Lecture 1 Introduction to econometrics

Xi Qu

Fall 2019

What is Econometrics?

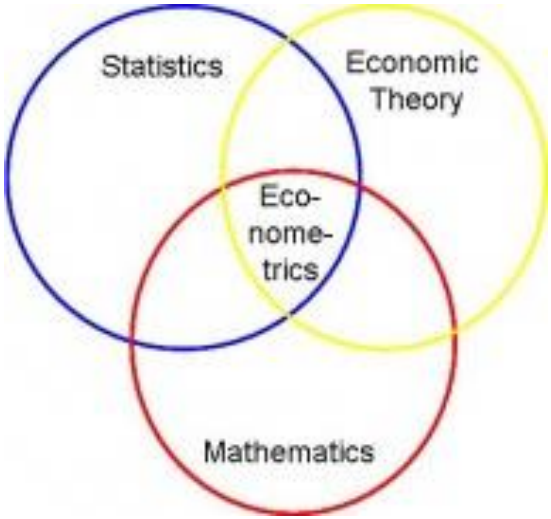
- Econometrics is the application of mathematics, **statistical methods**, and computer science, to economic data and is described as the branch of economics that aims to give **empirical content** to economic relations.

- Empirical research is a way of gaining knowledge by means of direct and indirect observation or experience. Empirical evidence (the record of one's direct observations or experiences) can be analyzed quantitatively or qualitatively.

- Econometrics allows economists to sift through mountains of data to extract simple relationships. (From Wiki)

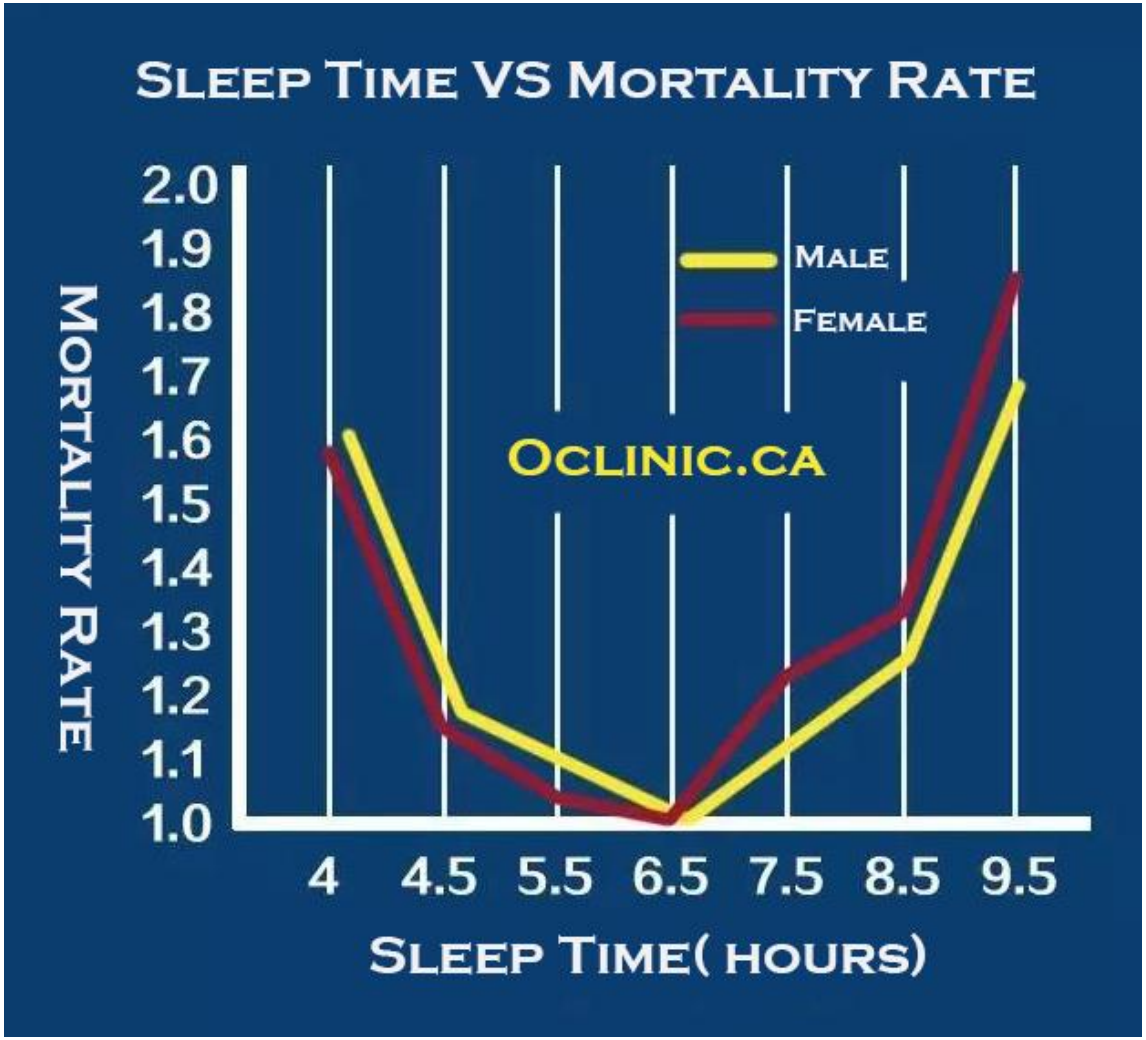
What is Econometrics?

