Introduction to Econometrics, Spring 2026

Lecture 0: Introduction

Zhaopeng Qu
Business School,Nanjing University
March 05 2026

Today's Agenda

1. Introduction to Scientific Framework of Rational Knowledge
2. What is Econometrics? (Why we need econometrics)
3. Why Econometrics is so important? (Who should take the course)
4. Course Logistics(How to study the course)
5. Introduction to Economic Data

Introduction: A Scientific Framework of Rational Knowledge

Question #1: Student's Performance and Class Size

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

Question #2: Discrimination in Labor Market

- Many types of discrimination in labor market:
  - Racial Discrimination
  - Gender Discrimination
  - Hukou Discrimination
  - Age Discrimination

- **Question**. How to prove the existence of discrimination?
- What is **discrimination**?
  - *unequal pay regardless of job?*
  - *equal job, unequal pay?*
  - *equal ability, unequal pay?*
  - *equal productivity, unequal pay?*
- **Question: How to quantitatively measure the degree of discrimination?**
- **The Challenge**: If we observe women earning less, how do we know it is discrimination rather than differences in occupation, hours, or experience etc.?

Question #3: Cigarette Taxes and Smoking

- A major public health concern worldwide.
  - medical expenses of caring for those made sick by smoking.
  - nonsmokers breathe secondhand cigarette smoke.

- Basic economics states:

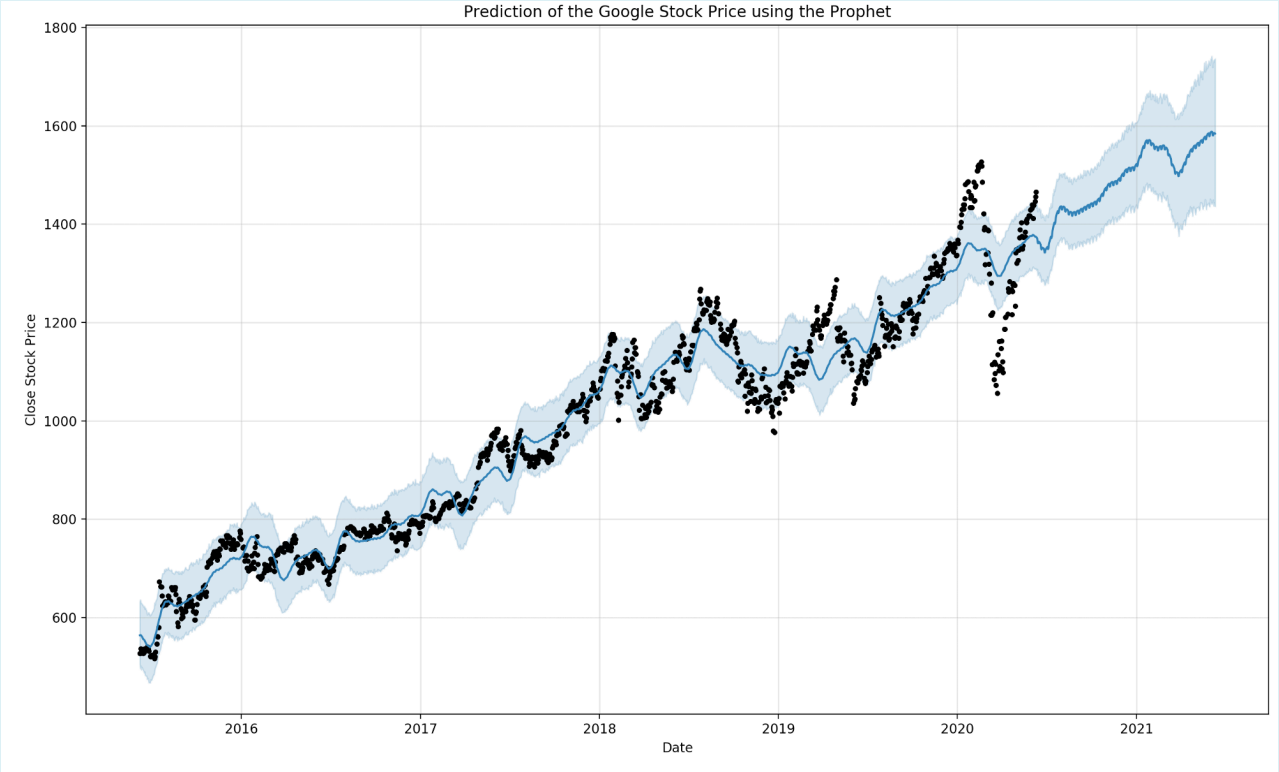
> **cigarette prices go up, consumption will go down.**

- But by how much?
  - *If the sales price goes up by 1%, by what percentage will the quantity of cigarettes sold decrease?*

- The percentage change in the quantity demanded resulting from a 1% increase in price is **the price elasticity of demand**.

- **Economic theory** can **rarely** provide numerical answers to such questions.

Question #4: How Will Stock Market Head?



- People have a natural desire to peek into the future.
  - *Will the stock market rise next month, and if so, by what magnitude?*
  - *What level of confidence can we place in our predictions?*

Quesitions Require Both Scientific Answers

- **Other Similar Questions**:
  - Air pollution and Health?
  - Credit regulation on housing price
  - Coupon on products sales
  - Trade War...
  - Pandemic...
- Living in an unprecedentedly complex and dynamic world, we need to make decisions based on
  - **Rational cognition**(理性认知)
  - **Scientific prediction**(科学预测)

A Scientific Framework for Rational Cognition

How to obtain rational knowledge(judgment)?

1. Anecdotes(轶事）or Intuition(直觉)
    - explore the world by our own experience or intuition.
2. Theory and Formalization(理论/逻辑推理）
    - formalization using systematical methodology: Hypothesis, Logical deduction...
3. Empirical Evidence(经验证据)
    - using data to make descriptions and statistical inference.

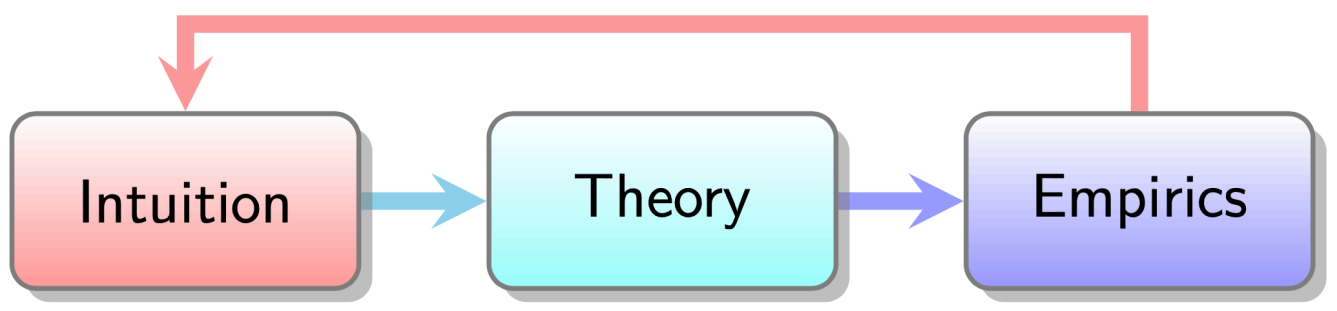
A Scientific Framework for Rational Cognition

An Example: Smoke and Mortality

- Anecdotes or Intuition
  - eg. "My grandmother smoked two packs a day and lived until she was 95 years old."
- Theory and Formalization
  - Because Cigarettes contain carcinogens(致癌物) such as nicotine, tar, and formaldehyde(尼古丁、焦油、甲醛等), then...
- Empirical Evidence
  - Collecting data through experiments or surveys, and then use statistical or econometrical methods to verify whether and how cigarettes can harm our health.
- **Conclusion**:
  - Cigarettes **harm our health** and **quantify the extent of the harm**.

