

# Introduction to Econometrics

## *Lecture 0: Introduction*

**Zhaopeng Qu**

**Business School, Nanjing University**

Feb.16,2023



# Outlines

- 1 Introduction: A Scientific Framework of Rational Knowledge
- 2 What is Econometrics?
- 3 Why and who should take the course?
- 4 Logistics
- 5 An Introduction to Economic Data
- 6 Homework(not required)

# Introduction: A Scientific Framework of Rational Knowledge

# A Complex and Dynamic World



- Living in an unprecedentedly complex and dynamic world, we need **rational knowledge(理性认知)** and **scientific prediction(科学预测)**.
- So far, economics you have learned suggests many important relationships, often with policy implications, but ***seldom suggests quantitative magnitudes of these relationships.***

# Question #1: Student's Performance and Class Size



- A Classical Issue in Economics of Education
  - Is there a gap of students' performance between large-size classes and small-size classes?
  - Turn it into an empirical or policy question: What is the quantitative effect of reducing class size on student achievement?
  - Like by 5 student per class? or 10?

## Question #2: Discrimination in the Market?

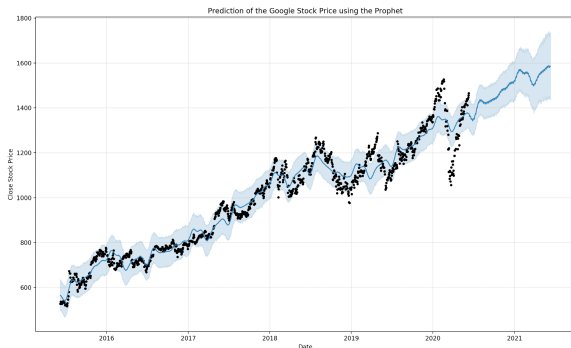


- Discrimination is the situation where people who are presumed to be equally productive or have equally productive capacity, get treated differently by the market.

## Question #2: Discrimination in the Market?

- Because of their gender, their religion, sexual orientation or the color of their skin, whatever even though these characteristics are irrelevant for the purpose being considered.
  - **Racial Discrimination**
  - **Gender Discrimination**
  - **Hukou Discrimination**
- To see if there is a difference in the probability of being denied **for otherwise identical** applicants and, if so, whether this difference is large or small.

# Question #3 How Will Stock Market Go?



- It seems that people always want a sneak preview of the future.
  - Will the stock market go up next month, and, if it does, by how much?
- A forecaster's job is to predict the future by using the past which means fortune.



## Other Similar Questions

- Air pollution and Health?
- Credit regulation on housing price
- Coupon on products sales
- Trade War...
- Pandemic...

*Living in such a messy world, as an educated person, we need **rational knowledge** or **scientific methods** to make decisions.*

# How can we obtain rational knowledge(judgment)?

- Anecdotes(轶事) or Intuition(直觉)
- Theory(理论/逻辑推理)
  - Systematical methodology: Hypothesis, Logical deduction...
- Empirics (数据实证)
  - statistical inference from data.

# An Example: Smoke and Mortality

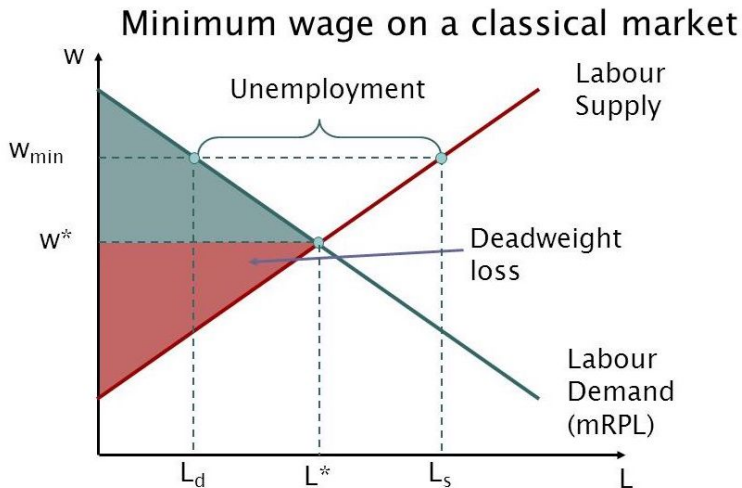
- Anecdotes(轶事) or Intuition(直觉)
  - eg. “My grandmother smoked two packs a day and lived until she was 95 years old.”
- Theory
  - Because Cigarettes contain carcinogens(致癌物) such as nicotine, tar, and formaldehyde(尼古丁、焦油、甲醛等), then...
- Empirics
  - Collecting data through experiments or surveys, and then use statistical or econometrical methods to verify whether and how cigarettes can harm our health.

