# Cross-Sectional Data

Pengpeng Yue

Version: Fall 2022

1. The Nature of Econometrics and Economic Data
   - What is Econometrics?
   - Steps in Empirical Economic Analysis
   - The Structure of Economic Data
   - Causality and the Notion of Ceteris Paribus in Econometric Analysis

2. The Simple Regression Model
   - Definition of the Simple Regression Model
   - Deriving the ordinary Least Squares Estimates
   - Properties of OLS on Any Sample of data

3. Heteroskedasticity
   - Consequences of Heteroskedasticity for OLS
   - Heteroskedasticity-Robust Inference after OLS Estimation
   - Testing for Heteroskedasticity

# Talk about your research interests

- Talk about your research interests
- Your recent works

# What is Econometrics?

## Example

*Imagine that you are hired by your state government to evaluate the effectiveness of a publicly funded job training program. Suppose this program teaches workers various ways to use computers in the manufacturing process. The twenty-week program offers courses during nonworking hours. Any hourly manufacturing worker may participate, and enrollment in all or part of the program is voluntary. You are to determine what, if any, effect the training program has on each worker's subsequent hourly wage.*

# What is Econometrics?

## Example

*Suppose you work for an investment bank. You are to study the returns on different investment strategies involving short-term U.S. treasury bills to decide whether they comply with implied economic theories.*

# What is Econometrics?

## Question

*What kind of data you would need to collect?*

# What is Econometrics?

**Note**

*How to use econometric methods to formally evaluate a job training program or to test a simple economic theory*

# What is Econometrics?

## Defintion

*Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy.*

# What is Econometrics?

**Note**

*The most common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product.*

# What is Econometrics?

## Note

*Econometric methods can be used in economic areas that have nothing to do with macroeconomic forecasting. For example, we will study the effects of political campaign expenditures on voting outcomes. We will consider the effect of school spending on student performance in the field of education.*

# What is Econometrics?

- Econometrics is the set of tools by which economists, and others in the social sciences, analyze data. We can use econometrics to estimate economic relationships, test economic theories, and evaluate government and business policy.
- For example, what if we are asked to study the effects of school spending on student performance? A simple correlation analysis might not be sufficient because causality can be difficult to infer.

# What is Econometrics?

- Econometrics is its own discipline (separate from statistics) because it focuses on problems inherent in analyzing data generated by individuals, firms, and other entities acting strategically, and interacting with one another.
- Data from controlled experiments - called experimental data - while common in the natural sciences, are harder to come by in the social sciences (although some exist).

# What is Econometrics?

- We usually have access to nonexperimental data, which are sometimes called observational or retrospective data. These are data sets collected in a passive manner, after we observe outcomes on individuals, firms, schools, and so on.
- We simply act as "observers" of what has happened and then try to learn from what we observe.

# Why Study Econometrics?

- Important to be able to apply economic theory to real world data.
- Theory may be ambiguous as to the effect of some policy change, and in any case theory rarely tells us how large the effect might be.
- Forecasting economic variables (inflation, interest rates, housing starts, and so on) is important, too.

# Steps in Empirical Economic Analysis

- In economics, theory and empirical analysis are both important. An empirical analysis uses data to test a theory, estimate an economic relationship, or determine the effects of a policy or intervention. Econometrics allows us to analyze data using formal statistical methods.

# How to structure an empirical economic analysis?

## Note

*It may seem obvious, but it is worth emphasizing that the first step in any empirical analysis is the careful formulation of the question of interest. The question might deal with testing a certain aspect of an economic theory, or it might pertain to testing the effects of a government policy. In principle, econometric methods can be used to answer a wide range of questions.*

# Econometric Methods

- In some cases, especially those that involve the testing of economic theories, a formal economic model is constructed. An economic model consists of mathematical equations that describe various relationships.

# Economic models

- Economists are well known for their building of models to describe a vast array of behaviors.
- For example, in intermediate microeconomics, individual consumption decisions, subject to a budget constraint, are described by mathematical models. The basic premise underlying these models is utility maximization.

# Utility Maximization

- The assumption that individuals make choices to maximize their well-being, subject to resource constraints, gives us a very powerful framework for creating tractable economic models and making clear predictions.
- In the context of consumption decisions, utility maximization leads to a set of demand equations.
- In a demand equation, the quantity demanded of each commodity depends on the price of the goods, the price of substitute and complementary goods, the consumer's income, and the individual's characteristics that affect taste.
- These equations can form the basis of an econometric analysis of consumer demand.

# Steps for a Successful Empirical Study

- Step 1. Be very precise in posing the question you hope to answer. For example, does attending lectures in college lead to better grades (on average)? If the severity of punishment for certain crimes increases, do crime rates fall on average?
- Step 2. Specify an economic model, or at least a conceptual model, to study the phenomenon of interest. Formal economic modeling (such as utility maximization) is often used, but one can get by with careful economic reasoning that is less formal.