A Scientific Workflow for Analysis

- By **Intuition**: Propose meaningful or interesting questions (questions that matter or that we care about)
- By **Theory**: Derive a preliminary conclusion or propose a testable hypothesis
- By **Empirics**: Use data and quantitative methods to test your theory or hypothesis



- Once we have a theory (or causal mechanism) that has been tested by empirical work, we can manipulate the cause to achieve the desired effect.

Scientific Methods for Economic Analysis

- **Use quantitative(定量) methods to answer both qualitative(定性) and quantitative(定量) questions.**
- **A numerical answer to the question**

  > whether the effect is positive or negative and how large the effect is?*

- **A measure of how precise the answer is**

  > how confident we are in the answer?

- It is the job of **Econometrics**.

Paradox of Education on Wages

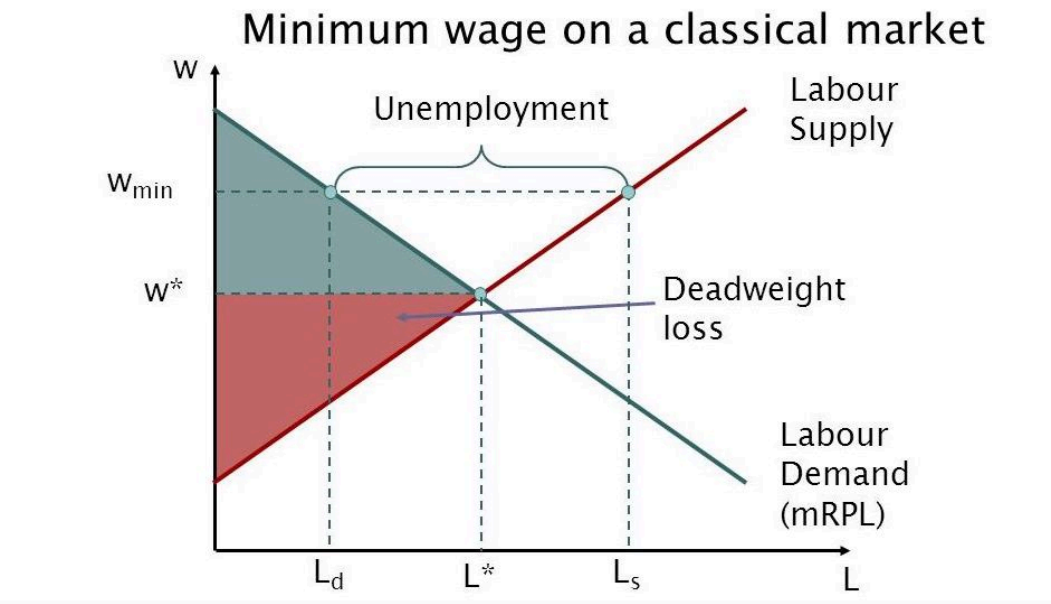
Human Capital v.s Signal

- A common phenomenon in labor markets can be observed across countries.
    - Higher education, Better pay!

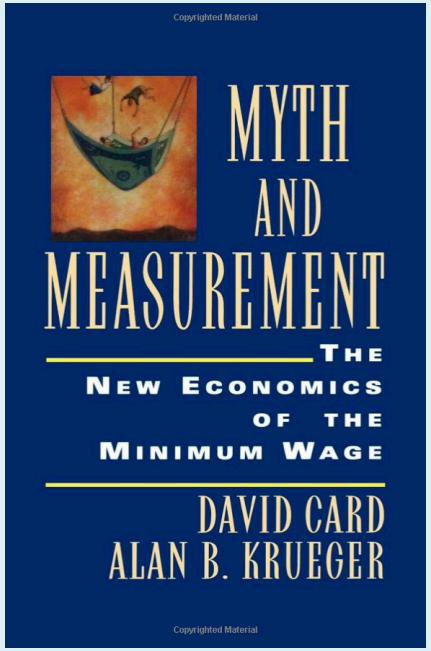
- Two classical theories to explain
    - **Human Capital**: Education improves work productivity.
    - **Signal**: Education does not increase the productivity. It simply serves as a signal of the individuals' innate ability.
- **Question**: which one is right?

Controversy of Public Policy

Minimum Wage and Unemployment



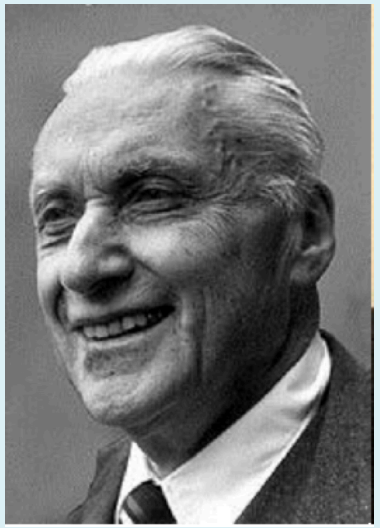
- The classical supply-demand model tell us that
  - Minimum wage will definitely increase unemployment.



- 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.

What Is Econometrics?

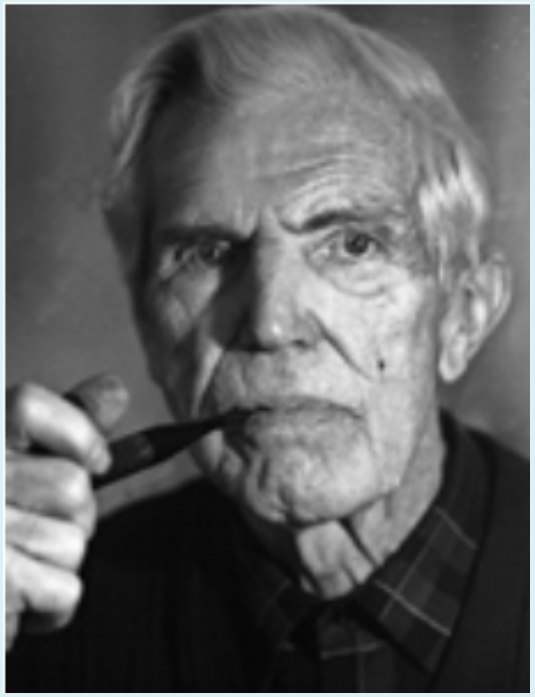
Econometrics: A Brief History



- The term is attributed to **Ragnar Frisch(1895-1973)** who is the 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.

> "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**". in *Econometrica*, 1933, volume 1, pages1-2.

Econometrics: A Brief History



- **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.". in *Econometrica*, 1944, volume 12, pages1-2.

Econometrics: A Brief History

James Stock(Havard)

Mark Watson(Princeton)

> "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 to Econometrics*, 4th edition.

The Two Missions of Econometrics

- **My Own View**

> Econometrics is a series of **scientific methods** for extracting economic logic from data. It serves **two fundamental missions**:

Mission 1: Causal Inference

- Estimating causal effects
- Policy evaluation
- **Core Question**: *"What is the effect of X on Y?"*

Mission 2: Scientific Prediction

- Forecasting future outcomes
- Pattern recognition and classification
- **Core Question**: *"What will Y be, given X or past Y?"*

> **Key Insight**: These are fundamentally different tasks. A model that predicts well may be useless for causal inference, and a credibly causal estimate may be a poor predictor. (Mullainathan & Spiess, 2017)

Two Missions: Why the Distinction Matters

Causal Inference

- **Goal**: Understand Causal Effects
- **Question**: What happens if we *change* X?
- **Example**: How much would reducing class size *improve* test scores?
- **Method**: RCT, IV, DID, RDD

Scientific Prediction

- **Goal**: Forecast outcomes
- **Question**: What Y will we *observe* given X?
- **Example**: Which students are *likely* to fail?
- **Method**: Machine learning, Time Series Analysis

> **This course focuses primarily on Causal Inference** — the credibility revolution toolkit. Prediction and forecasting (time series, machine learning) are equally important but require separate courses.