# Classical Question: Human Capital v.s Signal

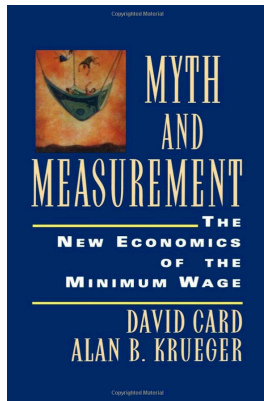
- A common phenomenon in labor markets can be observed across countries.
  - **Higher education, Better pay!**
- Two classical theory to explain
  - **Human Capital:** Education improves work productivity.
  - **Signal:** Education does not increase the productivity. It simply serves as a signal of the individuals' innate ability.
- **Question: which one is right?**

# Public Policy Debate: Minimum Wage and Unemployment

- A classical supply-demand model tell us



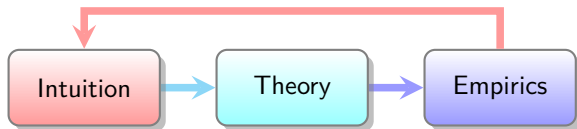
# Public Policy Debate: Minimum Wage and Unemployment



- One famous empirical evidence challenged the theory by David Card and Alan Krueger(1994)
- They found that increases in the minimum wage do **NOT** lead to job losses.

# A Scientific Workflow to Analyzing

- By Intuition: Propose meaningful or interesting questions (It does matter or we care about)
- By Theory: Obtain a preliminary conclusion or proposal an hypothesis
- By Empirics: use data and quantitative methods to test your theory or conclusion.



- *Once we have a theory (or cause) which has been testified by empirical works, then we can manipulate the cause to obtain the effect.*

# Theory, Empirics and Math

- Economic theories sole are not enough to explain social phenomenon.
  - It should be tested by empirical evidence.
  - When having competing theories, we need more test to justify which one is more likely right.
- When mechanism is clear, math is unnecessary, but empirical evidence is unexpendable.
  - Having a mathematical model is better, but sometimes not necessary...



# Quantitative Answers to Quantitative Questions

- Many decisions in economics, business and government hinge on understanding the relationship among variables in the world around us.
  - Economic theory may provide clues about the direction of the answer.
  - But decisions require **quantitative answers to quantitative questions**.
- Therefore, we have develop a framework and find a practical method that provide
  - a numerical answer to the question
  - a measure of how precise the answer is.
- *It is the job of **Econometrics***

# What is Econometrics?

# Introduction: Econometrics

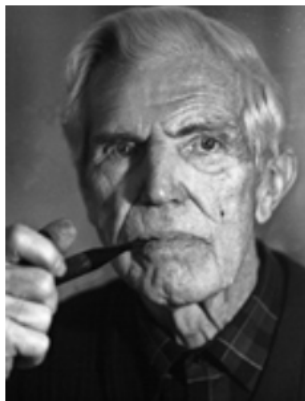


- The term is attributed to
  - **Ragnar Frisch(1895-1973);**
  - **1969 Nobel Prize co-winner**  
(the first year for Economics)
- Although the term coins by a combination of economics( “Econ-” ) and metrology( “Metrics” ), it is special enough in social science and science at that time.

# Introduction: Econometrics

- “Econometrics is by no means the same as *economic statistics*. Nor is it identical with what we call general *economic theory*, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous with the application of *mathematics to economics*. Experience has shown that each of these three view-points, that of *statistics, economic theory, and mathematics*, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. *And it is this unification that constitutes econometrics.*”
  - (Ragnar Frisch, *Econometrica*, 1933, volume 1, pages. 1-2)

# Introduction: Econometrics



- ***Trygve Haavelmo(1911-1999)***
- **1989 Nobel Prize winner**
- “The method of econometric research aims, essentially, at a conjunction of **economic theory** and **actual measurements**, using the theory and technique of **statistical inference** as a bridge pier.”  
(*Econometrica*, 1944, volume 12, pages. 1-2)

# Introduction: Econometrics



- *James Stock and Mark Watson(2014)*
  - “Ask a half dozen econometricians what econometrics is—you could get a half dozen different answers.”
  - “At a broad level, it is a **science and art** of using **economic theory** and **statistical techniques** to analyze **economic data**.”

