Introduction

Statistics and Econometrics

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Outline (Wooldridge, Chap. 1.1,1.3,1.4)

- Why statistics and econometrics?
- Course info
- Data structures and roadmap

Outline

- Why statistics and econometrics?
- Course info
- Data structures and roadmap

Statistics and Econometrics

- Statistics
  - Collection, analysis, interpretation, presentation, and organization of data

What is econometrics?

Goals of Statistical Analysis

- Goal of many statistical/machine learning analysis is prediction!

Goals of Statistical Analysis

- Goal of many statistical/machine learning analysis is prediction!
- Example: An Infamous Retail Tale

- The second-largest discount retailer in the US
- The first store opened in 1962
- Annual revenue $69.495 billion (2016)

"My daughter is still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

Goal: Expand sales by targeting consumers going through major life events, like pregnancy

Goals of Statistical Analysis

- Goal of many statistical/machine learning analysis is prediction!
- However, in many cases,

$$\text{prediction} \neq \text{decision making}$$

Goals of Statistical Analysis

- In many cases, there are gaps between prediction and making a decision
  - Example: customer "churn" - a customer abandons a company or service
    - Predict the probability of churn, and allocate interventions to those at the highest risk
    - A recent research in Journal of Marketing Research shows

> of churning. I consistently find that customers identified as being at the highest risk of churning are not necessarily the best targets for proactive churn programs. In particular, I find that the overlap between the group of customers with the highest sensitivity to the retention efforts and those with the highest risk of churn is approximately 50%; thus, the relationship

Source: E. Ascarza, 2018, Retention futility: Targeting high-risk customers might be ineffective, JMR

Goals of Statistical Analysis

- In many cases, there are gaps between prediction and making a decision
- In order to make a decision, we need to understand its causal effect

Definition (Causal Effect)

Holding all other relevant factors constant, how does variable y change if variable x changes?

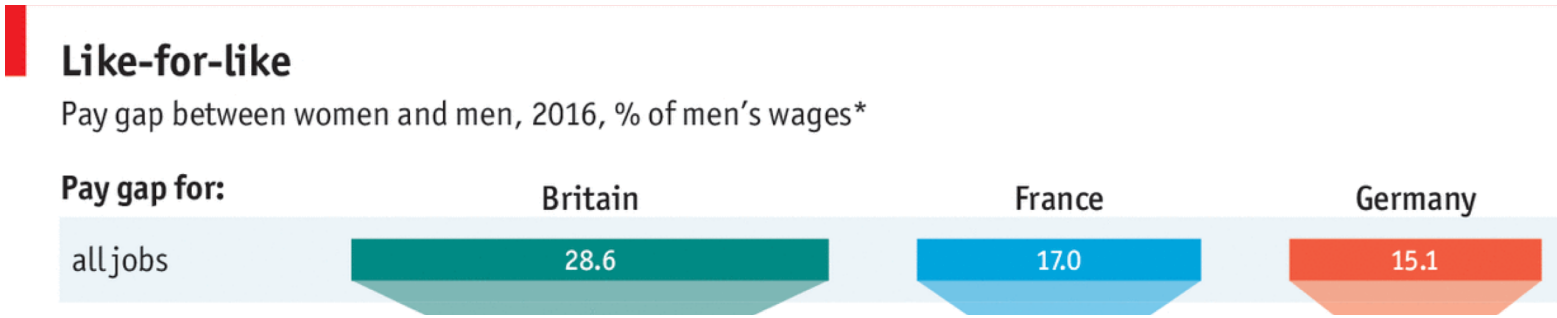
- Example: customer churn management
  - The causal effect of interventions on the probability of customer churn

What is Econometrics?

- Understanding casual effect of one variable on another
  - Focus of this course!
  - Sometimes, it can be "easy" if you have experimental data
    - Example: testing the effect of a new drug, A/B testing
  - Most cases, ideal "laboratories" are not available
    - Thus causality can be difficult to establish, as it is often not feasible to literally hold "all else equal"

Example: Gender Pay Gap

- Question: Does the pay gap below truly reflect the causal effect of gender?



Source: The Economist, 2017, Are women paid less than men for the same work?

This course

We will discuss how to construct linear regression models properly to infer causal effect of one variable on another, using non-experimental data

Outline

- Why statistics and econometrics?
- Course info
- Data structures and roadmap

Course Information

- Schedule
  - Two 1.5-hr multi-modal sessions per week
  - One 1-hr Zoom session/online self-study materials per week
    - Self-study materials for Week 1 & 2
    - Zoom sessions for Week 3, 4, & 5
  - Daily 1-hr office hour hosted by TAs
    - Running from 29 Oct. to 1 Dec. (except for Sundays)

Course Information

- Required textbook
  - Wooldridge, J.M. (2020), Introductory Econometrics: A Modern Approach, 7th Edition
- Assessment
  - Individual Assignments: 30%
    - One problem set each week
    - Mix of conceptual/theoretical and R coding questions
    - Will post online every Thursday, and due the next Wednesday
  - Final exam: 70%
    - 18 Dec 2020 (Friday), 10am-12pm
    - Four conceptual/theoretical questions; NO coding component
    - Close book, and 1-page formula sheet will be provided

Course Information

- Software for this course: R
  - Assignments are expected to be submitted in PDF or HTML compiled from a R markdown/notebook file
  - This is more of a methodology class than a coding class
  - Take advantage of the free access to Datacamp
- Extra weekly practise questions

Outline

- Why statistics and econometrics?
- Course info
- Data structures and roadmap

Data Structures

- Different data structures may require different methods
- Three major data structures
  - Cross-sectional data
  - Time series data
  - Panel data

Cross-Sectional Data

Cross-Sectional Data

- Each observation is a new individual, firm, etc.
  - Information collected at the same point in time
  - Minor timing differences usually ignored
- The order of observations does not matter (the index assigned to each observation is immaterial)
- Random sampling (observations are independent of each other) is desirable
  - Eg. Randomly draw 100 families from UK households and record income and other characteristics
  - What if wealthy families tend to decline to report?

Cross-Sectional Data

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Time Series Data

Time Series Data

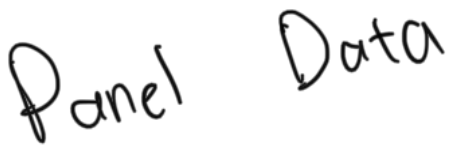
- Observations on variables are collected over time
  - Eg. stock prices, inflation, GDP, ...
- Chronological ordering of observations is important
  - Variables (eg. stock prices) tend to be related to their histories
  - Dependence in observations needs to be accounted for in models

Time Series Data

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Panel Data



- Observations follow the same units (individuals, families, firms, ...) over time
  - A time series for each cross-sectional unit
  - Eg. Data on individuals wage, education, union membership over 5 years
- Panel data is more difficult and expensive to obtain than cross sectional data
  - But has advantages in controlling unobserved factors and studying dynamic behavior

Panel Data

Example: Crime rates in 150 US cities: 1986 and 1990.
Data are stored by city and year.

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Roadmap

- Regression analysis with cross-sectional data
  - Basics: estimation, inference, analysis with dummy variables
  - More involved: model specification and data issues
- Advanced topics
  - Binary dependent variable models
  - Panel data analysis
  - Time series analysis*

*: Will be discussed in the Logistics and Supply Chain Analytics Module