1

INTRODUCTION:
Purpose, Scope, and Plan

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

2

Classical Methodology in Econometrics

- Formulation of theory or hypothesis,
- Specification of economic (mathematical) model,
- Specification of econometric model,
- Collecting data,
- Estimation of parameters,
- Hypothesis tests,
- Forecasting/Prediction,
- Evaluation of results for policy analysis or decision making

3

Economic Model

- The first step in an econometric analysis is to pose the question we wish to answer as precisely as possible.
- In most cases the question has important policy implications. For example,
- How does the class size affect student performance in primary schools? If the minimum wage increases by 10%, how much does unemployment (or employment) change? If the severity of punishment for certain crimes increases, do crime rates fall on average?
- The second step is to specify an economic model or a conceptual framework with careful economic reasoning. (utility maximization, profit maximization, cost minimization, etc.)

4

Example: An Economic Model of Crime

- Gary Becker is a famous economist and a Nobel prize winner who analyzed several topics that were once considered to be outside economics such as crime, racial discrimination, marriage and family organization using utility maximization framework.
- In his crime model, Becker considers an individual who decides how much time to spend in illegal activity (crime).
- We can think of this as a supply function for time in illegal activity. Of course this supply function includes zero, i.e, not spending any time in illegal activity.
- We can ask several research questions in an economic model. The next slide describes the formal economic model of crime.

ECONOMIC MODEL

Example 1 - Economic Model of Crime

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

$f()$ functional form (not yet specified)

Description of variables

- y: hours spent in criminal activities,
- x_1: "earnings" for an hour spent in criminal activity,
- x_2: hourly wage in legal employment,
- x_3: other income,
- x_4: probability of getting caught,
- x_5: probability of being convicted if caught,
- x_6: expected sentence if convicted,
- x_7: age

ECONOMIC MODEL vs. ECONOMETRIC MODEL

Economic Model Example 2 - Job Training and Worker Productivity

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

wage: hourly wage (in dollars)
educ: level of education (in years)
exper: level of workforce experience (in years)
training: weeks spent in job training.

Econometric Model: f Linear specification

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

ECONOMETRIC MODEL: Linear Specification

Econometric Model Example 2 - Job Training and Worker Productivity

$$wage = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 educ + \boldsymbol{\beta}_2 exper + \boldsymbol{\beta}_3 training + \boldsymbol{u}$$

Components of econometric model:

- **$\boldsymbol{u}$: random error term or disturbance term**
- Random error term u contains influence of factors that are not included in the model. It also contains unobserved factors such as innate ability or family background.
- No matter how comprehensive the specified model, there will always be factors that cannot be included in the econometric model. We can never eliminate u entirely (otherwise, we have a deterministic model which is not realistic).

ECONOMETRIC MODEL: Linear Specification

Econometric Model Example 2 - Job Training and Worker Productivity

$$wage = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 educ + \boldsymbol{\beta}_2 exper + \boldsymbol{\beta}_3 training + \boldsymbol{u}$$

Components of econometric model:

- $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$: parameters of the econometric model
- These are unknown constants.
- They describe the directions and strengths of the relationship between wage and factors affecting wage included in the model.
- For example, we may be interested in testing $H_0 : \beta_3 = 0$ which says that job training has no effect on wage.
- Because we do not know these beta coefficients, we need to collect data and estimate them using the methods we will learn in this class.

ECONOMETRIC DATA TYPES

The types of data that we use in econometric applications can be classified as follows:

- **Cross-sectional data**
- **Time series data**
- **Pooled cross-section**
- **Panel data (longitudinal data)**

Cross-sectional data

- consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time
- Significant feature: random sampling from a target population
- Generally obtained through official records of individual units, surveys, questionnaires (data collection instrument that contains a series of questions designed for a specific purpose)
- For example, household income, consumption and employment surveys conducted by the Turkish Statistical Institute (TUIK/TURKSTAT)

Cross-sectional data example: Wage Data (GRETL data set: wage1.gdt)

A Cross-sectional data set on wages and individual characteristics

Obs. No	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
6	8.75	16	9	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
524	4.67	15	13	0	1
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Time series data