# Introduction: Econometrics

- My View: **In general, a series of scientific methods to searching for economic logic from data.**
- It could include two jobs
  - *Making a causal inference*, such as
    - Testing economic theories.
    - Estimating causal effects.
    - Using data to give policy recommendations.
  - *Forecasting or predicting* future values
- More and more prevalence in
  - other social science such as *political science, sociology, legal and education studies* etc
  - and business practice, like the hottest one: **Big Data or Data Science.**

# Methodological Revolutions in Social Science

- Social science (firstly started by Economics) is experiencing **two methodological revolutions** over the past few decades.
- On the one hand, there is the **Credibility Revolution**
  - A movement that emphasizes the goal of obtaining secure **causal inference** (Angrist and Pischke, 2010)
- On the other hand, there is the **Big Data revolution**
  - A movement that emphasizes that how our increasing ability to produce, collect, store and analyze vast amounts of data is going to transform our understanding of the human affairs. (Schonberger, 2013)
- Obviously, Econometrics are playing **key roles** in both two revolutions.



# Credibility Revolution in Social Science

- Before the revolution, researchers frequently relied on attempts to statistically model the world to make inferences from **observational data**.
  - In essence, they would rely heavily on **ex post** statistical analysis to make causal inferences. Now we have acknowledged that this is not a “real” causal inference.
- The revolution centered around the idea that the only way to truly account for possible sources of bias is to remove the influence of all confounders **ex ante through better research design**.
  - Thus, since the revolution, researchers have attempted to design studies around **sources of random or as-if random variation**, either with **experiments** or what have become known as “**quasi-experimental**” designs.
- The revolution started from around **1990s pioneering in economics** then spread over other empirical social sciences such as sociology, political science, education, legal studies, etc, which has entirely changed empirical social science.

# Big Data Revolution in Social Science

- The name of Big Data comes from computer scientists working to do aggregation on data that is too big to fit on a single machine.
- The Big Data presents exciting opportunities for the study of social science, but at the same time its size and heterogeneity present significant challenges.
- Big Data analytic tools are highly correlated with econometrics
  - partially straight out of previous statistics and econometrics classes(Cluster, Regression and Matching)
  - some are totally new(such as decision trees and neural networks).

# Big Data, Data Science and Econometrics

- So many labels for what we do...
  - Econometrics
  - Data Mining/Big Data/Data Science
  - Machine Learning(ML)
  - Artificial Intelligence(AI)
- Along this spectrum, you move from heavy focus on what things you are measuring (what real phenomena they correspond to) to a more practical 'useful is true' pattern discovery approach.
- *The similarities are much bigger than any distinctions.*

# Econometrics: sub-fields

## • Theoretical Econometrics

- It is concerned with methods, both their properties and developing new ones.
- It is closely related to mathematical statistics, and it states assumptions of a particular method, its properties etc.
- We could call *theoretical econometricians* as the **producer** of econometrics.

## • Applied Econometrics

- More orientated to applied work, such as choice of technique and interpretation of research finding.
- But it should be also based on a **solid conceptual foundation** and some **practical experiences** plus a little bit **skills of computer**.
- Most of us are the **consumers** of econometrics.

# Wrap Up

- Econometrics is a collection of a series of scientific methods to searching for economic logics from data.
- It could include two jobs: *causal inference and prediction*.
- It plays key roles in revolutions of both social science and business practices.

# Why and who should take the course?

# Why Econometrics is so important?

- Several Common Questions about Econometrics?
  - Why we should study econometrics?
  - How is studying econometrics helpful in understanding social science?
  - Especially, can one excel in the economics without learning econometrics?
- The answer to the last question is simple.
  - ***NO!*** because one *hardly* to learn modern economics without the knowledge of econometrics.

# Why Econometrics is so important?

- Econometrics is one of **three core courses** required in almost every economics department worldwide.
- Econometrics is also an important research field
  - Many **Noble Prize winners**

Year	Name	Country
1969	Ragnar Frisch	Norway
1969	Jan Tinbergen	Netherlands
1973	Wassily Leontief	U.S.
1980	Lawrence R. Klein	U.S.
1989	Trygve Haavelmo	U.S.
2000	James.J Heckman	U.S.
2000	Daniel L. McFadden	U.S.
2003	Robert F. Engle	U.S.
2003	Clive W.J. Granger	U.S.



# 2021 Noble Prize in Economics

- Last year, I predicted that “**Joshua Angrist, David Card and Guido Imbens could be winners soon.**”