- Econometrics, literally "**economic measurement**," involves quantitative analysis of economic problems.
- It is the application of statistical methods to connect theoretical economic models to data.
- It is a bridge between abstract economic theory and real-world human economic activity.



What's you opinion of this article?

http://www.sohu.com/a/190976004_355456
http://www.orientalacupuncture.ca/2017/09/



Why do we need to learn econometrics?

- http://www.zhihu.com/question/22052661
- https://www.quora.com/How-useful-is-econometrics

- Metrics are not useful (and sometimes downright misleading) without context. However, it's much more useful than simply saying 'about', 'slightly more' or 'a lot more'.

- Econometrics is useful if we use it well.

Usefulness of econometrics

- A racial-discrimination lawsuit filed against Harvard by Students for Fair Admissions
- It is still on trial…

- Expert witnesses:
  - David Card (Berkeley, hired by Harvard, $750/hr)
  - Peter Arcidiacono (Duke, hired by SFFA, $450/hr)

- Both were given private admissions data and tasked with determining whether Harvard University discriminates against Asian Americans.

- "None of Professor Card's arguments are persuasive. His modeling choices are inconsistent with standard econometric practice and appear designed to understate the effect of race in the admission process generally, and on Asian-American applicants specifically…"

Comments of Econometrics

- **Hardness**
- "Econometrics is too mathematical; it's the reason my best friend isn't majoring in economics"
- "There are two things you are better off not watching in the making: sausages and econometric estimates"
- "For some reason, I always wanted to take econometrics in college. After reading the first page of this thread, thank god I didn't"

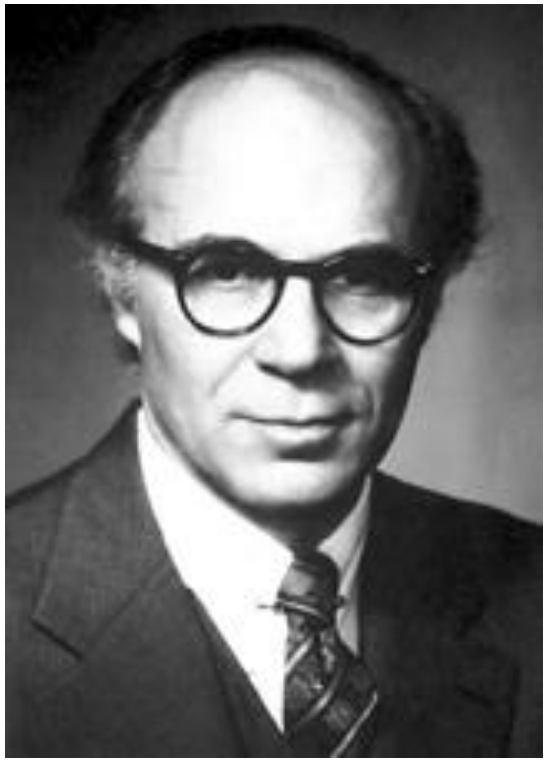
- **Importance**
- In economics today, with the exception of pure theory, almost every academic article and government study uses econometrics of some kind.
- Econometric skills play a special role in the training of economists. See **Nobel Laureates in Economics**
- https://www.nobelprize.org/nobel_prizes/facts/economic-sciences/index.html

Nobel laureates in Economics

- Econometric Fellows and Nobel Laureates in Economics: Of all 69 Nobel Prize Laureates between 1969 and 2011, only 9 of them were not also Fellows.