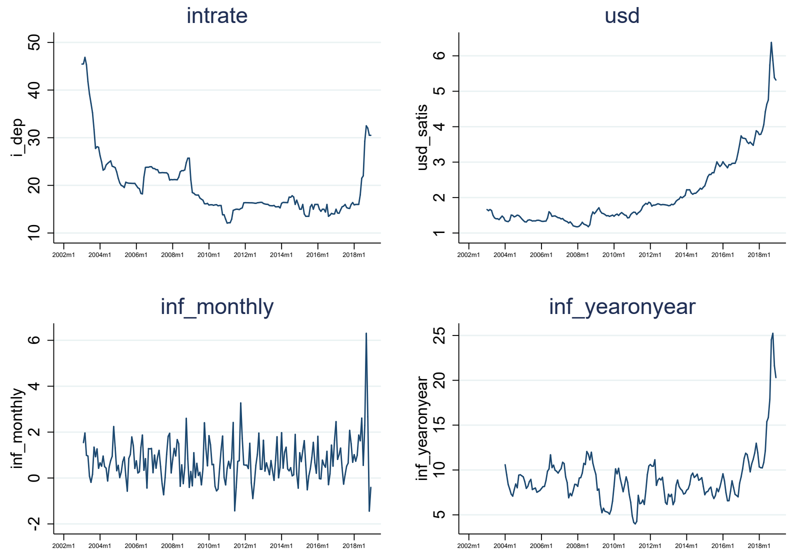
- consists of observations on a variable or several variables over time.
- Chronological ordering
- Frequency of time series data: hour, day, week, month, year
- Time length between observations is generally equal
- Examples of time series data include stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, and automobile sales figures.

A Time Series Data Example: Selected Turkish Monthly Macroeconomic Variables 2003m01-2018m12

(first 2 years displayed)

obsno	year	month	intrate	usd	cpi	inf_monthly	inf_yearon~r
1	2003	1	45.46	1.663822	94.77	.	.
2	2003	2	45.48	1.6310601	96.23	1.540572	.
3	2003	3	46.87	1.663371	98.12	1.964044	.
4	2003	4	45.32	1.6350447	99.09	.9885854	.
5	2003	5	41.75	1.4975657	100.04	.9587244	.
6	2003	6	39.17	1.4259443	100.12	.079968	.
7	2003	7	37.24	1.4038464	99.93	-.1897723	.
8	2003	8	35.25	1.403841	100.09	.1601121	.
9	2003	9	31.74	1.3789644	101.44	1.348786	.
10	2003	10	27.74	1.4281261	102.38	.9266561	.
11	2003	11	28.12	1.479556	103.68	1.269779	.
12	2003	12	28	1.4360091	104.12	.4243827	.
13	2004	1	26.22	1.3499002	104.81	.6626969	10.59407
14	2004	2	24.93	1.3302146	105.35	.515218	9.477294
15	2004	3	23.19	1.3222042	106.36	.9587091	8.397881
16	2004	4	23.41	1.3578829	106.89	.4983076	7.871632
17	2004	5	24.27	1.5085833	107.35	.430349	7.307077
18	2004	6	24.61	1.4959256	107.21	-.1304145	7.081502
19	2004	7	24.88	1.4552241	107.72	.4757019	7.795457
20	2004	8	25.14	1.4744442	108.54	.7612329	8.442402
21	2004	9	24.02	1.50557	109.57	.9489589	8.01459
22	2004	10	23.9	1.491942	112.03	2.24514	9.425669
23	2004	11	23.73	1.452395	113.5	1.312149	9.471451
24	2004	12	22.81	1.4000632	113.86	.3171806	9.35459
25	2005	1	21.5	1.3565421	114.49	.5533111	9.23576

Time Series Plot: 2003m01-2018m12



A Time Series Data Example: Selected Turkish Monthly Macroeconomic Variables

Variable Definitions:

- intrate: average interest rates on 3-month bank deposits
- usd: Exchange rate, USD/TL (Turkish Lira per US Dollar)
- cpi: Consumer Price Index (seasonally unadjusted)
- inf_monthly: percent change in cpi from the previous month

$$inf\_monthly = 100 * \frac{(cpi - cpi(-1))}{cpi(-1)}$$

  where $cpi(-1)$ is the cpi in the previous month (note that the first observation is missing)
- inf_yearonyear: percent change in cpi from the same month in the previous year

$$inf\_yearonyear = 100 * \frac{(cpi - cpi(-12))}{cpi(-12)}$$

  where $cpi(-12)$ is the cpi in the previous year (note that the first 12 observations are missing)

