit at multiple points in time.
- **Panel data:** generated by observing multiple units at multiple points in time.

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 1, Exercise 3.

You are asked to study the causal effect of hours spent on employee training (measured in hours per worker per week) in a manufacturing plant on the productivity of its workers (output per worker per hour). Describe:

- an ideal randomized controlled experiment to measure this causal effect;
- an observational cross-sectional data set with which you could study this effect;
- an observational time series data set for studying this effect; and
- an observational panel data set for studying this effect.

Get Data, Compute, Plot.

Get quarterly data on the GDP for the United States, going back to 1960 and up to the latest available observation.

1. Compute the growth rate for each quarter.
2. Compute the annualized growth rate for each quarter.
3. Plot the two series on the same axes in blue and red.
4. Write a proper citation for the data source.
5. Describe the data you have selected.
6. What software did you use?
7. What computation did you make to obtain an annualized growth rate?
8. What value do you get for 1960:Q3?
9. What was the biggest challenge?

Keywords

randomized controlled experiment; control group; treatment group; causal effect; prediction; forecast; experimental data; observational data; cross-sectional data; observation number; time series data; panel data; longitudinal data;