## The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021



© Nobel Prize Outreach. Photo:  
Paul Kennedy



© Nobel Prize Outreach. Photo:  
Risdon Photography

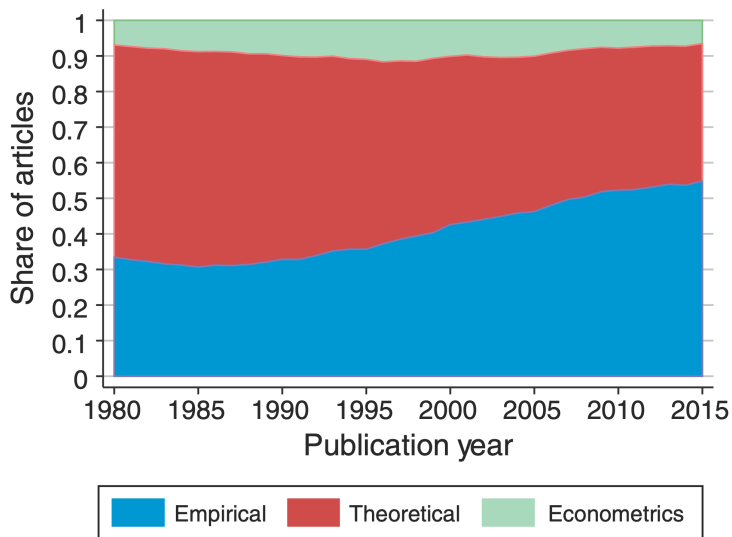


© Nobel Prize Outreach. Photo:  
Paul Kennedy

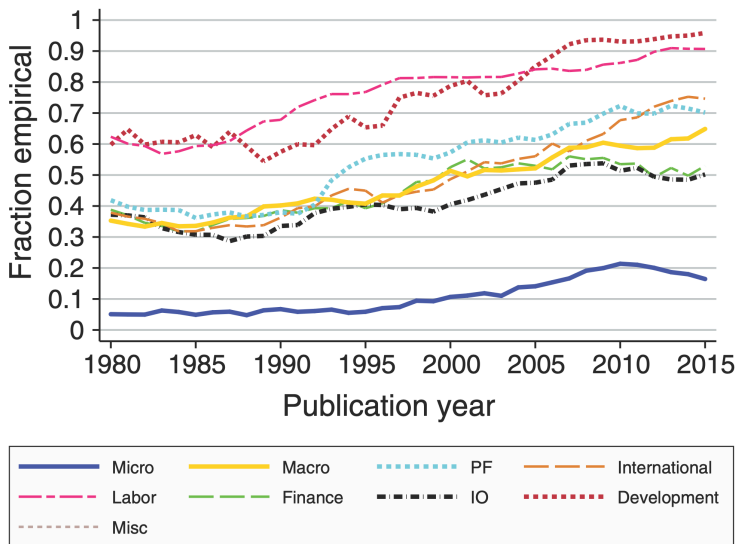
# Who pursue an academic career

- **Congratulations!** You are in one of the most promising and internationalizing areas in economic research of China.
  - Master econometrics knowledge will improve your research greatly.
  - Overwhelming majority of top journal articles are quantitative.

# The Increasing Proportion of Empirics (Angrist et al, 2017)



# The Increasing Proportion of Empirics (Angrist et al, 2017)



# Who enter industry job market

- Who want to work in industry: mastering econometrics can **help you get a good job!**
  - A lot of internet giants even hire economists to lead their special R&D department. Such as
    - Google
    - Microsoft
    - eBay
    - Baidu
    - Alibaba
    - Tencent
- Credit cards, POS terminals and ERP systems widely used in supermarkets, banks and factories also revolutionize the business (or management) practices in marketing, accounting, management operation etc.
- **Data Scientist** is the hottest job in consulting, business areas as well as financial industry.

# Who enter industry job market: Apple Job Wanted

## Economist/Core Data Scientist

Apple · Beijing, Beijing, China

Apply

Save

...

### Key Qualifications

Strong background in statistics or econometrics, regression analysis, causal inference, time series analysis, GLM, logistic regression, probability theory, regularization, interest in machine learning algorithms.

Develop internal visualization and modeling tools to facilitate data-driven decisions

Present results and other analytical findings to business partners

Strong statistical background and experience with causal inference, time series analysis (e.g. ARIMA, exponential smoothing, time series regression methods etc.), forecasting, and data analysis

Experienced R/Python programmer also proficient in other languages important to the ETL data pipeline (e.g. SQL)

Experience with data visualization packages (e.g. ggplot2, plotly) and advancing multiple projects at once on a tight schedule

Ability to share results with a non-technical audience

Experience in bayesian statistics and modeling (e.g. bayesian structural time series, dynamic linear models)

Advocate and practitioner of version control and reproducible code

Excellent verbal and written communication skills, in both Mandarin Chinese and English

### Description

- Work with various teams to understand business problems and develop data-driven solutions
- Build models to causal impact of new programs released
- Develop internal visualization and modeling tools to facilitate data-driven decisions
- Present results and other analytical findings to business partners