- – Nobel 1969, Jan Tinbergen, Ragnar Frisch;
- – Nobel 1980, Lawrence Robert Klein;
- – Nobel 1989, Trygve Haavelmo;
- – Nobel 2000, James Heckman, Daniel McFaddan;
- – Nobel 2003, Robert Engle, Clive Granger.

- Lawrence Klein received the 1980 Nobel Price “for the creation of econometric models and the application to the analysis of economic fluctuations and economic policies”.



Lawrence Klein
USA
Upenn

ECONOMETRICIANS

- Thomas Sargent and Christoper Sims received the 2011 Nobel Price in economics “for their empirical research on cause and effect in the macroeconomy”, more specifically, for developing new econometric models for estimating the link between economic policy measures and macroeconomic variables like GDP, inflation, and unemployment.

Photo: U. Montan

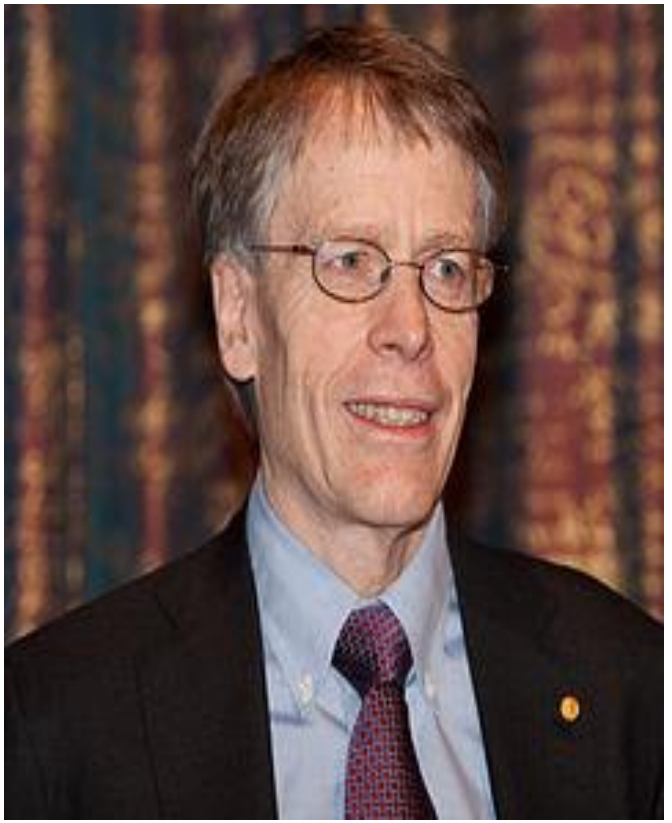
Thomas J. Sargent

Photo: U. Montan

Christopher A. Sims

2013 Lars Peter Hansen

- Hansen is best known as the developer of the econometric technique generalized method of moments (GMM) and has written and co-authored papers applying GMM to analyze economic models in numerous fields including labor economics, international finance, finance, and macro economics.



2015 Angus Deaton

- His research deals with how consumption depends on prices, puts consumption in relation to savings and incomes, and shows how **data on consumption** can be used to analyze welfare, poverty, and economic development.

Features of Econometrics

Econometrics is an integration of economic theory and statistics.

Different from Pure Economic theory

--Economic theory makes statements or hypotheses that are mostly qualitative in nature.

--Econometrics provides numerical estimates, gives empirical content to most economic theory.

Different from Statistics:

--**Statistics** is mainly concerned with collecting, processing, and presenting data in the form of charts and tables.

--**Econometrics**: being concerned with using the collected data to test economic theories.

Examples (no theory):

http://mt.sohu.com/20150629/n415818814.shtml

Gong, Li Zhang, Ziyi Li, Lianjie Li, Xiaolong Zhou, Jielun Xi, Jinping Peng, Liyuan

Ice cream vs drowning death

Different from Mathematical economics

--**Mathematical economics:** Main concern is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory.

--**Econometrics**: Mainly interested in the empirical verification of economic theory. And this conversion of mathematical into econometric equations requires a great deal of ingenuity and practical skill.

QUESTION.