Econometrics, Big Data, and Social Science

- There can be many labels for our work...
  - Econometrics(Causal Inference)
  - Statistics
  - Data Mining/Big Data/Data Science
  - Machine Learning(ML) and Artificial Intelligence(AI)
- Along this spectrum, the focus shifts from heavily emphasizing the phenomena being measured to a more practical approach of discovering patterns that are useful and true.
  - *The more we move to the right, the more we are interested in prediction and less in causality.*
  - *The more we move to the left, the more we are interested in causality and less in prediction.*
- From a data scientist's perspective, the **similarities** are much greater than the distinctions.

Econometrics: Sub-fields

- **Theoretical Econometrics**(理论计量经济学)
  - 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**(应用计量经济学)
  - More oriented toward applied work, such as choice of technique and interpretation of research findings.
  - But it should also be grounded in a solid conceptual foundation, practical experience, and some programming skills.
  - Most of us are the **consumers** of econometrics.

Econometrics: Structural vs. Reduced-Form

- **Structural Econometrics** (结构式)
- Build and estimate **explicit economic models** (utility, production, equilibrium)
- Requires strong theoretical assumptions
- Can perform **counterfactual** policy simulations
- Examples: discrete choice models, DSGE, auction models

- **Reduced-Form Econometrics** (简约式)
- Focus on **causal effects** without full structural model
- Relies on **research design** (IV, RDD, DID, RCT)
- Fewer assumptions, more "credible" identification
- Examples: returns to education, minimum wage effects

> **This course focuses on reduced-form methods — the "credibility revolution" in empirical economics. Structural methods require more advanced training.**

Economics in the Age of Big Data and AI

Economics in the Age of Big Data and AI

- Social sciences have experienced two methodological **revolutions** over the past few decades.

- **No.1: Credibility Revolution**
  - A movement that emphasizes the goal of obtaining secure causal inferences in social sciences.
  - The revolution started from around the 1990s, pioneering in economics, then spread over to other empirical social sciences such as sociology, political science,public policy, etc., which has entirely changed empirical social science and business research.

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

© Nobel Prize Outreach. Photo: Paul Kennedy
David Card
Prize share: 1/2

© Nobel Prize Outreach. Photo: Risdon Photography
Joshua D. Angrist
Prize share: 1/4

© Nobel Prize Outreach. Photo: Paul Kennedy
Guido W. Imbens
Prize share: 1/4

Economics in the Age of Big Data and AI

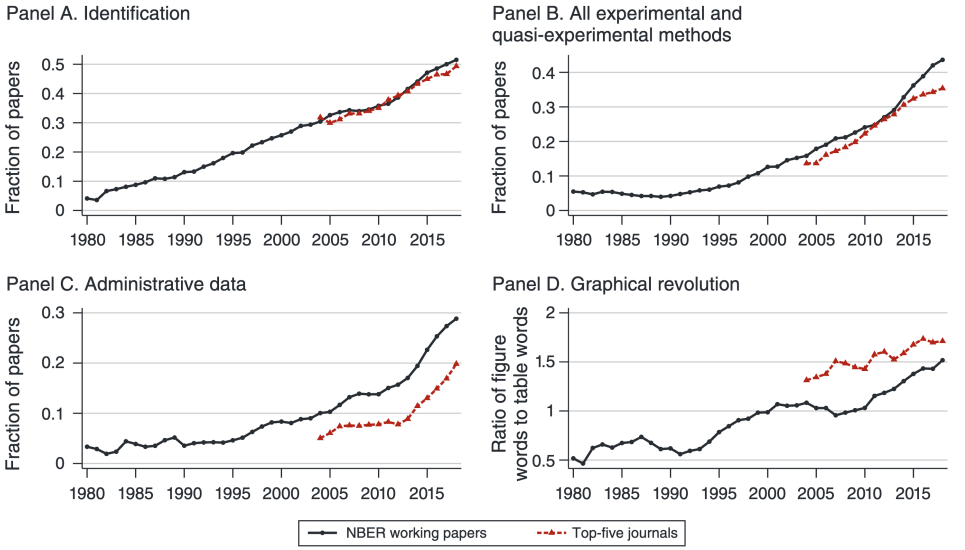


FIGURE 2. THE CREDIBILITY REVOLUTION

Notes: This figure shows different dimensions of the "credibility revolution" in economics: identification (panel A), all experimental and quasi-experimental methods (panel B), administrative data (panel C), and the graphical revolution (panel D). Panel D shows the ratio of the number of "figure" terms to the number of "table" terms mentioned. See Table A.I for a list of terms. The series show five-year moving averages.

Key words for CR

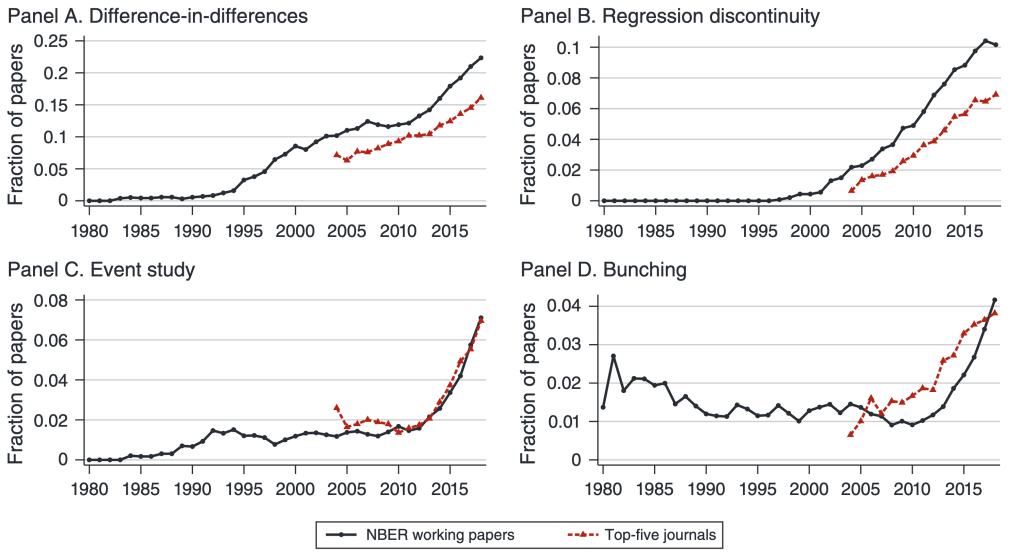


FIGURE 4. QUASI-EXPERIMENTAL METHODS

Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show five-year moving averages.

Quasi-experimental methods

- Currie, J., Kleven, H., & Zwiers, E. (2020). Technology and Big Data Are Changing Economics: Mining Text to Track Methods. AEA Papers and Proceedings, 110, 42–48

Economics in the Age of Big Data and AI

- **No.2: Big Data Revolution**
  - 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 and Cukier, 2014)

- Viktor Mayer-Schnberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work and Think

- Data sources and types are changing, which makes new methods to obtain, process, analyze and visualize data necessary.

