Ch01. The Nature of Econometrics and Economic Data

Ping Yu

HKU Business School
The University of Hong Kong

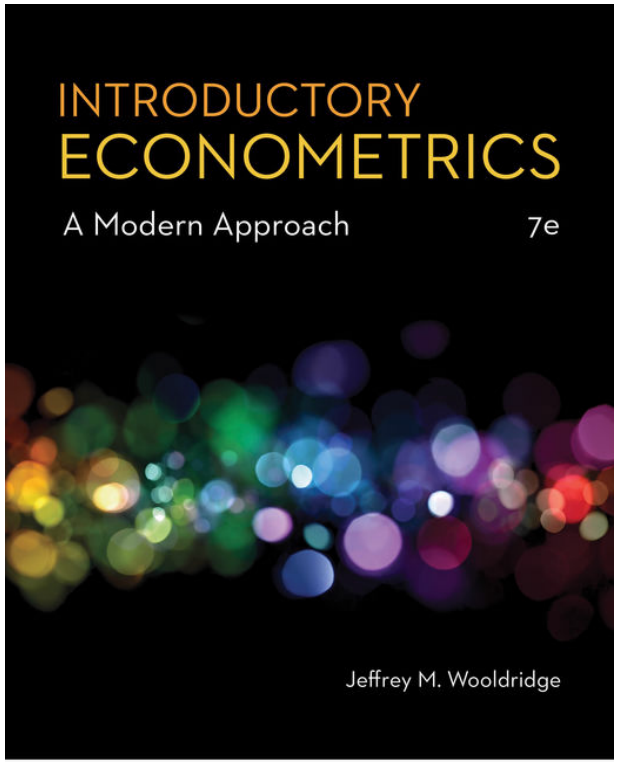
Course Information

- Instructor: Yu, Ping
- Email: pingyu@hku.hk
- Teaching Time:
  Session A: 9:30-10:15am, 10:30-11:15am, and 11:30-12:15pm, Monday;
  Session B: 1:30-2:15am, 2:30-3:15pm and 3:30-4:15pm, Tuesday.
- Teaching Location: Virtual by Zoom
- Office Hour: 3:00-4:00pm, Monday, Virtual by Zoom
  - I will NOT answer questions in email if the answer is long or is not easy to explain exactly by words. Please ask me through Zoom during my office hour.

- Tutor: Zhao, Jiuqi
- Email: zhaojq@hku.hk
- Teaching Time: TBA
- Teaching Location: TBA
- Office Hour: TBA
  - Any issues on administration (e.g., enrollment, time clash, ~~lab entrance~~, etc.) and assignments (e.g., clarification of problems) should contact the tutor.
  - Due to COVID-19, the lab is not open so you will solve computer exercises using Excel rather than STATA.

Information on the Content and Evaluation

- Textbook: Introductory Econometrics – A Modern Approach, 7th edition, Jeffery M. Wooldridge [photo here]
- Evaluation: HWs ($3 \times 6\% = 18\%$), Midterm Test (32%), Final Exam (50%)
- HW: HW must be typed. Turn in your HW on moodle on the due day. Late HW is not acceptable for whatever reasons. To avoid any risk, start your HW early.
- Tutorial: The answer key to the HWs and midterm would not be posted on moodle and will be discussed by the tutor. The tutorial class starts from week three. Tutorial questions will be posted on Moodle one week in advance. Tutorial questions are not assignments; there is no need to turn them in.
- Examination: Mimic HWs and tutorials. Closed book and closed note. A formula sheet would be provided for the midterm and final and posted on moodle before the midterm and final. Slides and Sections indexed by (*) and [Review] are not tested (but are related to HW). Excel is not tested. No past exams are provided (due to the university policy).
  - You must take the final to pass this course; if you cannot take the midterm, then the weight of midterm would be automatically shifted to the final.
  - Midterm: Oct. 25, Sunday, 10:00am-12:00noon, virtual by Zoom.
  - Final: TBA, two hours, virtual by Zoom.