# Example: Economic model of crime

## Example

*Gary Becker (a Nobel Prize winner), in 1968, wrote an influential article showing how criminal behavior can be modeled in a utility maximizing framework. The result is a supply function for time an individual spends in illegal activity (including, of course, not spending any time). Becker's analysis helps one decide the kind of factors one would ideally include in a study of factors that affect criminal behavior.*

# Example: Economic model of crime

- Certain crimes have clear economic rewards, but most criminal behaviors have costs. The opportunity costs of crime prevent the criminal from participating in other activities such as legal employment. In addition, there are costs associated with the possibility of being caught and then, if convicted, the costs associated with incarceration. From Becker's perspective, the decision to undertake illegal activity is one of resource allocation, with the benefits and costs of competing activities taken into account.

# Example: Economic model of crime

- Under general assumptions, we can derive an equation describing the amount of time spent in criminal activity as a function of various factors. We might represent such a function as

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \tag{1}$$

- $y$: hours spent in criminal activities, $x_1$:"wage" for an hour spent in criminal activity, $x_2$: hourly wage in 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, and $x_7$: age.
- This function depends on an underlying utility function, which is rarely known. This is the basis for an econometric analysis of individual criminal activity.

# Example: Job training and worker productivity

- To study the effects of job training on worker productivity - where productivity is measured by observed hourly wage - we can start with an equation such as

$$\text{wage} = f(\text{educ},\text{exper}, \text{training}) \tag{2}$$

- wage: hourly wage, educ: years of formal education, exper: years of workforce experience, and training: weeks spent in job training.

# Steps for a Successful Empirical Study

- Step 1: Carefully pose a question.
- Step 2: Specify an economic or conceptual model.
- Step 3: Turn the economic model into an econometric model.

**Note**

*After we specify an economic model, we need to turn it into what we call an econometric model.*

# Example: Economic model of crime

## Example

*The ambiguities inherent in the economic model of crime are resolved by specifying a particular econometric model:*

$$\begin{aligned} crime \;=\; & \beta_0 + \beta_1\, wage_{\,m} + \beta_2\, othinc \;+ \beta_3\, freqarr \;+ \beta_4\, freqconv \\ & + \beta_5\, avgsen \;+ \beta_6\, age \;+ \mu \end{aligned} \tag{3}$$

# the Econometric Model

- The choice of these variables is determined by the economic theory as well as data considerations. The term $\mu$ contains unobserved factors, such as the wage for criminal activity, moral character, family background, and errors in measuring things like criminal activity and the probability of arrest. We could add family background variables to the model, such as number of siblings, parents' education, and so on, but we can never eliminate $\mu$ entirely.
- In fact, dealing with this error term or disturbance term is perhaps the most important component of any econometric analysis.
- The constants $\beta_0$, $\beta_1$, ..., $\beta_6$ are the parameters of the econometric model, and they describe the directions and strengths of the relationship between crime and the factors used to determine crime in the model.

# the Econometric Mode

Our focus is on econometric models. Here is where we resolve certain difficulties and ambiguities concerning an economic model. For example, in a study of criminal behavior:

- How should we measure the probability of being caught committing a crime?
- What is the exact functional relationship among economic variables?
- How do we account for unobserved factors that make relationships among variables inexact?

# Steps for a Successful Empirical Study

- Step 1: Carefully pose a question.
- Step 2: Specify an economic or conceptual model.
- Step 3: Turn the economic model into an econometric model.
- Step 4: Collect data on the variables and use statistical methods to estimate the parameters, construct confidence intervals for the parameters, and test hypotheses.

# The Structure of Economic Data

- Cross-Sectional Data
- Time Series Data
- Pooled Cross Sections
- Panel or Longitudinal Data

# Cross-Sectional Data

- Data are collected on individuals, families, firms, schools, or some other units at a given point in time. (Or, at least, time does not play a crucial role. Interview dates for surveys may vary somewhat.)
- In this course, we will assume that a cross-sectional data set represents a random sample. That is, each unit in the population has the same chance of appearing the sample, and the draws are statistically independent of one another.

# Random Sampling

- For example, if we obtain information on wages, education, experience, and other characteristics by randomly drawing 500 people from the working population, then we have a random sample from the population of all working people. Random sampling is the sampling scheme covered in introductory statistics courses, and it simplifies the analysis of cross-sectional data.

# Sample Selection Problem

- Sometimes, random sampling is not appropriate as an assumption for analyzing cross- sectional data. For example, suppose we are interested in studying factors that influence the accumulation of family wealth. We could survey a random sample of families, but some families might refuse to report their wealth. If, for example, wealthier families are less likely to disclose their wealth, then the resulting sample on wealth is not a random sample from the population of all families. This is an illustration of a sample selection problem.

# Geographical Units

- Another violation of random sampling occurs when we sample from units that are large relative to the population, particularly geographical units. The potential problem in such cases is that the population is not large enough to reasonably assume the observations are independent draws.

# Random Sampling

- Random sampling (with replacement) generates observations that are independent and identically distributed.
- Intuitively, a random sample is representative of the population of interest, and gives us the best chance of learning about the population.