Pooled cross-section

- consists of cross-sectional data sets that are observed in different time periods and combined together
- At each time period (e.g., year) a different random sample is chosen from population
- Individual units are not the same
- For example if we choose a random sample 400 firms in 2002 and choose another sample in 2010 and combine these cross-sectional data sets we obtain a pooled cross-section data set.
- Cross-sectional observations are pooled together over time.

A Pooled Cross-sectional Data Example

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

Panel Data (longitudinal data)

- consists of a time series for each cross-sectional member in the data set.
- The same cross-sectional units (firms, households, etc.) are followed over time.
- For example: wage, education, and employment history for a set of individuals followed over a ten-year period.
- Another example: cross-country data set for a 20 year period containing life expectancy, income inequality, real GDP per capita and other country characteristics.

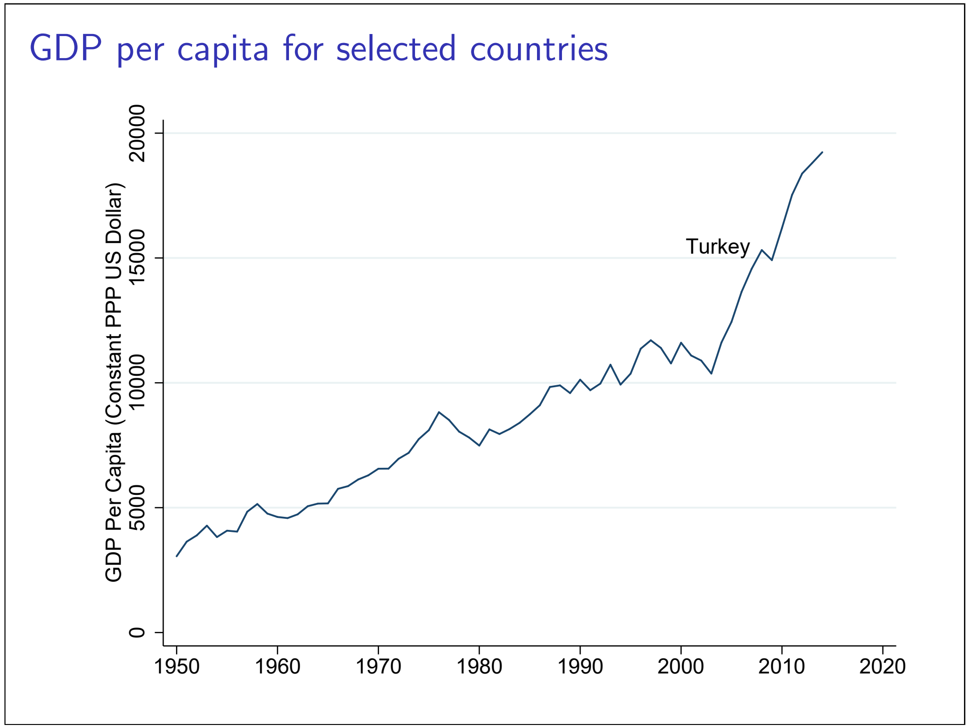
A Panel Data Example

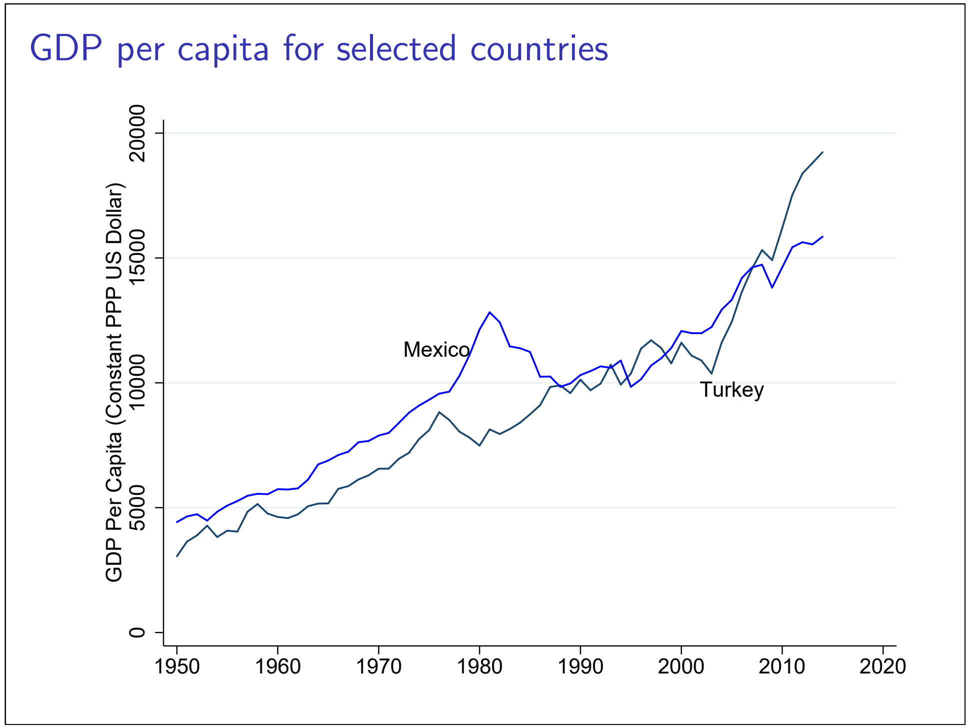
A Two-Year Panel Data Set on City Crime Statistics

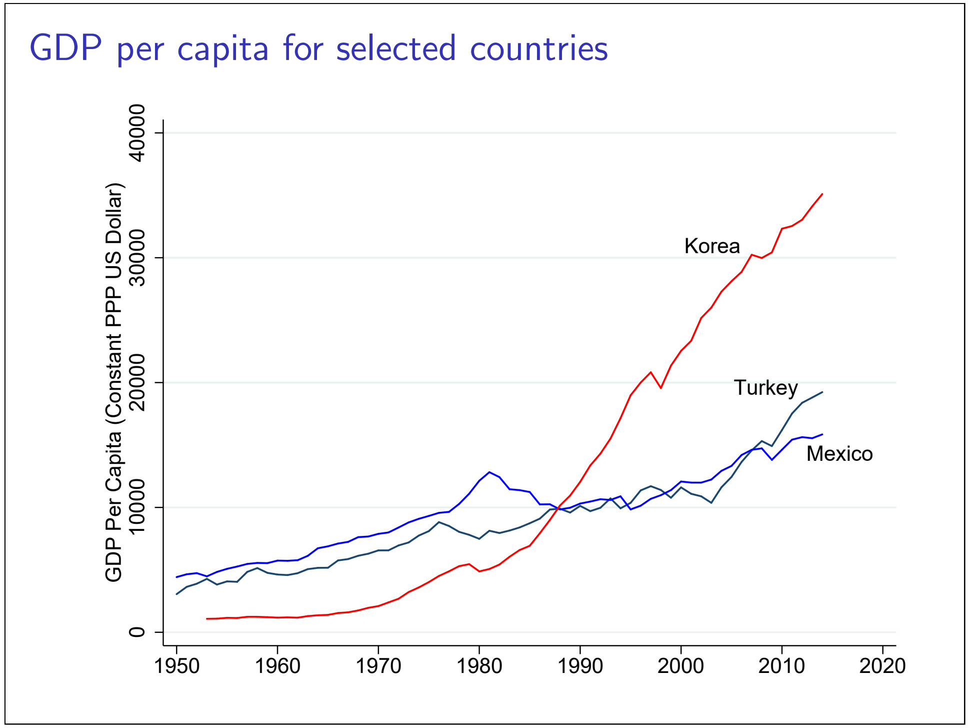
obsno	*city*	*year*	*murders*	*population*	*unem*	*police*
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150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