### Education & Experience

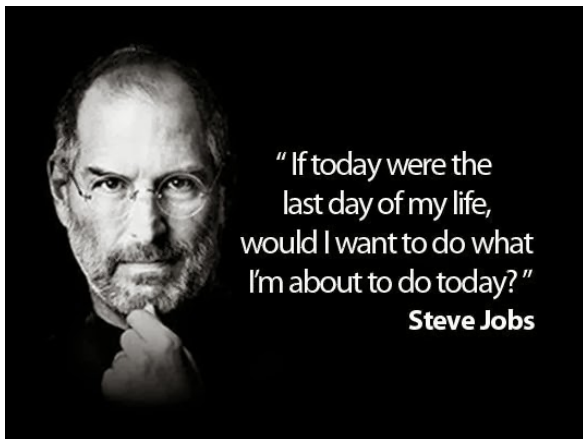
- PhD in Economics or related fields
- M.S. in related field with 5+ years experience applying statistical methods to business problems.

# Who look for fun

- Introduction to Econometrics could not be a boring and demanding variant of a mathematics course, but a **interesting and having fun** class.
- You can just enjoy it by thinking in an empiricist's way in your daily life.
  - have novel ideas or new perspectives about our world.
  - Econometrics is kind of a bible or philosophy of empirical economists.
- We will cover many very interesting stories
  - Interesting Examples
    - Eg. **Crime and Abortion** in *Freakonomics* written by Steven Levitt.
    - Eg. What is **the value to be the president's son(or daughter)**? in *Economic Gangster* written by Raymond Fisman and Edward Miguel.

# Whoever and Whatever

- **Whoever you would like to be or whatever you want**

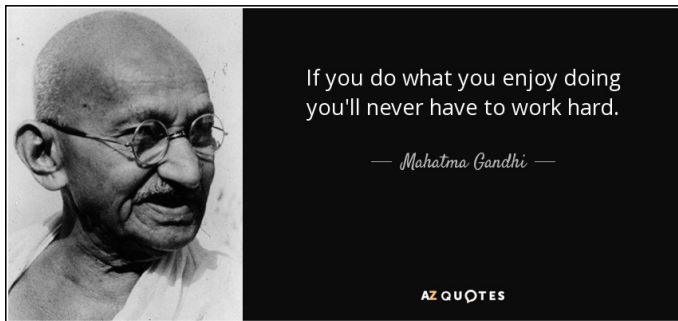


- *Every choice you make has an opportunity cost. Try your best to make a wise one!*



# Whoever and Whatever

- **Whoever you would like to be or whatever you want**



- *Enjoy doing something seriously and cultivate a special quality for yourself!*

# Hard and Soft Skills

- You **should** learn or improve several important skills during college.
  - **Hard Skills**
    - Language
    - Computer
  - **Soft Skills**
    - Critical Thinking
    - Presentation and Writing
    - Teamwork
- *Fortunately, you will learn/practice **almost all above skills** in this course*

# Conclusion

- In a word, *Introduction to Econometrics* is a **very important** and **interesting** course,
  - **Please think over before your take it!**
  - **Once you take it, please work hard on it!**
  - **And please enjoy working hard on it!**

# Logistics

# About Me, our TA and the Course

- My name is *Zhaopeng Qu*(曲兆鹏) <https://byelenin.github.io/>
  - Position and Affiliation: *Associate Professor, Institute of Population Studies, Business School.*
  - Research Fields: *Labor Economics and Applied Econometrics*
  - Office: Room 2017, Anzhong Building
  - Office Hour: Make an appointment in advance
  - Tel: 83621349
  - Email: [qu@nju.edu.cn](mailto:qu@nju.edu.cn)
- TAs: *Jiayi Cheng*(成佳怡)
  - 1st grade graduate students
  - good at *Stata*
  - Email: [intrometrics@gmail.com](mailto:intrometrics@gmail.com)(only for homework)
- Online resources
  - Our Course Website: [https://byelenin.github.io/Metrics\\_2023/](https://byelenin.github.io/Metrics_2023/)
  - Wechat group: discuss anything about the course if you want
- Other characteristics
  - 本科生院评定“千层次课程”(2019-)
  - 国际处评定“国际化课程”(2020-)

# The Procedure:

- **The First Part-Lectures by the instuctor**
  - Introduce the underlying theoretical problems briefly and focus on the empirical strategy heavily.
  - May provide some specific examples in classical papers with interesting topics in our fields.
- Everyone who take the course is **required attending the class.**

# The Procedure:

- **The Second Part: Recitation Sessions and Computer Labs by our TA**
  - Review some basic concepts and formulas in probability theory and statistics.
  - Review the homeworks if necessary.
  - Teach you basic skills of Stata.
- They are **optional** sessions that students are **NOT required** to attend.