- 1.Econometrics is the branch of economics that ___.

a. studies the behavior of individual economic agents in making economic decisions

b. develops and uses statistical methods for estimating economic relationships

c. deals with the performance, structure, behavior, and decision-making of an economy as a whole

d. applies mathematical methods to represent economic theories and solve economic problems.

- 2. An empirical analysis relies on _____to test a theory.

a. common sense

b. ethical considerations

c. data

d. customs and conventions

METHODOLOGY OF ECONOMETRICS

Traditional econometric methodology (empirical study)proceeds along the following lines:

STEP1**.** Statement of theory or hypothesis.

STEP2. Specification of the mathematical model of the theory.

STEP3. Specification of the statistical, or econometric model

STEP4. Obtaining the data.

STEP5. Estimation of the parameters of the econometric model.

STEP6. Hypothesis testing.

STEP7. Forecasting or prediction.

STEP8. Using the model for control or policy.

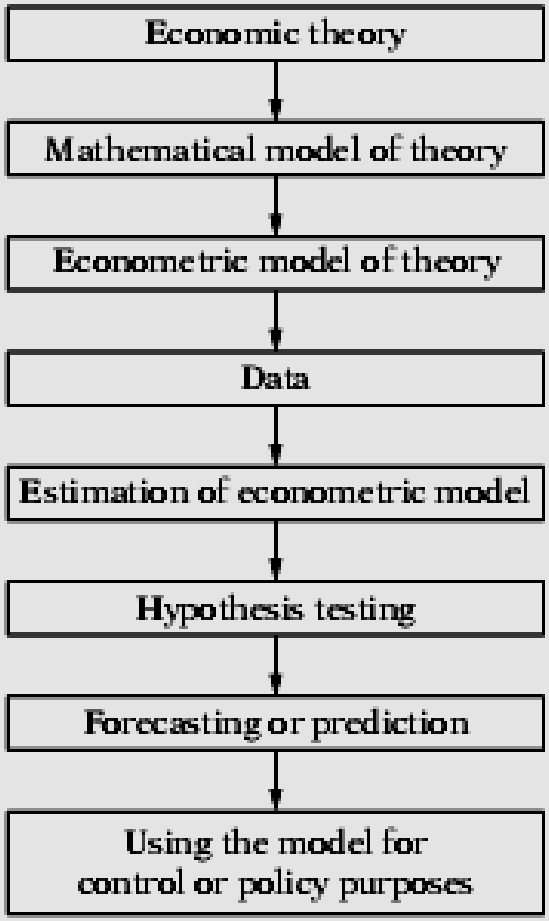


FIGURE I.4 Anatomy of econometric modeling.

STEP1. STATEMENT OF THEORY OR HYPOTHESIS

- **One Example** The well-known Keynesian theory of consumption

- Keynes postulated that the marginal propensity (natural tendency) to consume (MPC), the rate of change of consumption for a unit (say, a dollar) change in income, is greater than zero but less than 1.

Step 2. Specification of the Mathematical Model of Consumption

- Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of the functional relationship between the two.

- For simplicity, a mathematical economist might suggest the following form of the Keynesian consumption function:

$$Y = \beta_1 + \beta_2 X, \quad 0 < \beta_2 < 1 \qquad (1)$$

β_1 , β_2 : parameters of the model,

Y: consumption (expenditure) : dependent variable

X: income : explanatory variable.