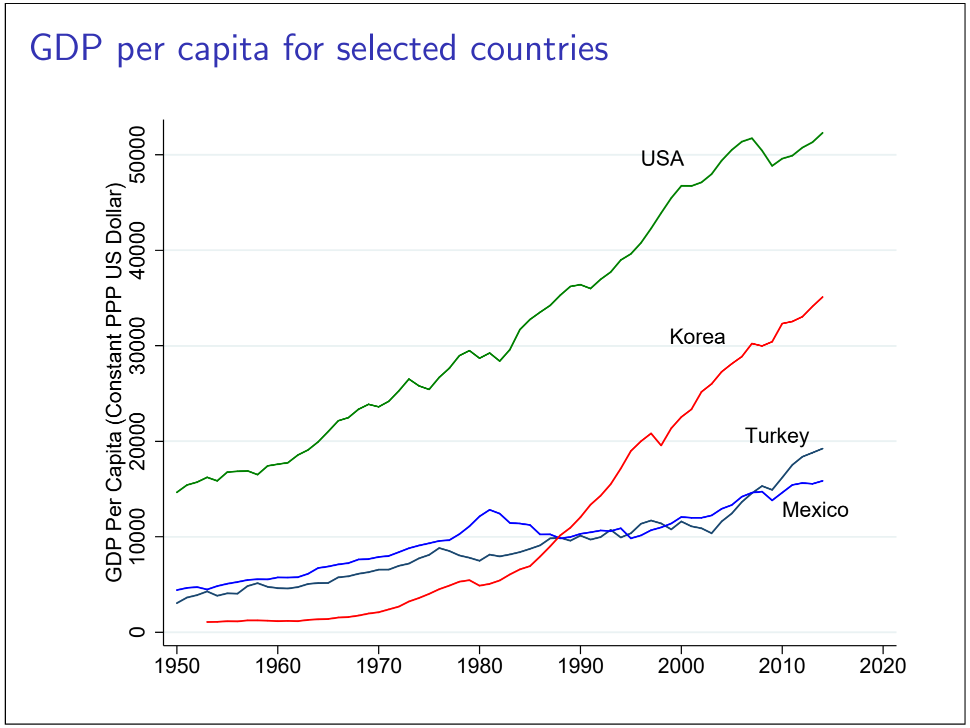
Panel Data (longitudinal data)

- In practice, we encounter two types of panel data: micro panels and macro panels
- Micro panels (large N small T): we have a large number (N) of cross-sectional units (consumers, firms, etc.) but small number of time periods (T)
- Macro panels (small N large T): we have a small number of cross-sectional units (e.g. countries) but large number of time periods (e.g., 50 years of observations)
- The following graphs display time series plots of GDP per capita for 4 countries.
- This data set can easily be extended by incorporating additional macroeconomic variables (price level, exchange rates, unemployment rate, etc.) as well as cross-sectional units (in this case, country).


GDP per capita for selected countries
GDP Per Capita (Constant PPP US Dollar)
0
5000
10000
15000
20000
1950
1960
1970
1980
1990
2000
2010
2020
Turkey


GDP per capita for selected countries
GDP Per Capita (Constant PPP US Dollar)
0
5000
10000
15000
20000
1950
1960
1970
1980
1990
2000
2010
2020
Mexico
Turkey


GDP per capita for selected countries
GDP Per Capita (Constant PPP US Dollar)
0
10000
20000
30000
40000
1950
1960
1970
1980
1990
2000
2010
2020
Korea
Turkey
Mexico


GDP per capita for selected countries
GDP Per Capita (Constant PPP US Dollar)
0
10000
20000
30000
40000
50000
1950
1960
1970
1980
1990
2000
2010
2020
USA
Korea
Turkey
Mexico

Causality and the Notion of Ceteris Paribus

- In testing economic theory usually our goal is to infer that one variable has a **causal effect** on another variable.
- Correlation may be suggestive but cannot be used to infer causality.
- You've probably heard the mantra "correlation is not causation": the rooster's crow in the morning does not cause the sun to rise (but they are highly correlated).
- Fundamental notion: **Ceteris paribus**: "other relevant factors being equal"
- Or "holding all other factors fixed"
- Most economic questions are ceteris paribus by nature.