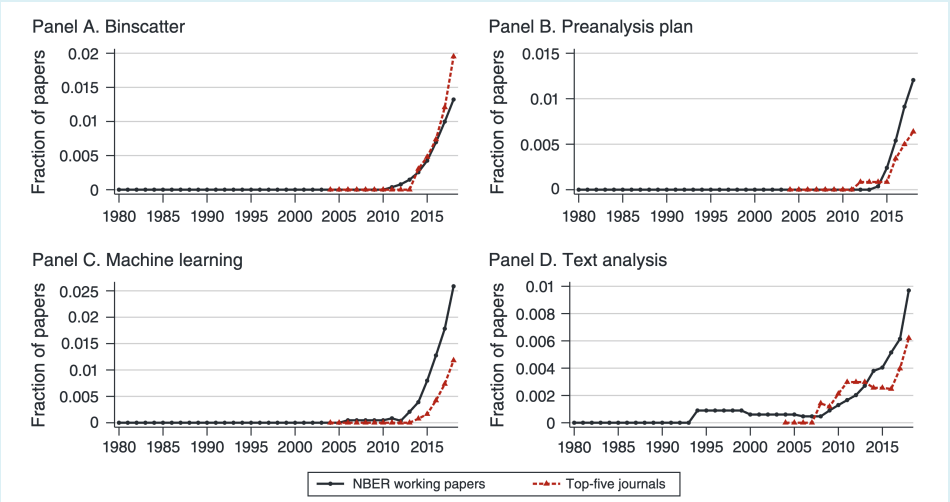


FIGURE 6. WHAT'S NEW?

Notes: This figure shows the fraction of papers referring to each method. See Table A.I for a list of terms. The series show five-year moving averages.

Currie, J et al(2020)

Economics in the Age of Big Data and AI

- Now we are facing the third revolution in social science: **No.3: AI Revolution**.

What is Artificial Intelligence (AI)?

- It is a field of **computer science** that aims to create machines that can perform tasks that typically require human intelligence.It includes several subfields, such as machine learning, natural language processing, computer vision, robotics, and expert systems.
- The most influential breakthrough in AI recently is in the space of **Generative AI** models or the **Large Language Models(LLM)**
  - which are designed to create **text or other forms of media** based on patterns and examples they have been trained on such as **ChatGPT** and many others.
- The ability of the AI model is still quickly evolving and upgrading.

Economics in the Age of Big Data and AI

- It is dramatically changing the way of obtaining, processing, analyzing and visualizing information and knowledge. So it is also changing the way of doing research.

Category	Task	Usefulness
Ideation and Feedback	Brainstorming	●
	Feedback	◐
	Providing counterarguments	◐
Writing	Synthesizing text	●
	Editing text	●
	Evaluating text	●
	Generating catchy titles & headlines	●
	Generating tweets to promote a paper	●
Background Research	Summarizing Text	●
	Literature Research	○
	Formatting References	●
	Translating Text	●
	Explaining Concepts	◐

The third column reports my subjective rating of LLM capabilities as of September 2023:
○: experimental; results are inconsistent and require significant human oversight
◐: useful; requires oversight but will likely save you time
●: highly useful; incorporating this into your workflow will save you time

Category	Task	Usefulness
Coding	Writing code	◐
	Explaining code	◐
	Translating code	●
	Debugging code	◐
Data Analysis	Creating figures	◐
	Extracting data from text	●
	Reformatting data	●
	Classifying and scoring text	◐
	Extracting sentiment	◐
	Simulating human subjects	◐
Math	Setting up models	◐
	Deriving equations	○
	Explaining models	◐

The third column reports my subjective rating of LLM capabilities as of September 2023:
○: experimental; results are inconsistent and require significant human oversight
◐: useful; requires oversight but will likely save you time
●: highly useful; incorporating this into your workflow will save you time

- Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. Journal of Economic Literature, 61(4), 1281–1317

Quickly Evolving and Upgrading LLMs

- From a **ChatBot** to **Reasoning and Programming Assistant**,a **Autonomous Agent**,and now **a group of agents working together** to achieve a common goal.

Table 1
Primary Usefulness of LLM Model Types for Research Tasks

Research Category	Traditional LLMs	Reasoning Models	Agentic Chatbots
Ideation & Feedback	Good for initial brainstorming.	**Best for structured feedback and identifying logical flaws.**	Actively scans literature for novelty and grounding.
Writing	**Excellent for drafting, summarizing, and rephrasing existing text.**	Ensures logical flow in complex arguments.	Incorporates real-time web information.
Background Research	*Shortcoming: No live web access.*	Synthesizes info from provided texts.	**Best for live web searches and up-to-date literature reviews.**
Coding	Generates basic code snippets.	**Excellent for writing and debugging complex algorithms.**	**Executes code, tests hypotheses, and interacts with data files.**
Data Analysis	*Shortcoming: Cannot execute code.*	Helps interpret data and suggest approaches.	**Best for end-to-end analysis: data cleaning, coding, visualization.**
Math	*Shortcoming: Unreliable for complex math.*	**Solves multi-step problems and formal proofs at PhD level.**	Can leverage external computational tools to ensure accuracy.
Promoting Research	Drafts initial promotional content.	Tailors summaries for specific audiences.	**Automates creating summaries and posts for multiple platforms.**

Note: The most useful model type in each research category is bolded.

- Korinek, A. (2025). Generative AI for Economic Research: Use Cases and Implications for Economists

Summary: Econometrics in the Age of AI

What AI Changes:

- **How** we conduct analysis
- Speed and scale of computation
- Accessibility of programming

What Remains Unchanged:

- **Why** we need causal inference
- **What** questions matter for policy
- The importance of research design

> **AI has transformed the "how" of econometrics, but not the "why" or the "what." The core objective — developing causal reasoning and research design skills — becomes even more important in the age of AI.**

- We need **more instead of less** knowledge of econometrics and its applications in the age of AI.

Why and Who Should Take This Course?

Why Is Econometrics So Important?

- Several Common Questions about Econometrics:
  - *Why should we study econometrics?*
  - *How does studying econometrics help us understand social science?*
  - *Most importantly, why, what, and how should we study econometrics in the age of AI?*
- Econometrics is **one of three core courses** required in almost every economics department worldwide.
  - The other two are **Microeconomics** and **Macroeconomics**.
- **Econometrics** is not only crucial in economic research, but also plays an increasingly important role throughout the social sciences and business research.

Why take the course?

The Purpose of the course

- Introduce you to modern econometrics in a way that is accessible to students who have not taken a prior course in econometrics.
  - Laying down a solid theoretical background in introductory level econometrics.
  - Learning about basic tools and the latest developments in the empiricist's toolbox.
  - Developing an ability to implement modern econometric methods to real-world data.
  - Cultivate a **critical thinking** about empirical studies and applications in social sciences.
- The course focuses more on intuition and practical examples, while **keeping the mathematical content to a minimum**.
  - Notably, **matrix notation is not used in the course**.

Why take the course?

Hopefully, taking this course can aslo help you:

- Master important skills in college studies:

- **Hard Skills**
  - **Language**
  - **Computer**
  - **Presentation and Writing**
- **Soft Skills**
  - **Critical Thinking**
  - **Teamwork**

- Fortunately, you could practice all above skills more or less in our class.

Why take the course?