Jeffery M. Wooldridge (1960-),
MSU, 1986UCSDPhD

- Other editions ($\geq$5th edition) are similar, so are perfect substitutes.

Course Policy

- In Class: (i) mute yourself during the class (I will mute all at the beginning of the class, but you can unmute yourself if you really want to say something); (ii) come to class and return from the break on time; (iii) for teaching efficiency, please do not ask me questions during the class; type in your questions through the chat function of Zoom during the class (to avoid forgetting) **OR** mark down your questions and ask me during the fifteen minutes breaks; (iv) speak English!
- Policy on Plagiarism: If judged as "plagiarism", you are in serious trouble. If a few students are judged to copy each other, each gets zero mark. I will not judge who copied whom. So **DO NOT** copy others and **DO NOT** be copied by others.
  - You may discuss with your classmates about HW, but **DO NOT** copy each other.
  - This policy applies to HW, midterm and final.
- Feedback: Any feedback to my teaching (e.g., the lecturer's English is hard to follow, technicalities are too hard to understand, the teaching should slow down, more interactions are required, etc.) is very welcome. I would incorporate your feedbacks in my future teaching during the semester. You can also give your feedbacks to the tutor so that the tutor can discuss them in tutorial classes.
- Substitute Sessions: Dr. Clement Wong will teach one session in the fall semester and two sessions in the spring semester for the same course. His sessions and mine are perfect substitutes with the same failure rate but some differences in empahsis on math.

Course Outline

- Week 1: Chapter 1 - Introduction
- Week 2, 3, 4: Chapter 2 - The Simple Regression Model
- Week 5, 6: Chapter 3 - Multiple Regression Analysis: Estimation
- Study Break: Midterm
  - Usually during the week after the break and covers Chapters 1-3.
- Week 7, 8: Chapter 4 - Multiple Regression Analysis: Inference
- Week 9: Chapter 6 - Multiple Regression Analysis: Further Issues
- Week 10: Chapter 7 - Regression Analysis with Qualitative Information
- Week 11: Chapter 8 - Heteroskedasticity
- Week 12: Chapter 10 - Basic Regression Analysis with Time Series Data
  - Because no HW problems are designed for Chapter 10, you'd better solve the problems at the end of Chapter 10 in the textbook to prepare for the final.
- Appendix B and C (Review of Probability and Statistics) are discussed only if required.
- I will roughly follow the textbook.

What is Econometrics?

What Is Econometrics?

- Econometrics = use of statistical methods to analyze economic data.
- Econometricians typically analyze nonexperimental data (or observational data/retrospective data, to emphasize that the researcher is a **passive** collector of the data).
- Typical goals of econometric analysis:
  - Estimating relationships between economic variables;
  - Testing economic theories and hypotheses;
  - Forecasting economic variables;
  - Evaluating and implementing government and business policy.

Steps in Empirical Economic Analysis

Steps in Empirical Economic Analysis

- An empirical analysis uses data to estimate a relationship or to test a theory, which are important in forecasting and policy analysis.
- Step 1: Economic model (this step is often skipped)
  - Maybe micro- or macro- models
  - Often use optimizing behavior, equilibrium modeling, $\cdots$
  - Establish relationships between economic variables
  - Example: utility maximization $\Longrightarrow$ demand equations:

$$D = f\left(\text{prices}, \text{income}, \text{taste}\right),$$

  where prices include the price of the good, and the prices of substitute and complementary goods.
- Step 2: Econometric model
- We will provide two examples for these two steps.