# Random Sampling

## Example

*To represent the undergraduate student population at a large university, we design a survey and collect information from 500 students. If we select these students randomly from all university students, we can treat the data as a random sample. If we only survey students entering, say, the engineering building, this is unlikely to be a random sample from the entire population.*

# Cross-sectional data

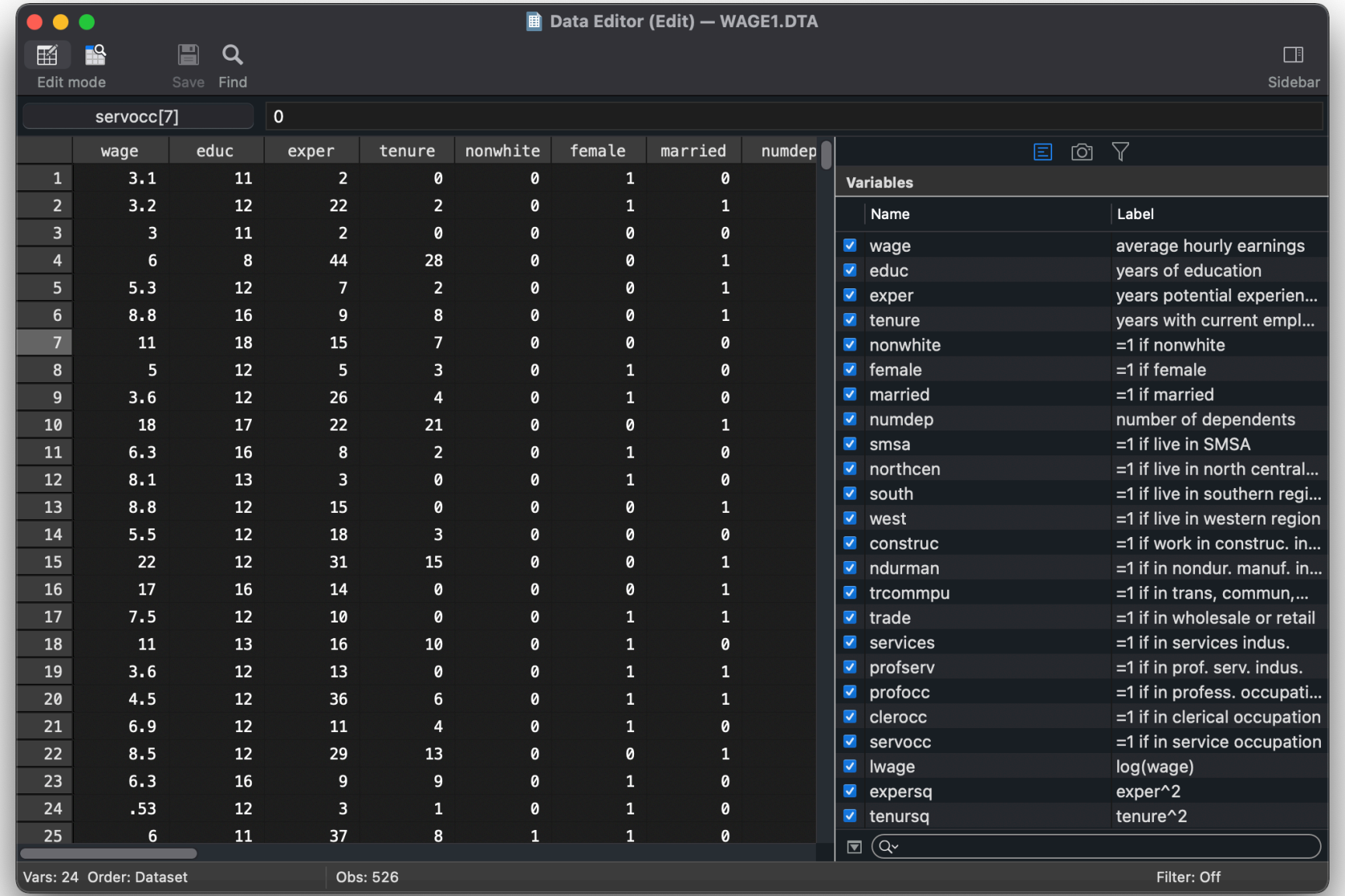
## Note

*Cross-sectional data are widely used in economics and other social sciences. In economics, the analysis of cross-sectional data is closely aligned with the applied microeconomics fields, such as labor economics, state and local public finance, industrial organization, urban economics, demography, and health economics. Data on individuals, households, firms, and cities at a given point in time are important for testing microeconomic hypotheses and evaluating economic policies.*

# Cross-sectional data: WAGE1.DTA

- WAGE1.DTA
- a cross-sectional data set on 526 working individuals for the year 1976
- The variables include wage (in dollars per hour), educ (years of education), exper (years of potential labor force experience), female (an indicator for gender), and married (marital status). These last two variables are binary (zero-one) in nature and serve to indicate qualitative features of the individual (the person is female or not; the person is married or not)