# Prerequisite

- I assume that you **should** be comfortable some basic concepts of **probability theory** and **statistics**, such as
  - Random variable
  - Expectation, variance and covariance
  - Probability density function, p.d.f. and cumulative distribution function, c.d.f
  - L.L.N and C.L.T
  - Estimator
  - Unbiased and Consistent
  - Asymptotic Normality
  - Hypothesis test
- I will not review these basic concepts and formulas in lectures, but **it will be reviewed by TA in the recite sections.** You should review it by yourself anyway.



# Reference Textbooks

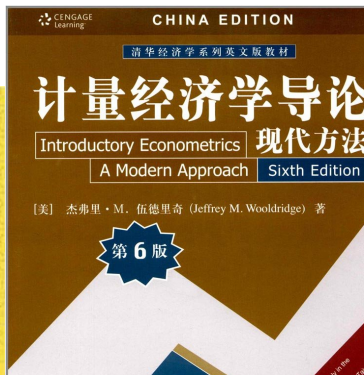
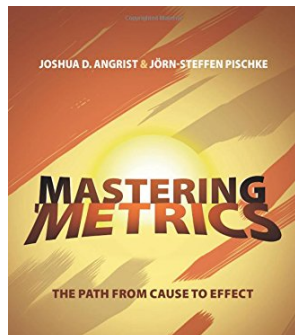
- Required Textbook: James H. Stock & Mark W. Watson, (2012). *Introduction to Econometrics*, 3rd Edition, Pearson Education.
  - 影印版：格致出版社 / 上海人民出版社。
  - 中文版：世纪出版集团 / 上海人民出版社。



# Reference Textbooks

- Supplementary textbook

- Joshua D. Angrist & Jorn-Steffen Pischke, (2014). *Mastering 'metrics: The Path from Cause to Effect*. Princeton University Press. (中译本:《精通计量:从原因到结果的探寻之路》,格致出版社,2018)
- Jeffrey M. Wooldridge,(2012). *Introductory Econometrics: A Modern Approach*, 5th or 6th Edition, South-Western College.



# Computing Tools

- The main computing tools used in the course are ***Stata*** and ***R***, optionally.
  - ***Stata*** is a complete, integrated statistical package that provides everything you need for data analysis, data management, and graphics.
  - ***R*** is a free software and useful for statistical study and econometrics.

# Reference Books for Software

- Stata Instruction
  - 任何一本中文或英文书。
- R Instruction
  - 任何一本中文或英文书。
  - too many resources online such as *Datacamp*

## Interesting Books for Reading

- Steven D. Levitt and Stephen J. Dubner, *SuperFreakonomics: Global Cooling, Patriotic Prostitutes, and Why Suicide Bombers Should Buy Life Insurance*, 2009. (中译本, 《超爆魔鬼经济学》, 斯蒂夫·列维特、斯蒂芬·都伯纳著, 中信出版社, 2010年1月。)
- Ian Ayres, *Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart*, 2007. (中译本《超级数字天才》, 伊恩·艾瑞斯著, 中国青年出版社, 2008年1月。)
- Raymond Fisman & Edward Miguel, *Economic Gangsters: Corruption, Violence, and the Poverty of Nations*, 2010. (中译本: 《经济黑帮: 腐败、暴力的经济学》, 中信出版社。)
- Abhijit V. Banerjee & Esther Duflo, *Poor Economics A Radical Rethinking of the Way to Fight Global Poverty*, 2011. (中译本: 《贫穷的本质: 我们为什么摆脱不了贫穷》, 中信出版社。)
- Angus Deaton, *The Great Escape: Health, Wealth, and the Origins of Inequality*, 2015. (中译本: 《逃离不平等: 健康、财富及不平等的起源》, 中信出版社。)

# How to learn

- Read required and related materials before lectures.
- Think about the content related with your personal knowledge and experience.
- Discuss with your classmates often.
- Whenever you have questions, ask them.
- Make good friends with data and Stata or R.
- Learning by doing is the best way to practice your skills.

# What I promise to offer you

- Prepare lectures as well as possible.
- One to one interaction on topics covered in the course, especially for your own topics.
- Help you start to using Stata or R to analyze some popular data sets in China.
- A good score?
  - It depends on you.

# What I expect to you

- Class participation with a little bit aggressive attitude.
  - More questions, more scores!
  - Interrupt me as often as necessary! (but I know most of you are not comfortable to this)
  - Got a dumb question? Please assume that you are the smartest person in the class, and you eventually will be!
- Finish your homework.
- Self-motived learn by doing.



# Two iron rules

- Don't ever cheat on assignments and exams.