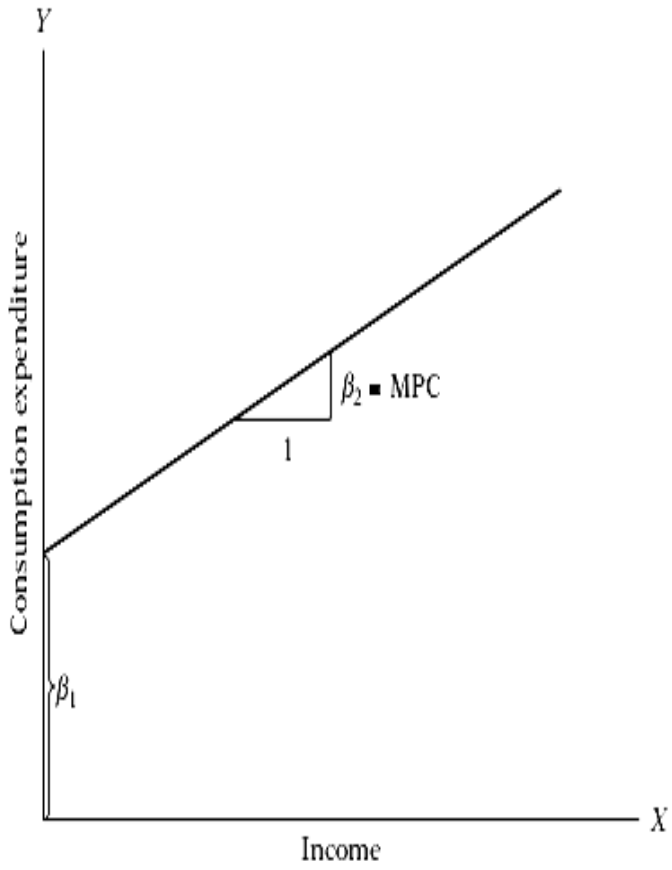


FIGURE I.1 Keynesian consumption function.

Keynesian theory of consumption

$ß_1$: the intercept.

$ß_2$:slope coefficients.

$ß_2$: The slope coefficient measures the MPC.

STEP3. SPECIFICATION OF THE ECONOMETRIC MODEL OF CONSUMPTION

- --To allow for the inexact relationships between economic variables, mathematical models should be modified into econometric models, by adding a disturbance, or error, term.

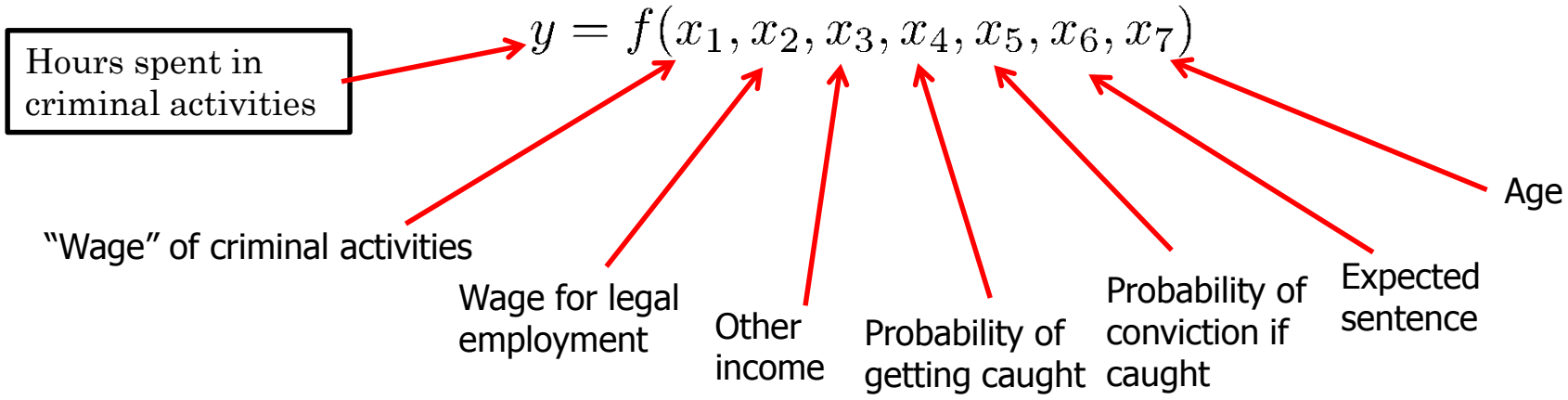
--**The Error Reasons**: in addition to income, other variables affect consumption expenditure. For example, size of family, ages of the members in the family, family religion, etc., are likely to exert some influence on consumption.

The econometric model: $Y = \beta_1 + \beta_2 X + u$ (2)

u: known as the disturbance, or error term, is a random (stochastic) variable that has well-defined probabilistic properties.