Economic Model of Crime

- Derives equation for criminal activity based on utility maximization:

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7),$$

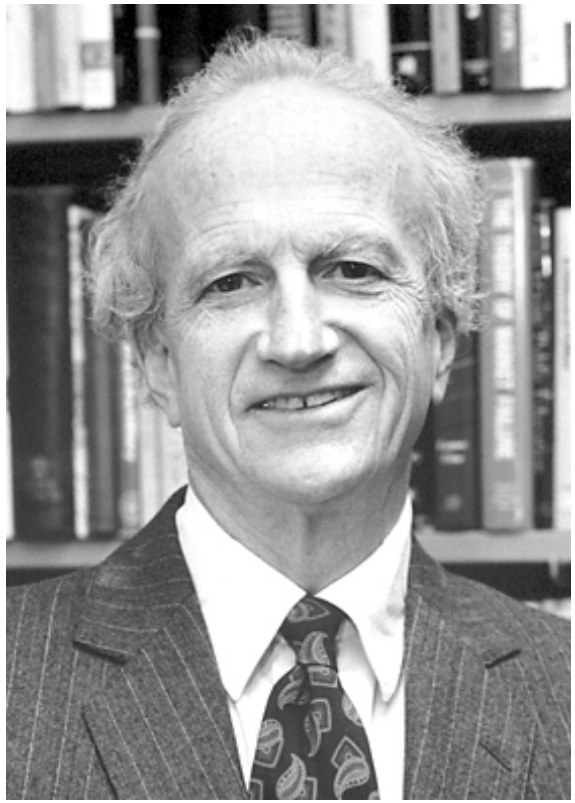
where

y = hours spent in criminal activities
x_1 = "wage" for an hour spent in criminal activity
x_2 = hourly wage for legal employment
x_3 = income other than from crime or employment
x_4 = probability of getting caught
x_5 = probability of being convicted if caught
x_6 = expected sentence if convicted
x_7 = age

- Functional form of relationship is not specified because it depends on an underlying utility function which is rarely known.
- Equation could have been postulated without economic modeling.

History of Crime Analysis

- Becker, G.S., 1968, Crime and Punishment: An Economic Approach, *Journal of Political Economy*, 76, 169-217.



Gary S. Becker (1930-2014),
Chicago, 1992NP, 1955ChicagoPhD

Econometric Model of Criminal Activity

- The functional form has to be specified.
- Variables may have to be approximated by other quantities:

$$\begin{aligned} crime &= \beta_0 + \beta_1 wage_m + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv \\ &\quad + \beta_5 avgsen + \beta_6 age + u, \end{aligned}$$

where

$crime$ = some measure of the frequency of criminal activity

$wage_m$ = the wage that can be earned in legal employment

$othinc$ = the income from other sources (assets, inheritance, and so on)

$freqarr$ = the frequency of arrests for prior infractions (to approximate the probability of arrest)

$freqconv$ = the frequency of conviction

$avgsen$ = the average sentence length after conviction

- u is unobserved determinants of criminal activity, e.g., moral character, wage in criminal activity, family background,$\cdots$

Job Training and Worker Productivity

- What is effect of additional training on worker productivity?
- Formal economic theory not really needed to derive equation:

$$wage = f(educ, exper, training)$$

where

wage = hourly wage
educ = years of formal education
exper = years of workforce experience
training = weeks spent in job training

- Other factors, e.g., productivity, may be relevant, but these are the most important.
- maybe productivity can be mostly explained by these three factors.

Econometric Model of Job Training and Worker Productivity

- An econometric model of job training might be

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u,$$

  where *educ*, *exper* and *training* are defined above, and u is unobserved determinants of the wage, e.g., innate ability, quality of education, family background,$\cdots$
- Most of econometrics deals with the specification of the error u.
- Econometric models may be used for hypothesis testing.
  - E.g., the parameter β_3 represents effect of training on wage.
  - How large is this effect? Is it different from zero?