For those pursuing an academic career:

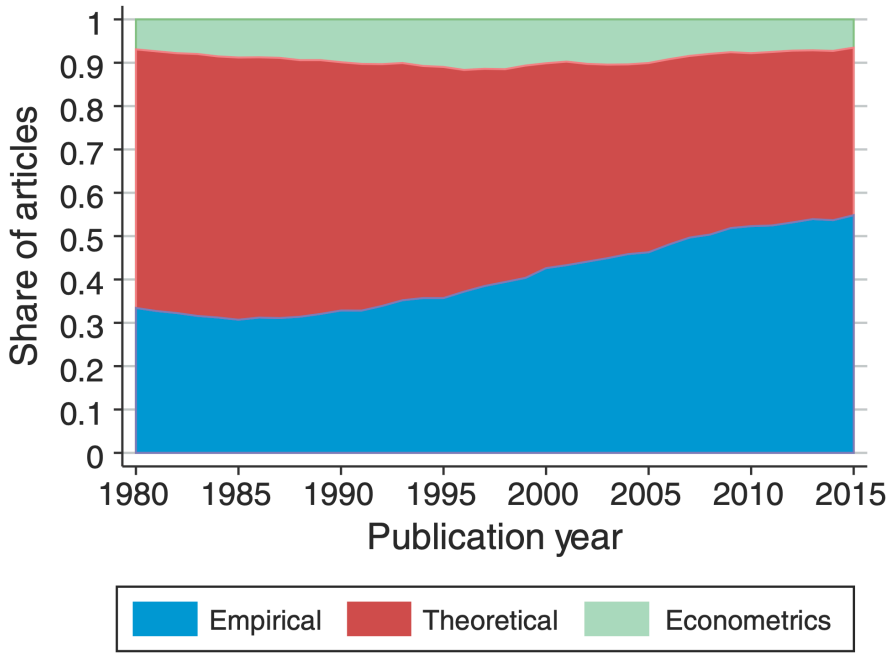


FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

Angrist et al(2017)

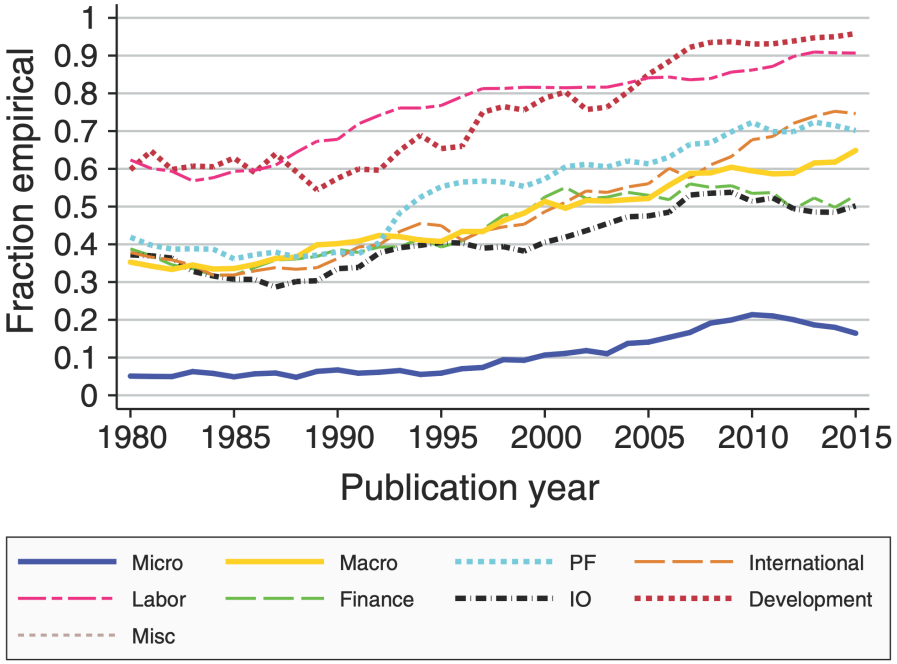


FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

Angrist et al(2017)

- The proportion of empirical studies in economics is **increasing more and more**.

Why take the course?

For Those Entering the Industry Job Market

- If you want to work in industry, mastering econometrics may 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,Tiktok
- **Data Analyst/Data Scientist** is **the hottest job** in consulting, business areas as well as financial industry right now.

Why take the course?

Industry Job Market: Apple Job Posting

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 provide business solutions
- Build models to causal impact of new programs release across different scenarios
- Develop internal visualization and modeling 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 econometric models to business problems.

Why take the course?

Industry Job Market: ByteDance Job Posting

国际电商–经济学家/数据科学家

上海 | 正式 | 产品 – 数据分析 | 职位 ID：A117677

职位描述

我们欢迎有创造力、探索精神、且具备基本经济学、统计学素养的人才加入，和业务方共创并推动项目的上线落地。我们的合作业务方包括推荐算法、产品、运营、资源管理等。

主要职责：

把商业问题转化为可解的模型问题。通过经济学视角的思考和科学的方法（因果推断、AB实验、求解理论模型、预测等）来推动搜推策略、产品功能、资源分配等相关决策:

1、因果性的衡量各类策略、政策的效果，衡量长期影响，并形成系统性的方法论；

2、对数据现象现象进行归因，对用户、商家的决策链路做深入探索，总结洞察和建议，帮助各决策方建立认知；

3、优化各类资源分配（流量、营销补贴）；

4、优化国际电商的生态环境，包括但不限于经营环境，用户体验，内容生态，供给生态，持续助力商达成长和用户增长。

职位要求

1、经济学、统计学、运筹学、金融学、或者其他的相关的量化学科背景；

2、掌握 R 或Python等至少一项数据分析必备的编程语言，以及基础SQL能力；

3、有一定的解决商业问题、构建可落地的系统性解决方案、复杂项目管理、协调多方决策的经验；

4、良好的写作沟通能力；

5、以下领域的相关的科研、或者业界项目经历: reduced-form 因果推断、预测、causal ML、劳动经济学、健康经济学、教育经济学、行为经济学、金融经济学、产业组织学。

投递

商业化数据科学家–因果推断

上海 | 正式 | 产品 – 数据分析 | 职位 ID：A56405

职位描述

1、通过积累日常使用经验、阅读相关学术论文和公开资料等，沉淀并向数科团队输出对因果推断方法论的深入理解和使用经验，澄清常见的使用误区，提供标准应用流程指南，以保障方法在团队内应用的科学性，提高使用效率；

2、关注具有方法论共性或场景共性的相似业务问题，主导专项探索，与其他业务方向数科同学紧密配合，从宏观视角优化资源分配效率或策略，优化产品策略，或针对相似问题抽象可复用的、普适的分析框架或解决方案，提升团队分析、决策效率；

3、对宏观战略问题进行拆解、定义，通过数据描述、可视化、挖掘、统计建模等方法，提炼有效的数据洞察和产品战略建议，指导科学的决策与迭代。

职位要求

1、本科以上学历，统计学、数学、计量经济学、数据科学、计算机等量化分析相关专业优先，硕士、博士优先；

2、具备扎实的统计学/计量经济学/机器学习/因果推断等数据科学理论基础及应用经验；精通SQL，熟练掌握Python/R中的一种，可进行数据清洗、可视化和分析；

3、具备快速学习能力，能够快速理解产品逻辑，并具备较强的逻辑思维能力，在较大不确定性的问题中可以构建分析框架，将数据转化为有效的商业洞察；

4、能够主动、独立思考的同时，具备良好的团队协作能力与责任心，善于与其他协作团队沟通，有主人翁意识；

5、具备强烈的好奇心与自我驱动力，乐于接受挑战，追求极致和创新，富有使命感。

投递

Why take the course?

For Those Seeking Public Sector Jobs (Shang'an Movement)

- To be honest with you, the course may **NOT** help you succeed in the examination directly.
- However, in the long run, it provides valuable knowledge and skills that offer a broader understanding of the subject matter and enhance your critical thinking abilities.
- It ultimately benefits your career, as well as our people, our country, and the world.