# Two iron rules



- **Don't ever snitch your teachers to help political persecutions.**

# Evaluation

- Class Participation (10%)
- Homework(30%)
- Team Project: A research proposal(20%)
- Final Exam: (40%)

Welcome contact me and TAs



# An Introduction to Economic Data

# Introduction

- Data Structure
  - Cross-sectional data
  - Time series data
  - Pooled cross-sectional data
  - Panel data

# 1. Cross-Sectional Data: (Major Focus)

- Units: individuals, households, firms, cities, states, countries, etc.
- Data on *multiple* agents at a *single* point in time

$$\{x_i, y_i, \dots\}_{i=1}^N; N = \text{Sample Size}$$

- Usually obtained by random sampling from the underlying population. It means

$$\{x_i, y_i \perp x_j, y_j\}, i \neq j \in N$$

- Cross-sectional data are widely used in economics and other social sciences:
  - labor economics, public finance, industrial economics, urban economics, health economics...

# 1. Cross-Sectional Data: (Major Focus)

**TABLE 1.1** Selected Observations on Test Scores and Other Variables for California School Districts in 1999

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

*Note:* The California test score data set is described in Appendix 4.1.

- $x = STRatio$ ;  $y = TestScore$ ;  $N = 420$



## 2. Time Series Data:(Minor Cover)

- Observations on a variable (or several variables) over time, thus data on a *single* agent at *multiple* points in time

$$\{x_t, y_t \dots\}_{t=1}^T; T = \text{Sample Size}$$

- Examples:
  - stock prices
  - money supply
  - consumer price index(CPI)
  - gross domestic product(GDP)
  - automobile sales
- Economic observations can rarely be assumed to be independent across time. So we have to account for the dependent nature of economic time series.
  - Data frequency: minutes, hourly, daily, weekly, monthly, quarterly, annually.

## 2. Time Series Data:(Minor Cover)

**TABLE 1.2** Selected Observations on the Growth Rate of GDP and the Term Spread in the United States: Quarterly Data, 1960:Q1–2013:Q1

Observation Number	Date (year:quarter)	GDP Growth Rate (% at an annual rate)	Term Spread (% per year)
1	1960:Q1	8.8%	0.6%
2	1960:Q2	-1.5	1.3
3	1960:Q3	1.0	1.5
4	1960:Q4	-4.9	1.6
5	1961:Q1	2.7	1.4
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
211	2012:Q3	2.7	1.5
212	2012:Q4	0.1	1.6
213	2013:Q1	1.1	1.9

*Note:* The United States GDP and term spread data set is described in Appendix 14.1.

- $x_t = \text{Date}(\text{quarter})$ ;  $y_t = \text{GDP Growth Rate}$ ;  $N(T) = 213$

### 3. Panel(or Longitudinal) Data(Minor Cover)

- Time series for each cross-sectional member in the data set, thus data on multiple agents at multiple points in time.
- The same cross-sectional units (individuals, firms, countries, etc.) are followed over a given time period.

$$\{X_{it}, Y_{it}\dots\}_{i=1, t=1}^{NT}$$

- Advantages of panel data:
  - Controlling for (time-invariant) unobserved characteristics
  - Consideration of the effects of lag variables

### 3.Panel(or Longitudinal) Data(Minor Cover)

**TABLE 1.3** Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation Number	State	Year	Cigarette Sales (packs per capita)	Average Price per Pack (including taxes)	Total Taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	\$1.022	\$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
528	Wyoming	1995	112.2	1.585	0.360

*Note:* The cigarette consumption data set is described in Appendix 12.1.

•  $x_{it} = \text{Total Taxes}_{it}$ ;  $y_{it} = \text{Cigarette Sales}_{it}$ ;  $N \times T = 48 \times 11 = 528$

## 4.Pool(Repeat) Cross-Sectional Data(Not Cover)

- Pooled cross sections can be generated by combining **two or more** years cross-sectional data.
- Cross-sectional data in each year is independent with other years.
  - It means that data does not track the respondent multiple times.
  - But the data come from a same population in different time.
- For it has both cross-sectional and time series features, so allows consideration of changes in key variables over time.
- Simple pooling may also be used when the number of observations of a single cross section is small.
- It is widely used in:
  - Before-after comparisons of a government policy
  - Cohort studies
  - Cross-sectional analyses

## 4.Pool Cross-Sectional Data(Not Cover)

**TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices**

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

•  $x_{ijt} = hprice_{i,1993}, hprice_{j,1995}; y_{ijt} = proptex_{i,1993}, proptex_{j,1995};$

•  $N = N_i + N_j = 250 + 270 = 520$

# Corresponding sub-fields or sub-course

- Micro-Econometrics(微观计量经济学)
  - Cross-Sectional
  - Pool Cross-Sectional
  - Short Panel(large N, small T)
- Macro-Econometrics (宏观计量经济学)
  - Times series
  - Long Panel(small N, large T)