The Structure of Economic Data

The Structure of Economic Data

- Econometric analysis requires data.
- Different kinds of economic data sets:
  - Cross-sectional data
  - Time series data
  - Pooled cross sections
  - Panel/Longitudinal data
- Econometric methods depend on the nature of the data used.
  - Use of inappropriate methods may lead to misleading results.

a: Cross-Sectional Data

- Sample of individuals, households, firms, cities, states, countries, or other units of interest at a given point of time/in a given period.
- Cross-sectional observations are more or less independent.
  - E.g., random sampling from a population.
- Ordering of observations is not important.
- Sometimes "pure" random sampling is violated, e.g., units refuse to respond in surveys, or if sampling is characterized by clustering.
- Typical applications: applied microeconomics.

Cross-Sectional Data on Wages and Other Characteristics

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

Indicator variables (1=yes, 0=no)

Observation number

Hourly wage

Cross-Sectional Data on Growth Rates and Country Characteristics

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbabwe	2.30	17	6

Growth rate of real per capita GDP

Government consumption as percentage of GDP

Adult secondary education rates

b: Time Series Data

- Time series data consist observations of a variable or several variables over time.
  - E.g., stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, automobile sales,$\cdots$
- Time series observations are typically serially correlated.
- Ordering of observations conveys important information.
- Data frequency: daily, weekly, monthly, quarterly, annually,$\cdots$
- Typical features: trends and seasonality.
- Typical applications: applied macroeconomics and finance.

Time Series Data on Minimum Wages and Related Variables

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

Average minimum wage for given year

Average coverage rate

Unemployment rate

Gross national product

c: Pooled Cross Sections

- Two or more cross sections are combined in one data set.
- Cross sections are drawn independently of each other.
- Pooled cross sections often used to evaluate policy changes.
- Example:
  - Evaluate effect of change in property taxes on house prices.
  - Random sample of house prices for the year 1993.
  - A **new** random sample of house prices for the year 1995.
  - Compare before/after (1993: before reform, 1995: after reform).