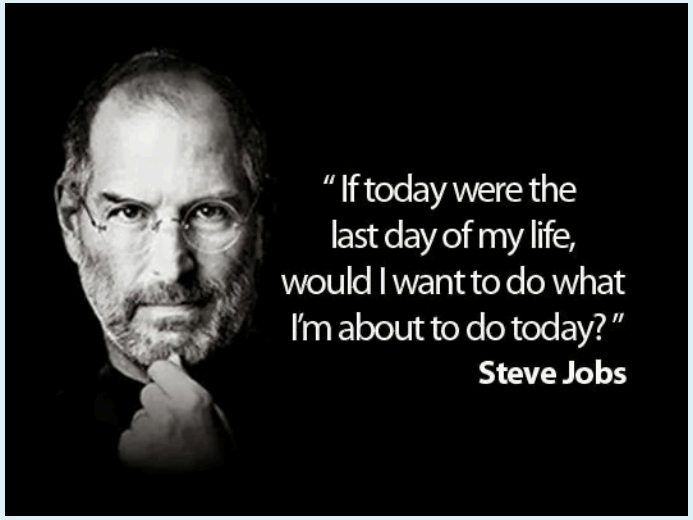
Why take the course?

For Those Who Seek Intellectual Enjoyment

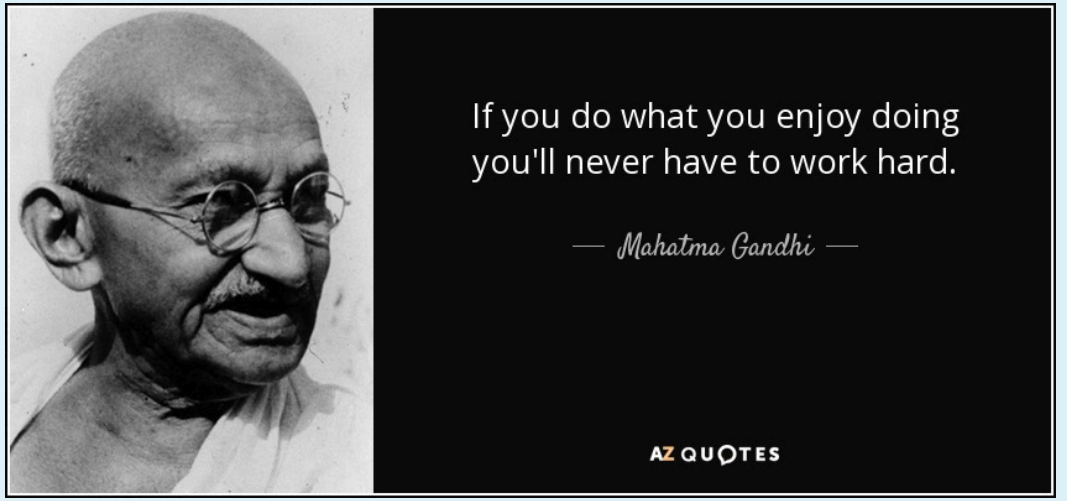
- This course need not be a boring and demanding variant of a mathematics course — it can be an interesting and enjoyable class.
- Help yourself enjoy life by adopting an empiricist's mindset in your daily activities.
  - novel ideas or new perspectives on our world.
- Also covering several interesting and insightful stories like
  - Eg. **Crime and Abortion** in *Freakonomics* written by Steven Levitt.
  - Eg. What is **the economic 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**.



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

Wrap up

- In essence, Intro 'Metrics is an **essential and intriguing yet challenging** course.
  - **Please think it over before enrolling!**
  - **Once committed,please work hard on it!**
  - **And remember, enjoy the process of working hard!**

Course Logistics

About Me, our TA, and the Course

- My name is Zhaopeng Qu(曲兆鹏)
  - Associate Professor, Institute of Population Studies, Business School.
  - Research Fields: Labor Economics and Applied Econometrics
  - Email: qu@nju.edu.cn
- Our TA: Xiaotian YU(虞晓天)
  - Ph.D. students in the second grade, proficient in Stata.
  - Recite some basic concepts and formulas in probability theory and statistics.
  - Help you with Stata during lab sessions.
  - Responsible for evaluating your homework.
- Online Resources
  - **Our Course Website**: https://byelenin.github.io/Metrics2026/
  - **Wechat group**: You can create one (though it's not required) to discuss anything related to the course.
- 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-)

Prerequisite

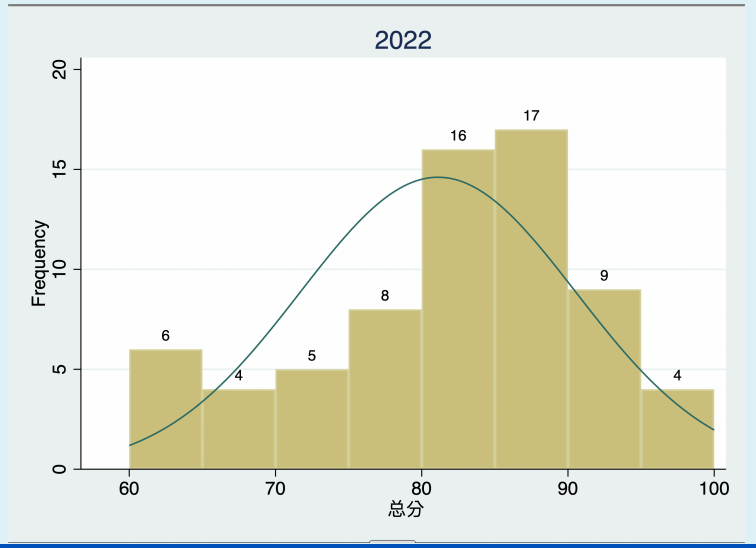
- I assume that you are comfortable with some basic concepts in **probability theory and statistics**, such as
  - Random variable,expectation, variance and covariance
  - Probability density function, p.d.f. and cumulative distribution function, c.d.f
  - L.L.N and C.L.T
  - OLS and other Estimators
  - Unbiased and Consistent
  - Asymptotic Normality
  - Interval estimation and hypothesis test
- I will **NOT** review these basic concepts and formulas in lectures, but it will be **reviewed by our TA** in the recite sections. You should review it by yourself anyway.

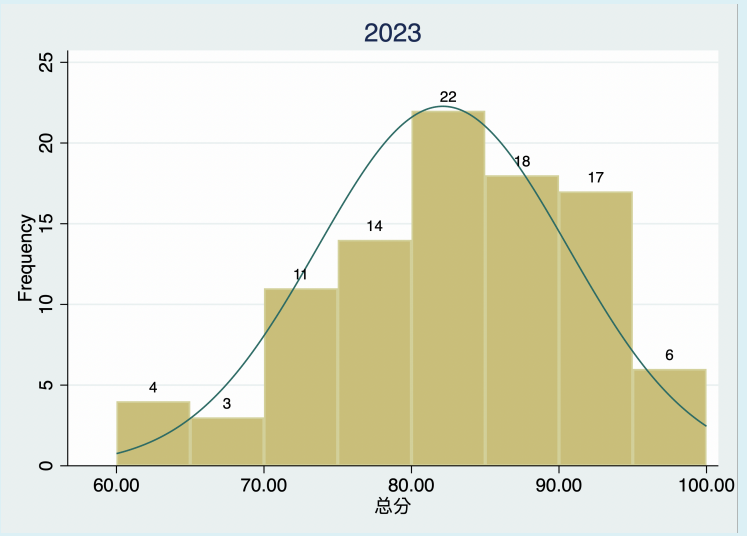
The Procedure

- The First Part: **Lectures by the instructor**
  - 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 takes the course is **required to attend lectures**.
- The Second Part: **Recitation Sessions and Computer Labs by our TA**
  - Review some basic concepts and formulas in probability theory and statistics.
  - Review your homework if necessary.
  - Teach you basic skills of Stata/R.
- While they may be beneficial to your studies, these sessions are **optional and students are not required to attend**.