Causality and the Notion of Ceteris Paribus

- For example, in analyzing consumer demand, we are interested in knowing the effect of changing the price of a good on its quantity demanded, while holding all other factors (such as income, prices of other goods, and individual tastes) fixed.
- If other factors are not held fixed, then we cannot know the causal effect of a price change on quantity demanded.
- Therefore, the relevant question in econometric analysis is "do we control sufficient number of factors?"
- Are there other factors that are not included in the model?
- Can we say that other components are held fixed?
- In most serious applications the number of factors is immense so the isolation of the effect of any particular variable may seem hopeless. But, if properly used, econometric methods can help us determine ceteris paribus effects.

Ceteris Paribus Example: Effects of Fertilizer on Crop Yield

- Suppose the crop is wheat. We are interested in measuring the impact of fertilizer on wheat yield (production).
- Obviously there are several factors that affect the production of wheat such as rainfall, quality of soil and presence of parasites.
- We need to control these factors in order to determine the ceteris paribus impact of fertilizers.
- To do this we can devise the following experiment: divide the land into equal pieces (such as one acre) and apply different amounts of fertilizer to each land plot and then measure the wheat yield.
- This gives us a cross-sectional data set where observation unit is land plot.
- We can apply statistical methods to this data set to measure the impact of fertilizers on crop yield.

Ceteris Paribus Example: Effects of Fertilizer on Wheat Yield

- How do we know the results of this experiment can be used to measure the ceteris paribus effect of fertilizer?
- Can we be sure that all other factors (quality of land plots for example) are held fixed?
- It is generally very difficult to observe the quality of soil.
- But we can still use ceteris paribus notion
- Amounts of fertilizers should be assigned to land plots independently of other plot features such as quality of plots
- In other words, other characteristics of plots should be ignored when deciding on fertilizer amounts.

29

Ceteris Paribus Example: Measuring the Return to Education

- Question: How can we measure 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?
- This is also a ceteris paribus question: all other factors are held fixed while another year of education is given to the person.
- There are several factors other than education that affect wages: experience, tenure, innate ability, gender, age, region, marital status, etc.

30

Example: Measuring the Return to Education

- Similar to fertilizer example we can design the following hypothetical experiment:
- Social planner has the ability to assign any level of eduction to any person.
- The planner chooses a random group of individuals from population and randomly assigns each person an amount of education: some are given high school education, some are given 4-year college education, some get none, etc.
- Subsequently the planner measures wages for each individual.
- This type of experiments sometimes called "Randomized Controlled Trials" (RCT)

31

Example: Measuring the Return to Education

- If levels of education are assigned independently of other personal characteristics that may affect productivity (such as innate ability or experience) then we can measure the impact of education on wages correctly.
- Notice that in such an experiment innate ability is uncorrelated with education levels.
- Of course such an experiment is impossible to conduct due to ethical and physical reasons. Just consider allocating random education levels to a random group of 6-year-old children and measuring their wage level 20 years later? Is it ethical?
- Even though we cannot obtain an experimental data (at least in the case of wage equation), we can obtain observational data set that contains information on wages, education, experience and other personal characteristics (e.g. from TUIK household employment surveys)

32

Ceteris Paribus Example: Measuring the Return to Education

- People choose their education levels. Thus, individual characteristics will be correlated with the level of education.
- For example, people with more innate ability tend to have higher levels of education.
- Workers with higher levels of education tend to have higher wages.
- It becomes difficult to isolate the impact of education from the impact of innate ability on wages.
- How much of this effect comes from education? How much from innate ability?

Ceteris Paribus Example: The Effect of Law Enforcement on City Crime Levels

- Does the presence of more police officers on the street deter crime?
- Ceteris paribus question: If a city is randomly chosen and given, say, ten additional police officers, by how much would its crime rates fall?
- Or: If two cities are the same in all respects, except that city A has ten more police officers than city B, by how much would the two cities' crime rates differ?