Pooled Cross Sections on Housing Prices

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

Property tax

Size of house in square feet

Number of bathrooms

Before reform

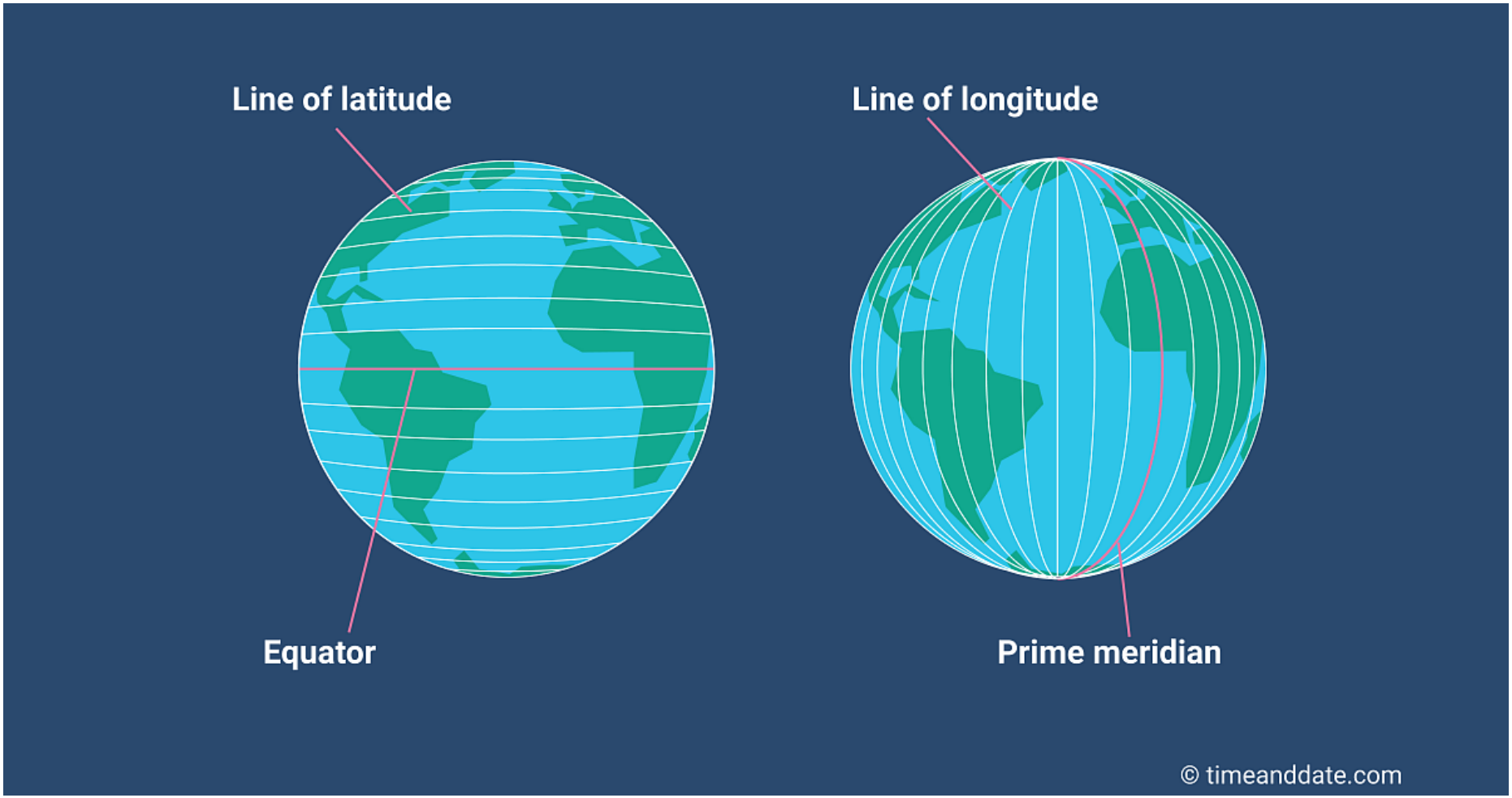
After reform

- A 1/2 bathroom usually means a bathroom with a toilet and sink but without bathtub/shower.

d: Panel or Longitudinal Data

- The **same** cross-sectional units are followed over time.
  - Latitude vs. Longitude [figure here]
- Panel data have a cross-sectional and a time series dimension.
- Panel data can be used to account for time-invariant unobservables.
- Panel data can be used to model lagged responses.
- Example:
  - City crime statistics; each city is observed in two years.
  - Time-invariant unobserved city characteristics may be modeled.
  - Effect of police on crime rates may exhibit time lag.

Latitude and Longitude



Two-Year Panel Data on City Crime Statistics

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

Each city has two time series observations

Number of police in 1986

Number of police in 1990

Causality and the Notion of Ceteris Paribus in Econometric Analysis

Causality and the Notion of Ceteris Paribus in Econometric Analysis

- Definition of causal effect of x on y:

  How does variable y change if variable x is changed
  but all other relevant factors are held constant.