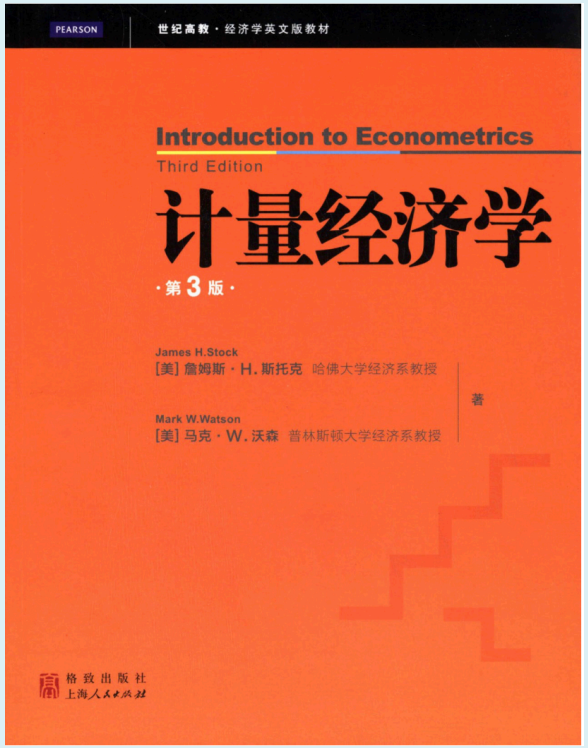
Evaluation

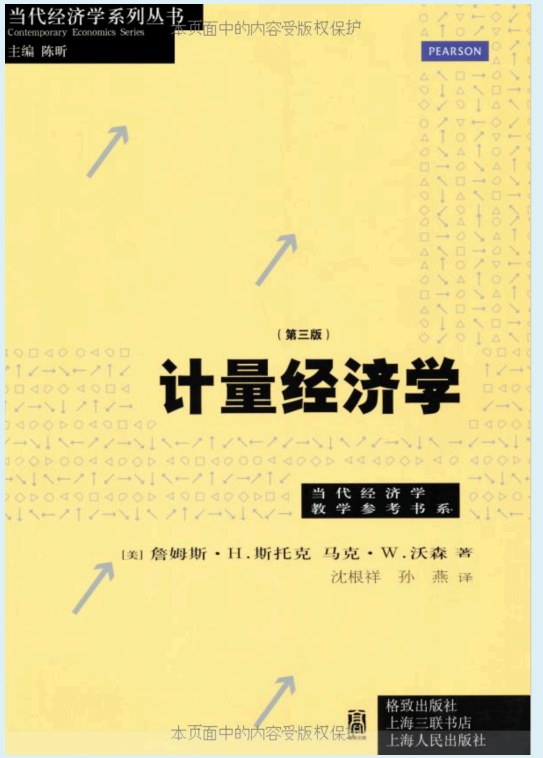
- The final grade will be based on the following components:
  - **Class Participation(10%)**
  - **Homework(30%)**
  - **Final Exam(40%)**
  - **Team Project: A research proposal(20%)**
- Total Score Distributions in 2022 and 2023





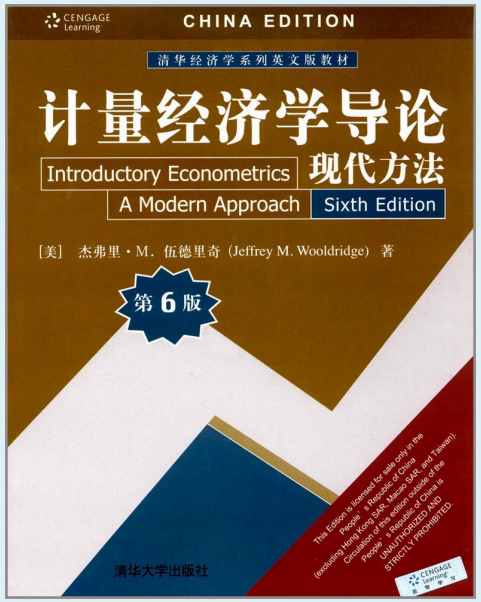
Required Textbook

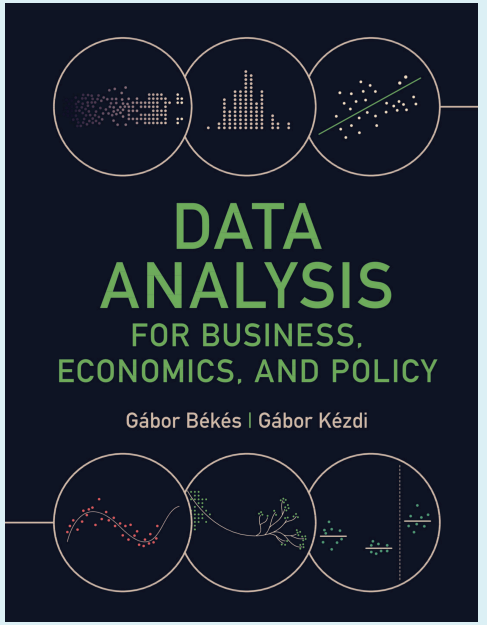




- James H. Stock & Mark W. Watson, (2012). **Introduction to Econometrics**, 3rd Edition, Pearson Education.
  - 影印版/中文版：格致出版社／上海人民出版社。

Supplementary Textbooks





- 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. （中译本：《现代计量经济学导论》，中国人民大学出版社，2013）
- Gábor Békés & Gábor Kézdi(2021). Data Analysis for Business, Economics, and Policy. Cambridge University Press.

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》,中信出版社。）
- 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》，中国青年出版社。)
- 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。）
- Abhijit V. Banerjee & Esther Duflo, Good Economics for Hard Times,2019.(中译本:《好的经济学》，中信出版社。)

Computing Tools

- The main computing tools used in the course is **Stata** instead of **R**.

Stata

- Pro: It is a powerful tool for data analysis and econometric modeling.
- Con: It is not free.

R

- Pro: It is a powerful tool for data analysis and econometric modeling and it is free.
- Con: It is not as accurate in econometric modeling as Stata.

Promise and Expectation

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!
- Finish your homework.
- Self-motivated learning by doing.

Two Iron Rules

- **Don't ever cheat on your assignments!**
- How to use AI tools properly in your study and research? please follow my instruction.

- **Don't ever snitch your teachers to help political repression!**

Welcome contact me

An Introduction to Economic Data

Two Axioms of Data Analysis

- **Axiom 1**: Any economy can be seen as a **stochastic process** governed by a certain probability law.
  - The economy's future state is not deterministic but can be described in terms of probabilities.
- **Axiom 2**: Economic phenomena, often summarized in form of data, can be interpreted as a **realization** of this stochastic data generating process.
  - By studying historical data, we can infer patterns, trends, and the probability distributions that describe the stochastic process, thereby gaining insights into the economy's underlying dynamics.
- It highlights the importance of probabilistic models in economics and provides a theoretical basis to use statistical tools and models to analyze economic data.

What is Data

- Data is a collection of facts or information, which can be presented in various forms such as *numbers*, *tables*, *words*, *graphs*, *pictures*, or even *sounds* and *videos*.
- And it can be processed and analyzed to produce knowledge and insights either by itself or after structuring, cleaning, and analysis.
- Data is most straightforward to analyze if it forms a single **data table**(a matrix).
  - It consists of **observations**(观测值) and **variables**(变量).
  - Observations are also known as cases, or row.
  - Variables are sometimes called **features** or covariates.
- Normally, in a data table the *rows are the observations*, *columns are variables*.