Example of Crime

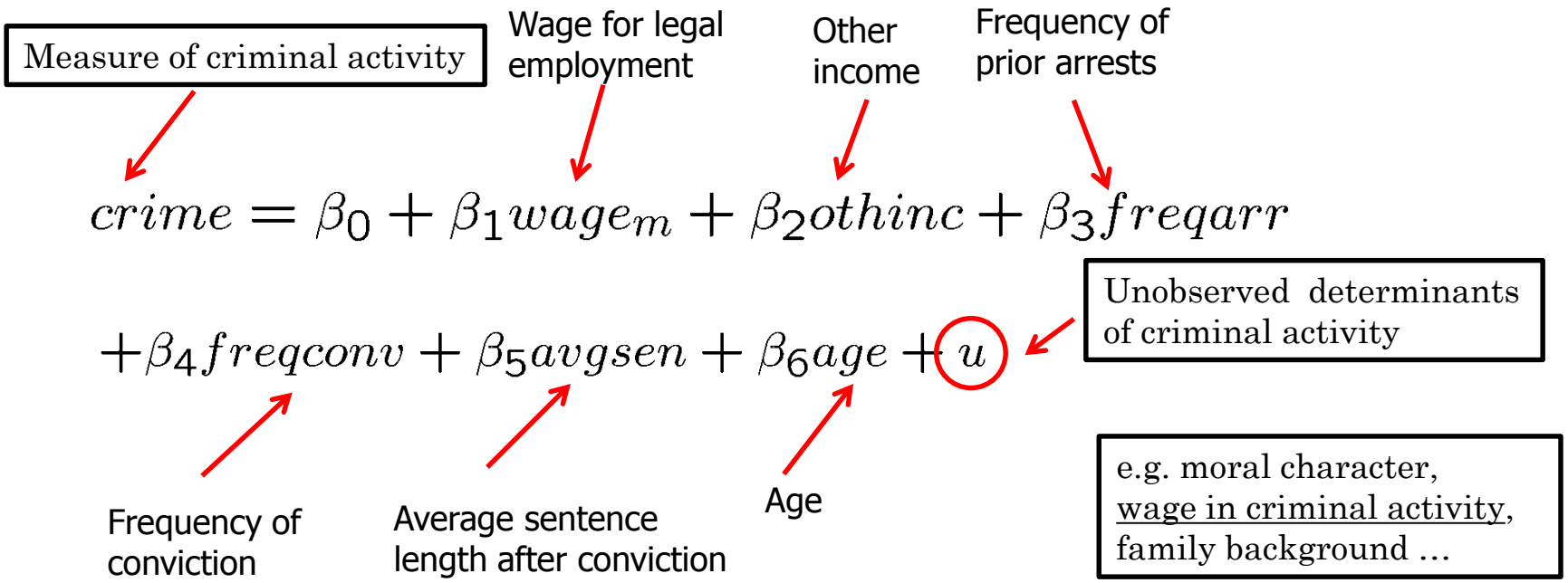
- **Economic model of crime (Becker (1968))**
  - Derives equation for criminal activity based on utility maximization



  - Functional form of relationship not specified
  - Equation could have been postulated without economic modeling