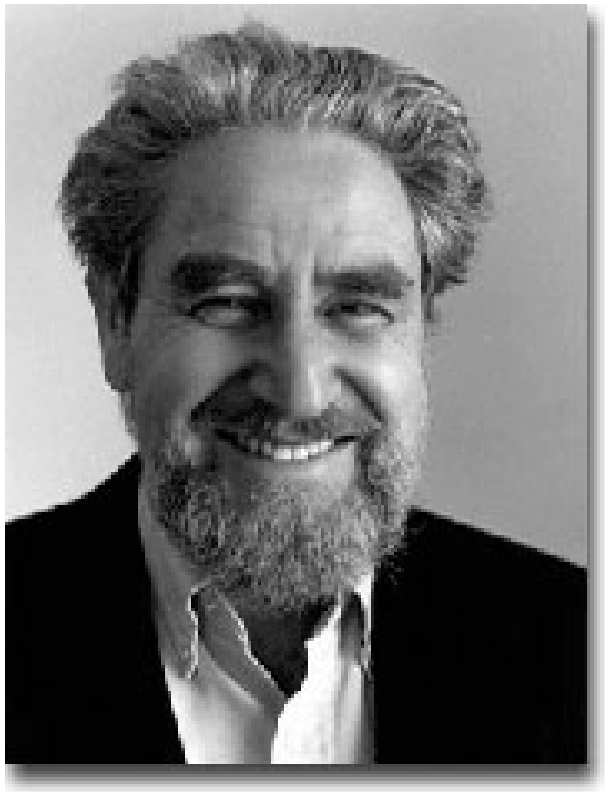
- Most economic questions are ceteris paribus questions.
- It is important to define which causal effect one is interested in.
- It is useful to describe how an experiment would have to be designed to infer the causal effect in question.
- Let's check four examples below; the first three deal with cross-sectional data at various levels of aggregation (e.g., at the individual or city levels), and the last one deals with time series data.
- The key point for all examples is that we need to randomly assign x such that all other relevant factors are balanced. (see Chapter 2 for a rigorous treatment).

Effects of Fertilizer on Crop Yield

- The question: By how much will the production of soybeans increase if one increases the amount of fertilizer applied to the ground?
- Implicit assumption: all other factors that influence crop yield such as quality of land, rainfall, presence of parasites etc. are held fixed.
- Experiment:
  - choose several one-acre plots of land;
  - randomly assign different amounts of fertilizer to the different plots;
  - compare yields.
- Experiment works because amount of fertilizer applied is unrelated to other factors influencing crop yields.

History of Production Function Estimation

- Griliches, Z., 1957, Specification Bias in Estimates of Production Functions, *Journal of Farm Economics*, 39, 8-20.



Zvi Griliches (1930-1999),
Harvard, 1957ChicagoPhD

Measuring the Return to Education

- The question: If a person is chosen from the population and given another year of education, by how much will his or her wage increase?
- Implicit assumption: all other factors that influence wages such as experience, family background, intelligence etc. are held fixed.
- Experiment:
  - choose a group of people;
  - randomly assign different amounts of education to them (infeasible!);
  - compare wage outcomes.
- Problem without random assignment: amount of education is related to other factors that influence wages (e.g. intelligence).

Effect of Law Enforcement on City Crime Levels

- The question: If a city is randomly chosen and given ten additional police officers, by how much would its crime rate fall?
  - Alternatively: If two cities are the same in all respects, except that city A has ten more police officers, by how much would the two cities crime rates differ?
- Experiment:
  - randomly assign number of police officers to a large number of cities;
  - compare crime rates.
- In reality, number of police officers will be determined by crime rate (simultaneous determination of crime and number of police).

Effect of the Minimum Wage on Unemployment

- The question: By how much (if at all) will unemployment increase if the minimum wage is increased by a certain amount (holding other things fixed)?
- Experiment:
  - Government randomly chooses minimum wage each year and observes unemployment outcomes;
- Experiment will work because level of minimum wage is unrelated to other factors determining unemployment;
- In reality, the level of the minimum wage will depend on political and economic factors that also influence unemployment.

Testing Predictions of Economic Theories

- Economic theories are not always stated in terms of causal effects, but they often have predictions that can be tested using econometric methods.
- For example, the expectations hypothesis states that long term interest rates equal compounded **expected** short term interest rates,

$$(1+r_{lt})^n = \left(1+r^e_{year\,1}\right)\left(1+r^e_{year\,2}\right)\cdot\cdots\cdot\left(1+r^e_{year\,n}\right),$$

  where r_{lt} is the long term interest rate, and $r^e_{year\,i}$ is the expected interest rate of year i.
- An implication is that the interest rate of a three-months T-bill should be equal to the expected interest rate for the first three months of a six-months T-bill; otherwise, there would be arbitrage (why?).
- This implication can be tested using econometric methods.