# Sources

- Traditional Collecting Way:
  - Survey(调查)
  - Administrative data(官方业务数据)
- Collecting Data in Digital Times:
  - Mass or Big data:
    - Online documents
    - Social Media
    - Geolocations or Geographic data
  - High Frequency Data
    - Stock, future or other financial tractional data



# Big Data: Volume

- Big in volume in terms of
  - the number of observations(size  $n$ )
  - the number of variables(dimension  $p$ )
- Eg. Wechat(微信) in 2018
  - 1.082 billion active users every month
  - 45 billion messages every day
  - 0.41 billion video calls every day

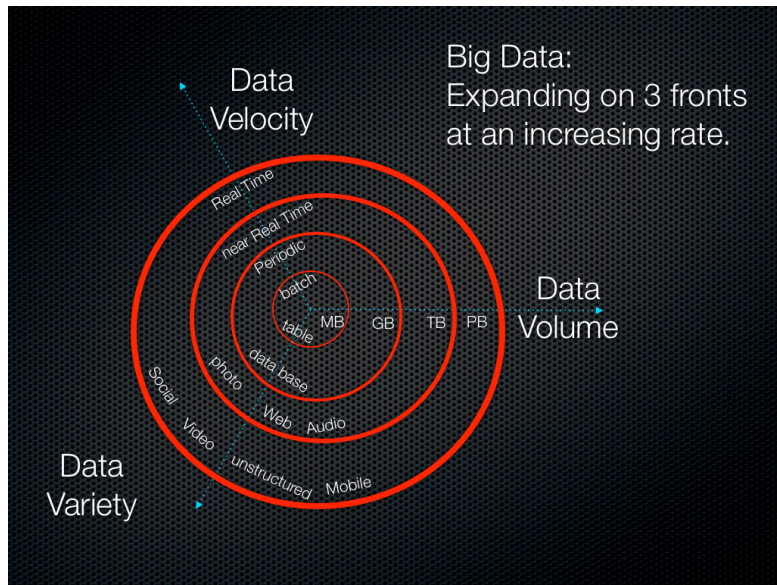
# Big Data: Variety

- Big in variety in terms of
  - classical data(numbers in tables)
  - photo/audio/video
  - text
  - map
  - sensors

# Big Data: Velocity

- Velocity is the measure of how fast the data is coming in.
  - Eg. Facebook(脸书) users upload more than 900 million photos a day in 2016
  - Eg. Wechat(微信) in 2018
    - 45 billion messages every day
    - 0.41 billion video calls every day

# Big Data: 3Vs



# Big Data means to analytics

- Big in the number of observations: size  $n$

$$n \rightarrow +\infty$$

- Big in the number of variables: dimension  $p$

$$p \rightarrow +\infty$$

- Tools in Big Data are
  - partially straight out of previous statistics and econometrics classes(Cluster, Regression and Matching)
  - some totally new(such as decision trees and neural networks)

# Data in China

## • Survey Data

- China Household Income Project(CHIP)
- China Family Panel Survey(CFPS)
- China Health and Retirement Longitudinal Study(CHARLS)

## • Administrative Data:

- Census: 全国人口普查数据; 全国 1% 人口抽样调查
- 工业企业数据库
- 海关交易数据库
- 工商企业登记数据库

## • Online Big Data:

- Shopping data on Taobao,JD,Tmall(淘宝、京东、天猫...)
- Movie Data on Douban.com(豆瓣电影数据)
- Air Quality: PM2.5(空气质量数据)
- Night-Lights Data(夜间灯光数据)
- Land Transaction Markets(土地交易市场数据)

# Homework(not required)

# Homework

- 到如下数据库网站选择其一登记注册

- China Household Income Project(CHIP): 中国居民收入调查
- China Health and Nutrition Survey(CHNS): 中国健康与营养调查
- China Family Panel Survey(CFPS): 中国家庭追踪调查
- China Health and Retirement Longitudinal Study(CHARLS): 中国健康养老追踪调查
- Chinese General Social Survey(CGSS): 中国综合社会调查
- China Labor-force Dynamics Survey(CLDS): 中国劳动力动态调查
- China Household Financial Survey(CHFS): 中国家庭金融调查



# Homework

- 了解抽样调查的目的和主要内容，以及抽样范围、方式、样本量等等基本信息，判断该数据属于哪种数据结构。
- 下载调查的问卷，详细了解调查有哪些具体的信息。
  - 首先确定自己感兴趣的问题，然后到问卷中寻找。
  - 或者先看问卷，找到自己感兴趣的具体信息。
  - 下载相应数据，进行初步的数据清理和统计分析（待上机课之后）
  - 为**研究项目**做准备。