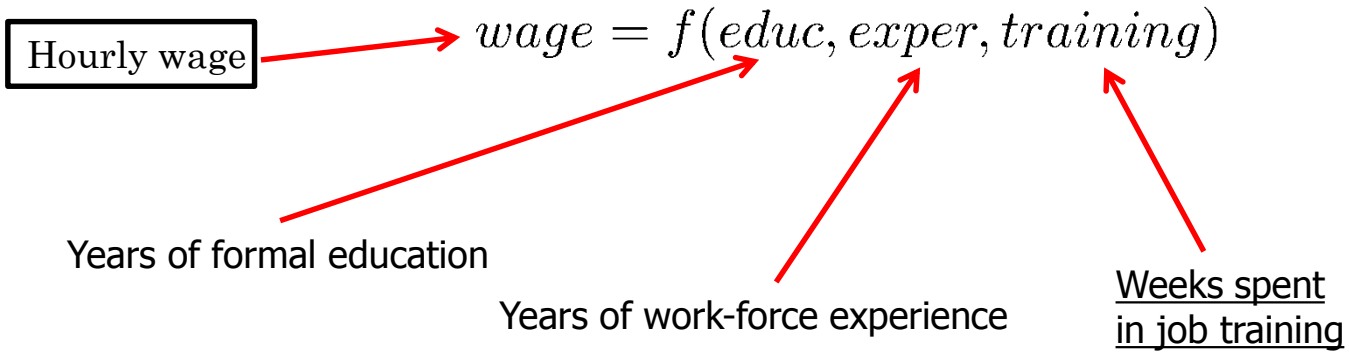
Econometric model of criminal activity

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



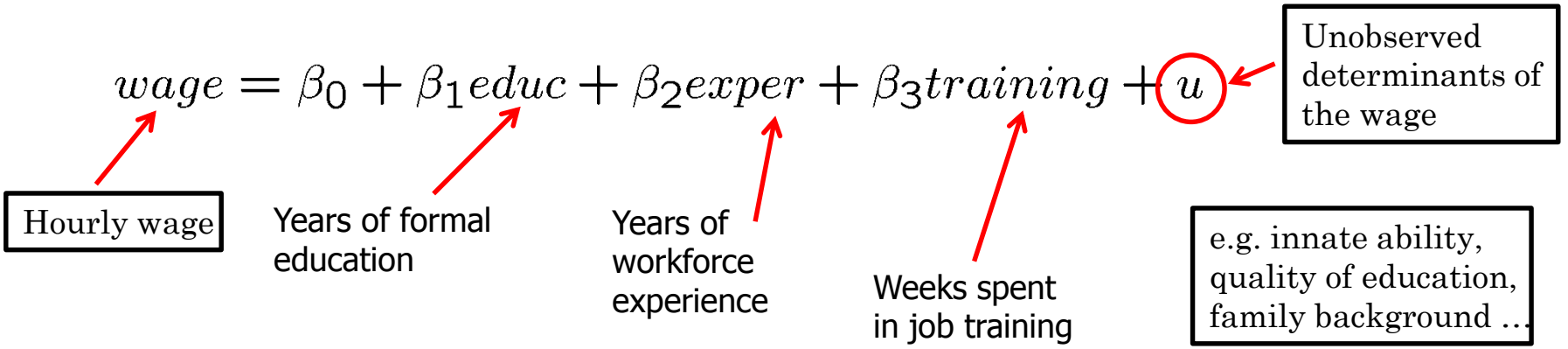
Example of Job Training

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



  - Other factors may be relevant, but these are the most important (?)

- **Econometric model of job training and worker productivity**



- **Most of econometrics deals with the specification of the error** u
- **Econometric models may be used for hypothesis testing**
  - For example, the parameter β_3 represents effect of training on wage
  - How large is this effect? Is it different from zero?

Question

- 3. The term ‘u’ in an econometric model is usually referred to as the _____.

a. error term

b. parameter

c. hypothesis

d. dependent variable

- 4. The parameters of an econometric model _____.

a. include all unobserved factors affecting the variable being studied

b. describe the strength of the relationship between the variable under study and the factors affecting it

c. refer to the explanatory variables included in the model

d. refer to the predictions that can be made using the model

STEP4. OBTAINING DATA

- Observation mechanisms
  - Passive, nonexperimental
  - Active, experimental
  - The 'natural experiment'

- Data types
  - Cross section
  - Pure time series
  - Pooled data

    A special type of pooled data: Panel – longitudinal data

Data Source

- First hand data
- Three Steps:
1. Questionnaire construction
2. Sample Selection
3. Data collection
- Second hand data
- Many macro and industrial data are available online or from yearbooks.
- Data on individuals, households, and enterprises are usually difficult to obtain.

Types of Data – Cross Sectional

- 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
- For example, **pure random sampling** from a population

- Sometimes, pure random sampling is violated, e.g. units refuse to respond in surveys, or if sampling is characterized by clustering

- Cross-sectional data typically encountered in applied microeconomics

Examples of Cross-sectional Data

- **Cross-sectional data set 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

Types of Data – Time Series

- Time series data has a separate observation for each time period – e.g. stock prices
- Data frequency: daily, weekly, monthly, quarterly, annually, …
- Time series observations are typically **serially correlated** (not a random sample)
- Ordering of observations conveys important information
- Trends and seasonality will be important
- Typical applications: applied macroeconomics and finance

Examples of Time Series Data

- **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

Types of Data – Pooled cross sections

- Two or more cross sections are combined in one data set
- Cross sections are drawn independently of each other

  - 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)

EXAMPLE OF POOLED CROSS SECTION DATA

- **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

Panel or longitudinal data

- The same cross-sectional units are followed over time
- Panel data have a cross-sectional and a time series dimension

  - 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

EXAMPLE OF PANEL DATA

- **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

Question

- 5. A data set that consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time, is called a(n) _____.

a. cross-sectional data set

b. longitudinal data set

c. time series data set

d. experimental data set

- 6. Which of the following is an example of time series data?

a. Data on the unemployment rates in different parts of a country during a year.

b. Data on the consumption of wheat by 200 households during a year.

c. Data on the gross domestic product of a country over a period of 10 years.

d. Data on the number of vacancies in various departments of an organization on a particular month.