A Simple Example: CA School Data

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

Variables

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
. . .	. . .	. . .	. . .	. . .
ID of Observations			One case	
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.

Data: Basic Characteristics

- First, we need to distinguish between the **population** and the **sample**.
  - **Population**: the entire group of individuals, objects, or events of interest.
  - **Sample**: a subset of the population that is supposed to be representative of the population.
- Second, we need to know **the unit of observation** and **the unit of analysis**.
  - **Unit of Observation**: the entity about which data are collected.
  - **Unit of Analysis**: the entity about which analysis/inferences are made.

A Simple Example: CA School Data

Topic: the effect of **class size** on **student achievement** in California public schools.

- Ideally,
  - **Population**: all the students in the CA public school system.
  - **Sample**: a subset of the students in the CA public school system by a **random sampling**.
- **Unit of Observation** and **Unit of Analysis**: each student.
- In reality, based on the availability of data and the way data is collected
  - **Population**: all the schools(districts), which can be equivalent to all the students in the CA public school system.
  - **Sample**: a subset, 420, of the all schools(districts) by a random sample.
- **Unit of Observation** and **Unit of Analysis**: each school(district)
- Actually, the unit of observation and unit of analysis can be **different**.
  - eg.the unit of observation is the student, but the unit of analysis are both the student and class.

Data: Sources and Types

Data Sources

- Traditional Collecting Methods:
  - Statistical Reports or Documents
  - Survey or Census
  - Administrative Data
  - Lab or Field Experimental Data
- Collecting Data in Digital Times:
  - Online Transactions or Activities
  - Social Media
  - Geolocations or Geographic Data
  - Online Documents or Texts
- Always need to figure out
  - what is the **population and sample** of the data.
  - what is the **unit of observation and unit of analysis** of the data.

Data: Sources and Types

Data Quality

- Including
  - Content
  - Accuracy
  - Completeness
  - Consistency
- **Garbage in, Garbage out**
  - **Prioritize data, then methods.**

Ethical and Legal Issues

- Including
  - Privacy and Confidentiality
  - Data Security
  - Data Ownership
  - Data Sharing and Open Data

Data Types

Experimental V.S. Observational

- **Experimental** data come from experiments designed to evaluate a treatment or policy or to investigate a causal effect.
- **Observational** data come from non-experimental settings, such as surveys,administrative records and other sources.

Data Structure

- Cross-sectional data
- Time series data
- Panel/longitudinal data
- Pool-cross sectional 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

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

- 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.

- **Questions?**: observations,variables,the sample size?

$$x_i = STRatio_i;\ y_i = TestScore_i;\ N = 420$$

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

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

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

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.

- **Questions?** observations,variables,the sample size?

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

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.

$$\{x_{it}, y_{it} \cdots\}_{i=1,t=1,}^{NT}$$

- 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.

- **Questions?**: observations,variables,the sample size?

$$x_{it} = Total\ Taxes_{it};\ y_{it} = Cigarette\ Sales_{it};\ N \times T = 48 \times 11 = 528$$

4.Pool 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.
  - While the data come from a same population in different time,the data does not necessarily track the respondent multiple times.
- 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:
  - Cohort studies
  - Difference-in-differences analyses
  - 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

- **Questions?**: observations,variables,the sample size?

$$x_{ijt} = hprice_{i,1993}, hprice_{j,1995}; y_{ijt} = proptex_{i,1993}, proptex_{j,1995}; N = N_i + N_j = 250 + 270 = 520$$

Variables or Features in Data

- Variables are the characteristics or properties of the observations in the data set, which can be broadly classified into two types:
- **Quantitative**: continuous or discrete numerical values. These numbers represent a quantity or amount.
  - **Scale**: the difference between two values is **meaningful and consistent** in most cases.
- **Qualitative**: specific numbers used to represent different groups or categories.
  - Here, the scale is mostly **meaningless or arbitrary**.
- The number of variables is the number of columns in the data set,represents by p,
  - which is normally **less than** the number of observations in the **traditional data set**, represents by N

$$p \ll N$$

  - However, in the **big data** era, the number of variables can be **larger than** the number of observations, represents by $p \gg N$

Data Types and Sub-Econometrics

Micro-Econometrics(微观计量经济学)

- Cross-Sectional
- Pool Cross-Sectional
- Short Panel(**large N, small T**)

Macro-Econometrics(宏观计量经济学或者时间序列)

- Times series
- Long Panel(**small N, large T**)

Big Data Econometrics(Machine Learning Econometrics)

- High-dimensional data(**larger p**, **large N**, **large T**)
  - eg. variables selection to avoid the curse of dimensionality
- Unstructured data(**unstructured N**, **unstructured p**, **unstructured T**, **many missing values**)
  - eg. extract information from text data, image data, audio data, video data, etc.

Typical Data sets in China

Survey Data(抽样调查数据):

- China Family Panel Survey(CFPS)
- China Health and Retirement Longitudinal Study(CHARLS)

Administrative Data(行政数据):

- Census: 全国人口普查数据、全国1%人口抽样调查
- China Industrial Survey Data: 工业企业数据库
- Chinese Custom Transaction Data: 海关交易数据库
- National Industrial and Commercial Enterprises Database:全国工商企业登记数据库:

Typical Data sets in China

Online Big Data:

- Transaction data on Tabao,JD,Tmall(淘宝、京东、天猫...）
- Movie Data on Douban.com(豆瓣\猫眼电影数据)
- Night-Lights Data(夜间灯光数据) and Air Quality: PM2.5(空气质量数据）
- Geolocations Data(地理位置数据）: Baidu Map, Didi, Mobike, Ofo...
- Social Media Data(微博、微信、知乎、豆瓣、贴吧、论坛、博客、新闻、评论、问答、社交网络...）
- Online Job Market Data(51job.com,lagou.com,zhipin.com...)
- Online Financial Data(Stocks, Bonds, Funds, Futures, Options, Forex, Cryptocurrency...)
- Official documents and News(新闻报道、政府公告、会议记录、政策文件、法律法规、司法判决...)

Homework(not required)

Homework

- Register on one of the following database websites and download the data set.
  - 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
  - China Urban Labor Survey(CULS):中国城市劳动力调查

Homework

- Understand the purpose and main content of the sample survey, as well as basic information such as the scope of sampling, the method, and the sample size. Determine the data structure to which this data belongs.(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?)
- Download the survey questionnaire and comprehend the specific information it contains.(下载调查的问卷，详细了解调查有哪些具体的信息)
  - Identify the questions you are interested in and locate them in the questionnaire.(首先确定自己感兴趣的问题，然后到问卷中去寻找)
  - Alternatively, review the questionnaire first and find the specific information of interest.(或者先看问卷，找到自己感兴趣的具体信息)
- Download the corresponding data and perform preliminary data cleaning and statistical analysis (to be completed after the computer class)(下载相应数据，进行初步的数据清理和统计分析）
- Prepare for the research proposal project at the end of semester(为期末的研究项目做准备）

References

- Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2017). Economic Research Evolves: Fields and Styles. American Economic Review, 107(5), 293–297. https://doi.org/10.1257/aer.p20171117
- Most figures from our textbook:
  - Stock, J. H., & Watson, M. W. (2015). *Introduction to Econometrics*. Pearson Education.
- Other figures are from the following websites: Wikipedia,Economists,Google Scholar,Nobel Foundation, 51job.com etc.