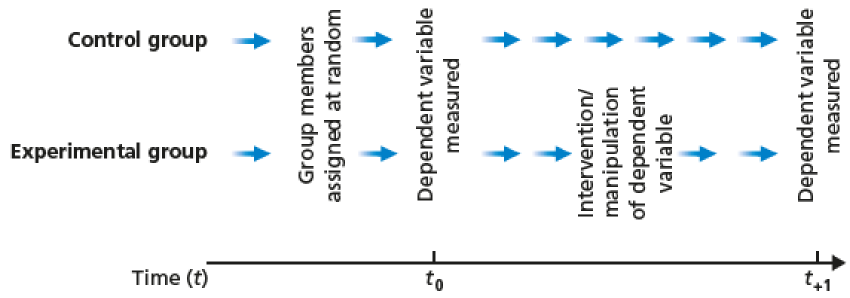
Ceteris Paribus Example: The Effect of Law Enforcement on City Crime Levels

- It almost impossible to find two cities identical in all respects.
- But this is not necessary in econometric analysis.
- We just need to know if the data on crime rates and number of police officers can be viewed as experimental.
- In most cases this is not the case, data is observational.
- The size of police force is determined by city authorities who probably take into account several other city characteristics.
- The problem is a little bit more complex: Does the size of police force affect the amount of crime or vice versa?
- The amount of crime and police force are simultaneously determined.

More on Randomized Controlled Trials (RCT)

- It is frequently dubbed that RCT is the gold standard in natural sciences to establish the causal relations among variables. This approach is routinely used in medicine and increasingly so in social sciences including economics, psychology, and sociology.
- In a basic RCT, the sample (individuals, or patients) are randomly assigned to two groups: **treatment** group and **control** group.
- The fundamental idea is that since the two groups are essentially the same, other than the treatment status, **all differences must be attributable to the treatment**.
- The treatment may be a new drug in a clinical study, or a new policy in social studies.
- The goal is to estimate the average treatment effect (ATE).
- The most crucial part of an RCT is randomization.

Classical Experimental Design



Note that the crucial factor in RCTs is the random assignment of subjects into treatment and control groups (Source of the figure: Saunders et al. (2015) Research Methods for Business Students, p.180)

Example: Do diets work?

- Randomized trial: randomly assign obese individuals to specific diets (Atkins, Ornish, etc.)
- Measure and record individual characteristics before implementing the diet
- Start the diet and follow individuals over a period of time (a year, or longer)
- At the end of the period measure weights and other risk factors (cholesterol levels, etc)
- Compare before and after
- H_0: there is no difference in average weights
- H_1: adherence to diets leads to weight loss

Development Economics Example: The Pakistan Early Child Development Scale Up Trial (UNICEF)

- Purpose: evaluate the effectiveness of integrating early childhood interventions (nutrition and child development advice etc.) to improve early child health outcomes
- Control group: received basic health and nutrition services
- Intervention group I: enhanced nutrition group received counseling, advice, and nutrition supplements.
- Intervention group II: early childhood development group received coached practice for child development including monthly visits and group meetings
- Intervention group III: received a combination of both the early childhood development and the enhanced nutrition interventions

Development Economics Example: The Pakistan Early Child Development Scale Up Trial (UNICEF)

- Children are randomly assigned to each group and the baseline data are collected by a different team to minimize the bias in assessment.
- Results indicate that at 24 months of age, all three of the intervention groups had significantly greater cognitive, language and motor scores compared to the control group
- RCT suggests that early child 'stimulation and care for development' intervention, delivered by Lady Health Workers to families with infants and young children aged 0-24 months living in rural Sindh in Pakistan, has beneficial outcomes on child development (cognitive, language, motor and social emotional development) and child growth.
- External validity: can we generalize these results to other settings?
- Internal validity: are the assumptions required for establishing cause-effect relationship satisfied? (for example, random assignment to control and treatment groups)

Quasi-Experiments (or natural experiments)

- Quasi-experiments lack random assignment but mimic the lab experiments: it is possible to measure the outcome variable in pretest and posttest
- Two similar but non-equivalent groups: program (treatment) vs comparison (control) group
- There are several potential outcomes in such an environment.
- A quasi-experiment will still use an experimental group(s) and a control group, but the researcher will not randomly assign participants to each group
- Differences in participants between groups may be minimized by the use of matched pairs.

Quasi-Experiments

- Quasi-experiments can be useful in evaluating social or economic programs
- Program evaluation is the field of study that concerns estimating the effect of a program, a policy, or some intervention or treatment (see Stock and Watson, 2010, Introduction to Econometrics, ch.13, p.469)
- Widely used quasi-experimental methods in econometrics include
- Matched-pair analysis,
- Propensity score matching
- Difference-in-differences
- Regression discontinuity designs