# Cross-sectional data: WAGE1.DTA




| | wage | educ | exper | tenure | nonwhite | female | married | numdep |
|---|---|---|---|---|---|---|---|---|
| 1 | 3.1 | 11 | 2 | 0 | 0 | 1 | 0 | |
| 2 | 3.2 | 12 | 22 | 2 | 0 | 1 | 1 | |
| 3 | 3 | 11 | 2 | 0 | 0 | 0 | 0 | |
| 4 | 6 | 8 | 44 | 28 | 0 | 0 | 1 | |
| 5 | 5.3 | 12 | 7 | 2 | 0 | 0 | 1 | |
| 6 | 8.8 | 16 | 9 | 8 | 0 | 0 | 1 | |
| 7 | 11 | 18 | 15 | 7 | 0 | 0 | 0 | |
| 8 | 5 | 12 | 5 | 3 | 0 | 1 | 0 | |
| 9 | 3.6 | 12 | 26 | 4 | 0 | 1 | 0 | |
| 10 | 18 | 17 | 22 | 21 | 0 | 0 | 1 | |
| 11 | 6.3 | 16 | 8 | 2 | 0 | 1 | 0 | |
| 12 | 8.1 | 13 | 3 | 0 | 0 | 1 | 0 | |
| 13 | 8.8 | 12 | 15 | 0 | 0 | 0 | 1 | |
| 14 | 5.5 | 12 | 18 | 3 | 0 | 0 | 0 | |
| 15 | 22 | 12 | 31 | 15 | 0 | 0 | 1 | |
| 16 | 17 | 16 | 14 | 0 | 0 | 0 | 1 | |
| 17 | 7.5 | 12 | 10 | 0 | 0 | 1 | 1 | |
| 18 | 11 | 13 | 16 | 10 | 0 | 1 | 0 | |
| 19 | 3.6 | 12 | 13 | 0 | 0 | 1 | 1 | |
| 20 | 4.5 | 12 | 36 | 6 | 0 | 1 | 1 | |
| 21 | 6.9 | 12 | 11 | 4 | 0 | 1 | 0 | |
| 22 | 8.5 | 12 | 29 | 13 | 0 | 0 | 1 | |
| 23 | 6.3 | 16 | 9 | 9 | 0 | 1 | 0 | |
| 24 | .53 | 12 | 3 | 1 | 0 | 1 | 0 | |
| 25 | 6 | 11 | 37 | 8 | 1 | 1 | 0 | |

# Cross-sectional data: WAGE1.DTA

| variable name | storage type | display format | value label | variable label |
|---|---|---|---|---|
| wage | float | %8.2g | | average hourly earnings |
| educ | byte | %8.0g | | years of education |
| exper | byte | %8.0g | | years potential experience |
| tenure | byte | %8.0g | | years with current employer |
| nonwhite | byte | %8.0g | | =1 if nonwhite |
| female | byte | %8.0g | | =1 if female |
| married | byte | %8.0g | | =1 if married |
| numdep | byte | %8.0g | | number of dependents |
| smsa | byte | %8.0g | | =1 if live in SMSA |
| northcen | byte | %8.0g | | =1 if live in north central U.S |
| south | byte | %8.0g | | =1 if live in southern region |
| west | byte | %8.0g | | =1 if live in western region |
| construc | byte | %8.0g | | =1 if work in construc. indus. |
| ndurman | byte | %8.0g | | =1 if in nondur. manuf. indus. |
| trcommpu | byte | %8.0g | | =1 if in trans, commun, pub ut |
| trade | byte | %8.0g | | =1 if in wholesale or retail |
| services | byte | %8.0g | | =1 if in services indus. |
| profserv | byte | %8.0g | | =1 if in prof. serv. indus. |
| profocc | byte | %8.0g | | =1 if in profess. occupation |
| clerocc | byte | %8.0g | | =1 if in clerical occupation |
| servocc | byte | %8.0g | | =1 if in service occupation |
| lwage | float | %9.0g | | log(wage) |
| expersq | int | %9.0g | | exper^2 |
| tenursq | int | %9.0g | | tenure^2 |

# Time Series Data

## Defintion

*A time series data set consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, and automobile sales figures.*

# Time Series Data

## Note

*Because past events can influence future events and lags in behavior are prevalent in the social sciences, time is an important dimension in a time series data set. Unlike the arrangement of cross-sectional data, the chronological ordering of observations in a time series conveys potentially important information.*

# Time Series Data: an example

- A key feature of time series data that makes them more difficult to analyze than cross-sectional data is that economic observations can rarely, if ever, be assumed to be independent across time. Most economic and other time series are related, often strongly related, to their recent histories.

# Time Series Data

## Example

*For example, knowing something about the gross domestic product from last quarter tells us quite a bit about the likely range of the GDP during this quarter, because GDP tends to remain fairly stable from one quarter to the next. Although most econometric procedures can be used with both cross-sectional and time series data, more needs to be done in specifying econometric models for time series data before standard econometric methods can be justified.*