Question

- 7. Which of the following refers to panel data?

a. Data on the unemployment rate in a country over a 5-year period

b. Data on the birth rate, death rate and population growth rate in developing countries over a 10-year period.

c. Data on the income of 5 members of a family on a particular year.

d. Data on the price of a company's share during a year.

- 8. Which of the following sections usually comes first in a research report?

a. state of problem

b. description of data

c. review of literature on the topic

d. economic model

Data clean

- Browse the data, looking for abnormal numbers.
- Missing data, outliers

- Calculate summary statistics
- n, min, max, mean, variance, standard deviation, range, skewness, kurtosis, quantiles (median)

- Distributions of variables
- Frequency distribution, histogram, pdf, cdf

- Relationship between variables
- X-y diagram, covariance, correlation coefficient

STEP5. ESTIMATION OF THE ECONOMETRIC MODEL(ESSENTIAL PART)

- The statistical technique of regression analysis is the main tool used to obtain the estimates.

- Historical origin of the term "*Regression*":
- This term was introduced by Francis Galton (1886).
- Galton found that although there was a tendency for **tall parents** to have **tall children** and for **short parents** to have **short children**, the average height of children born of parents of a given height tended to **move or "regress" toward the average height** in the population as a whole.
- In other words, the average height of sons of a group of tall fathers was less than their fathers' height and the average height of sons of a groups of short fathers was greater than their fathers' height.

STEP6. HYPOTHESIS TESTING

- **Keynes expected the MPC to be positive but less than 1.**

- In our example, $\tilde{Y}_i = -184.08 + 0.7064X_i$ (3), so the estimate of the MPC is about 0.70.

- Before we accept this finding as confirmation of Keynesian consumption theory, we must inquire whether this estimate is sufficiently below unity to convince us that this is not a chance occurrence or peculiarity of the particular data we have used.

- In other words, is 0.70 statistically less than 1? If it is, it may support Keynes' theory.

STEP7. FORECASTING OR PREDICTION

- If the chosen model does not refute the hypothesis or theory under consideration, we may use it to predict the future value(s) of the dependent variable Y on the basis of known or expected future value(s) of the explanatory, or predictor, variable X.

- To illustrate, suppose we want to predict the mean consumption expenditure for 1997. The GDP value for 1997 was 7269.8 billion dollars. Putting this GDP figure on the right-hand side of (3), we obtain:

- Y1997 =-184.0779+0.7064(7269.8) = 4951.3167 (4)

STEP8. USE OF THE MODEL FOR CONTROL OR POLICY PURPOSES

- Ex. 1. Government is interested in estimating returns (payoff) to education. Only worth investing if benefit > cost of provision.
- Ex. 2. Companies need to know whether and by how much increased advertising expenditure generates increased sales.

- Suppose further the government believes that consumer expenditure of about 4900 (billions of 1992 dollars) will keep the unemployment rate at its current level of about 4.2 percent (early 2000). What level of income will guarantee the target amount of consumption expenditure?
- 4900 =-184.0779+0.7064X
- X = 7197, an income level of about 7197 (billion) dollars, will produce an expenditure of about 4900 billion dollars.

Importance of Causality

- **Causality and the notion of ceteris paribus**

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**

Examples of Causal Effect

- **Causal effect of fertilizer on crop yield**
  - "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

- **Measuring the return to education**
  - "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)

Question

- 9. Which of the following is true?

a. A variable has a causal effect on another variable if both variables increase or decrease simultaneously.

b. The notion of 'ceteris paribus' plays an important role in causal analysis.

c. Difficulty in inferring causality disappears when studying data at fairly high levels of aggregation.

d. The problem of inferring causality arises if experimental data is used for analysis.

- 10. Of the following steps in conducting empirical economic research, which one should be performed last?

a. Find appropriate data that can be used for estimation

b. Build an economic model guided by economic theory

c. Evaluate and analyze the consequences and implications of the results

d. Estimate parameters and test hypotheses

After today's class, What's you opinion of this picture?