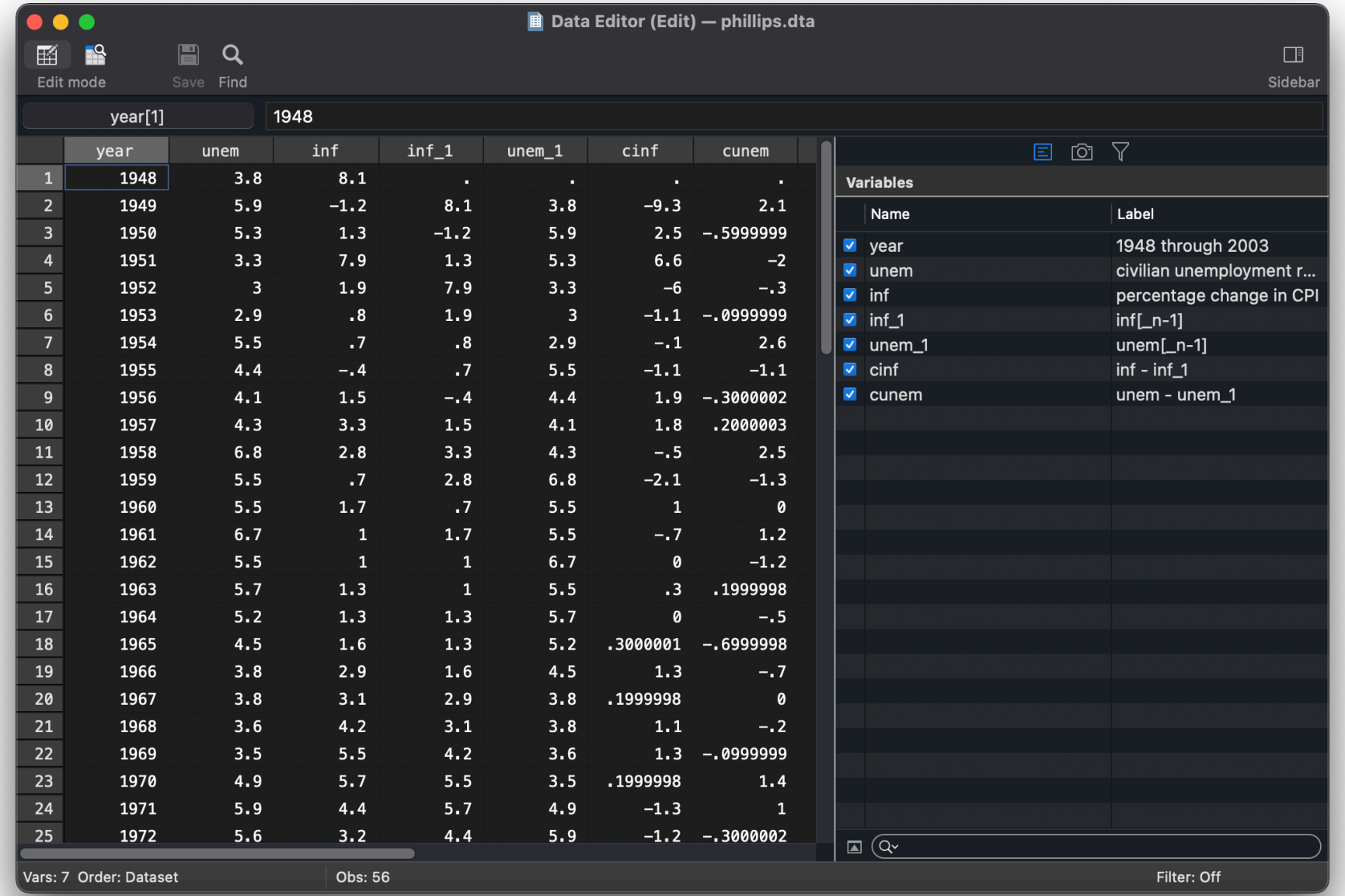
# Time Series Data: data frequency

- Another feature of time series data that can require special attention is the data frequency at which the data are collected. In economics, the most common frequencies are daily, weekly, monthly, quarterly, and annually. Stock prices are recorded at daily intervals (excluding Saturday and Sunday). The money supply in the U.S. economy is reported weekly. Many macroeconomic series are tabulated monthly, including inflation and unemployment rates.

# Time Series Data: seasonal pattern

- Many weekly, monthly, and quarterly economic time series display a strong seasonal pattern, which can be an important factor in a time series analysis. For example, monthly data on housing starts differ across the months simply due to changing weather conditions. We will learn how to deal with seasonal time series later.

# Cross-sectional data: phillips.DTA




| | year | unem | inf | inf_1 | unem_1 | cinf | cunem |
|---|---|---|---|---|---|---|---|
| 1 | 1948 | 3.8 | 8.1 | . | . | . | . |
| 2 | 1949 | 5.9 | -1.2 | 8.1 | 3.8 | -9.3 | 2.1 |
| 3 | 1950 | 5.3 | 1.3 | -1.2 | 5.9 | 2.5 | -.5999999 |
| 4 | 1951 | 3.3 | 7.9 | 1.3 | 5.3 | 6.6 | -2 |
| 5 | 1952 | 3 | 1.9 | 7.9 | 3.3 | -6 | -.3 |
| 6 | 1953 | 2.9 | .8 | 1.9 | 3 | -1.1 | -.0999999 |
| 7 | 1954 | 5.5 | .7 | .8 | 2.9 | -.1 | 2.6 |
| 8 | 1955 | 4.4 | -.4 | .7 | 5.5 | -1.1 | -1.1 |
| 9 | 1956 | 4.1 | 1.5 | -.4 | 4.4 | 1.9 | -.3000002 |
| 10 | 1957 | 4.3 | 3.3 | 1.5 | 4.1 | 1.8 | .2000003 |
| 11 | 1958 | 6.8 | 2.8 | 3.3 | 4.3 | -.5 | 2.5 |
| 12 | 1959 | 5.5 | .7 | 2.8 | 6.8 | -2.1 | -1.3 |
| 13 | 1960 | 5.5 | 1.7 | .7 | 5.5 | 1 | 0 |
| 14 | 1961 | 6.7 | 1 | 1.7 | 5.5 | -.7 | 1.2 |
| 15 | 1962 | 5.5 | 1 | 1 | 6.7 | 0 | -1.2 |
| 16 | 1963 | 5.7 | 1.3 | 1 | 5.5 | .3 | .1999998 |
| 17 | 1964 | 5.2 | 1.3 | 1.3 | 5.7 | 0 | -.5 |
| 18 | 1965 | 4.5 | 1.6 | 1.3 | 5.2 | .3000001 | -.6999998 |
| 19 | 1966 | 3.8 | 2.9 | 1.6 | 4.5 | 1.3 | -.7 |
| 20 | 1967 | 3.8 | 3.1 | 2.9 | 3.8 | .1999998 | 0 |
| 21 | 1968 | 3.6 | 4.2 | 3.1 | 3.8 | 1.1 | -.2 |
| 22 | 1969 | 3.5 | 5.5 | 4.2 | 3.6 | 1.3 | -.0999999 |
| 23 | 1970 | 4.9 | 5.7 | 5.5 | 3.5 | .1999998 | 1.4 |
| 24 | 1971 | 5.9 | 4.4 | 5.7 | 4.9 | -1.3 | 1 |
| 25 | 1972 | 5.6 | 3.2 | 4.4 | 5.9 | -1.2 | -.3000002 |

# Cross-sectional data: phillips.DTA

```
  obs:            56
 vars:             7                          29 Jul 2005 11:58
------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
year            int     %9.0g                 1948 through 2003
unem            float   %9.0g                 civilian unemployment rate, %
inf             float   %9.0g                 percentage change in CPI
inf_1           float   %9.0g                 inf[_n-1]
unem_1          float   %9.0g                 unem[_n-1]
cinf            float   %9.0g                 inf - inf_1
cunem           float   %9.0g                 unem - unem_1
------------------------------------------------------------------------------
Sorted by:
```

# Pooled Cross Sections

- Some data sets have both cross-sectional and time series features. For example, suppose that two cross-sectional household surveys are taken in the United States, one in 1985 and one in 1990. In 1985, a random sample of households is surveyed for variables such as income, savings, family size, and so on. In 1990, a new random sample of households is taken using the same survey questions. To increase our sample size, we can form a pooled cross section by combining the two years.

# Pooled Cross Sections

- Pooling cross sections from different years is often an effective way of analyzing the effects of a new government policy. The idea is to collect data from the years before and after a key policy change. As an example, consider the following data set on housing prices taken in 1993 and 1995, before and after a reduction in property taxes in 1994. Suppose we have data on 250 houses for 1993 and on 270 houses for 1995.

# Pooled Cross Sections

- A pooled cross section is analyzed much like a standard cross section, except that we often need to account for secular differences in the variables across the time. In fact, in addition to increasing the sample size, the point of a pooled cross-sectional analysis is often to see how a key relationship has changed over time.

# Panel Data

## Defintion

*A panel data (or longitudinal data) set consists of a time series for each cross-sectional member in the data set.*

# Panel Data

## Example

*As an example, suppose we have wage, education, and employment history for a set of individuals followed over a ten-year period. Or we might collect information, such as investment and financial data, about the same set of firms over a five-year time period. Panel data can also be collected on geographical units. For example, we can collect data for the same set of counties in the United States on immigration flows, tax rates, wage rates, government expenditures, and so on, for the years 1980, 1985, and 1990.*

# Panel Data

## Note

*The key feature of panel data that distinguishes them from a pooled cross section is that the same cross-sectional units (individuals, firms, or counties in the preceding examples) are followed over a given time period.*

# Panel Data

- Superficially, a panel data set has a structure similar to a pooled cross section. The key difference is that with a panel data set the same units (people, houses, schools, and so on) are followed over time.

# Panel data analysis is a more advanced topic

- Because panel data require replication of the same units over time, panel data sets, especially those on individuals, households, and firms, are more difficult to obtain than pooled cross sections.
- Not surprisingly, observing the same units over time leads to several advantages over cross-sectional data or even pooled cross-sectional data.

# Advantages of Panel data

- The benefit that we will focus on in this text is that having multiple observations on the same units allows us to control for certain unobserved characteristics of individuals, firms, and so on. The use of more than one observation can facilitate causal inference in situations where inferring causality would be very difficult if only a single cross section were available.
- A second advantage of panel data is that they often allow us to study the importance of lags in behavior or the result of decision making. This information can be significant because many economic policies can be expected to have an impact only after some time has passed.

# Causality

- The concept of causality is key in econometrics. How can we know that more spending causes better student performance (on average)? How can we know another year of education causes an increase in wages (on average)? Finding correlations in data might be suggestive but is rarely conclusive.

# the Notion of Ceteris Paribus

- Crucial to establishing causality is the notion of ceteris paribus: "all (relevant) factors equal." If we succeed –via statistical methods –in "holding fixed" other relevant factors, then sometimes we establish that changes in one variable (say, education) in fact "cause" changes in another variable (wage).

# EXAMPLE: Does a new fertilizer increase soybean yield?

## Example

*One way to determine the causal effect of fertilizer is to conduct an experiment. Choose several one-acre plots of land, apply different amounts of fertilizer to the plots, and subsequently record soybean yields. This experiment does not hold other relevant factors fixed because, for example, the land quality differs across plots.*

# EXAMPLE: Does a new fertilizer increase soybean yield?

**Note**

*It is impossible to truly hold all other factors fixed, but here is an important point: If the fertilizer amounts are assigned independently of other land factors that affect yield, then we will be able to use simple regression analysis to get a "good" estimate of the causal effect.*

# EXAMPLE: What is the value of another year of education on one's earnings?

## Example

*We can imagine the type of experiment we would have to run to obtain experimental data. At birth, each child is randomly given a highest grade that he/she must complete –no more, no less. This is like randomly assigning fertilizer amounts (years of schooling) to plots of land (individuals). Then, we eventually record, hourly or monthly or annual earnings.*

# EXAMPLE: What is the value of another year of education on one's earnings?

- The experiment is not feasible, and would be morally repugnant, anyway.
- For problems such as measuring the value of education, we must usually rely on observational data. We can, for very large random samples of people, collect information on education and earnings.
- The problem for inferring causality from, say, a simple correlation analysis is that individuals and their parents largely determine the amount of schooling. Probably on average people who are smarter or more capable choose to become better educated. But more capable people would earn more, on average, than less capable individuals.
- Seeing a positive correlation between earnings and schooling need not imply that it is due to schooling. Other confounding factors (such as intelligence and past experience) could explain most of the difference in earnings.

# EXAMPLE: What is the value of another year of education on one's earnings?

- The problem of individuals influencing their education levels is an example of self-selection. As another example, suppose we want to study the effects of attending college lectures on performance in a course. If the better students, on average, also attend lectures more frequently, a simple correlation analysis can be misleading. The students self-select into how much they attend lecture.
- Self-selection is often a serious concern in the social sciences.

# EXAMPLE: What is the value of another year of education on one's earnings?

- We can draw a parallel with the fertilizer example. What if an assistant decided (perhaps unknown to us) to put more fertilizer on the better plots of land? Then we might see a positive correlation between yield and fertilizer that might have nothing to do with the fertilizer but rather land quality.

# EXAMPLE: What is the effect of the minimum wage on employment?

## Example

*Studies of the minimum wage often use time series data (or even panel data). Suppose we want to study the effects using monthly data from California. We would have variation in the real minimum wage even in the stretches that the nominal minimum wage does not change.*

# EXAMPLE: What is the effect of the minimum wage on employment?

- The state legislature adjusts the minimum wage. Sometimes it is forced to by federal minimum wage laws. What if the state is more likely to increase the minimum during good economic times, that is, when the employment rate is high? Then, in the data, we might actually see a positive correlation between employment and the minimum wage.
- We will learn to use multiple regression analysis to convincingly (we hope) estimate causal effects.

# Definition of the Simple Regression Model

- We begin with cross-sectional analysis and will (eventually) assume we can collect a random sample from the population of interest.
- We begin with the following premise, once we have a population in mind. There are two variables, x and y, and we would like to "study how y varies with changes in x."
- We have seen examples: x is amount of fertilizer, and y is soybean yield; x is years of schooling, y is hourly wage.

# Definition of the Simple Regression Model

We must confront three issues:

- 1. How do we allow factors other than x to affect y? There is never an exact relationship between two variables (in interesting cases).
- 2. What is the functional relationship between y and x?
- 3. How can we be sure we a capturing a ceteris paribus relationship between y and x (as is so often the goal)?

# Definition of the Simple Regression Model

- Consider the following equation relating y to x:

$$y = \beta_0 + \beta_1 x + \mu \tag{4}$$

  which is assumed to hold in the population of interest.
- This equation defines the simple linear regression model (or two-variable regression model).
- The term "regression" has historical roots in the "regression-to-the-mean" phenomenon.

# Definition of the Simple Regression Model

- y and x are not treated symmetrically. We want to explain y in terms of x. From a causality standpoint, it makes no sense to "explain" past educational attainment in terms of future labor earnings.
- As another example, we want to explain student performance (y) in terms of class size (x), not the other way around.

# Definition of the Simple Regression Model

- The terms "explained" and "explanatory" (y and x) are probably best, as they are the most descriptive and widely applicable. But "dependent" and "independent" are used often. ("Independent" here should not be confused with the notion of statistical independence.)

# Definition of the Simple Regression Model

- We mentioned the error term or disturbance, $\mu$, before. The equation

$$y = \beta_0 + \beta_1 x + \mu \tag{5}$$

  explicitly allows for other factors, contained in u, to affect y.
- This equation also addresses the functional form issue (in a simple way). Namely, y is assumed to be linearly related to x. We call $\beta_0$ the intercept parameter and $\beta_1$ the slope parameter. These describe a population, and our ultimate goal is to estimate them.

# Definition of the Simple Regression Model

- The equation also addresses the ceteris paribus issue. In

$$y = \beta_0 + \beta_1 x + \mu \tag{6}$$

  all other factors that affect y are in $\mu$. We want to know how y changes when x changes, holding $\mu$ fixed.
- Let $\Delta$ denote "change." Then holding u fixed means $\Delta\mu = 0$. So

$$\begin{aligned} \Delta y &= \beta_1 \Delta x + \Delta u \\ &= \beta_1 \Delta x \text{ when } \Delta u = 0 \end{aligned}$$

  This equation effectively defines $\beta_1$ as a slope, with the only difference being the restriction $\Delta\mu = 0$.

# Yield and Fertilizer

## Example

*A model to explain crop yield to fertilizer use is*

$$yield = \beta_0 + \beta_1 \, fertilizer + u \tag{7}$$

*where u contains land quality, rainfall on a plot of land, and so on. The slope parameter, $\beta_1$, is of primary interest: it tells us how yield changes when the amount of fertilizer changes, holding all else fixed.*

## Note

*The linear function is probably not realistic here. The effect of fertilizer is likely to diminish at large amounts of fertilizer.*

# Definition of the Simple Regression Model

## Example

$$Wage and Education\ wage = \beta_0 + \beta_1 educ + u \tag{8}$$

*where u contains somewhat nebulous factors ("ability") but also past workforce experience and tenure on the current job.*

$$\Delta wage = \beta_1 \Delta educ \text{ when } \Delta u = 0 \tag{9}$$

## Question

*Is each year of education really worth the same dollar amount nomatter how much education one starts with?*

# Definition of the Simple Regression Model

- We said we must confront three issues:
  1. How do we allow factors other than x to affect y?
  2. What is the functional relationship between y and x?
  3. How can we be sure we a capturing a ceteris paribus relationship between y and x?
- We have argued that the simple regression model

$$y = \beta_0 + \beta_1 x + \mu \tag{10}$$

addresses each of them.

# Definition of the Simple Regression Model

- This seems too easy! How can we hope to generally estimate the ceteris paribus effect of y on x when we have assumed all other factors affecting y are unobserved and lumped into u?
- The key is that the simple linear regression (SLR) model is a population model. When it comes to estimating $\beta_1$ (and $\beta_0$) using a random sample of data, we must restrict how u and x are related to each other.

# Definition of the Simple Regression Model

- But x and u are properly viewed as having distributions in the population. For example, if $x = educ$ then, in principle, we could figure out its distribution in the population of adults over, say, 30 years old. Suppose u is cognitive ability. Assuming we can measure what that means, it also has a distribution in the population.
- What we must do is restrict the way in when u and x relate to each other in the population.

- First, we make a simplifying assumption that is without loss of generality: the average, or expected, value of u is zero in the population:

$$E(u) = 0$$

  where $E(\cdot)$is the expected value (or averaging) operator.
- Normalizing "land quality," or "ability," to be zero in the population should be harmless. It is.

- The presence of $\beta_0$ in

$$y = \beta_0 + \beta_1 x + \mu$$

allows us to assume $E(u) = 0$. If the average of u is different from zero, we just adjust the intercept, leaving the slope the same.If$\alpha_0 = E(u)$ then we can write

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0)$$

where the new error, $u - \alpha_0$, has a zero mean.

- The new intercept is $\beta_0 + \alpha_0$. The important point is that the slope, $\beta_1$, has not changed.

## Question

*How do we need to restrict the dependence between u and x?*

- We could assume u and x uncorrelated in the population:

$$\text{Corr}(x, u) = 0$$

- Zero correlation actually works for many purposes, but it implies only that u and x are not linearly related. Ruling out only linear dependence can cause problems with interpretation and makes statistical analysis more difficult.

- An assumption that meshes well with our introductory treatment involves the mean of the error term for each slice of the population determined by values of x:

$$E(u \mid x) = E(u), \text{ all values } x$$

  where $E(u|x)$ means "the expected value of u given x."
- We say u is mean independent of x.

- Suppose u is "ability" and x is years of education. We need, for example,

$$E(\text{ ability } \mid x = 8) = E(\text{ ability } \mid x = 12) = E(\text{ ability } \mid x = 16) \tag{11}$$

- so that the average ability is the same in the different portions of the population with an $8^{th}$ grade education, a $12^{th}$ grade education, and a four-year college eduction.
- Because people choose education levels partly based on ability, this assumption is almost certainly false.

- Suppose u is "land quality" and x is fertilizer amount. Then $E(u|x) = E(u)$ if fertilizer amounts are chosen independently of quality. This assumption is reasonably but assumes fertilizer amounts are assigned at random.
- Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u|x) = 0$ (a normalization) gives

$$E(u \mid x) = 0, \text{ all values } x$$

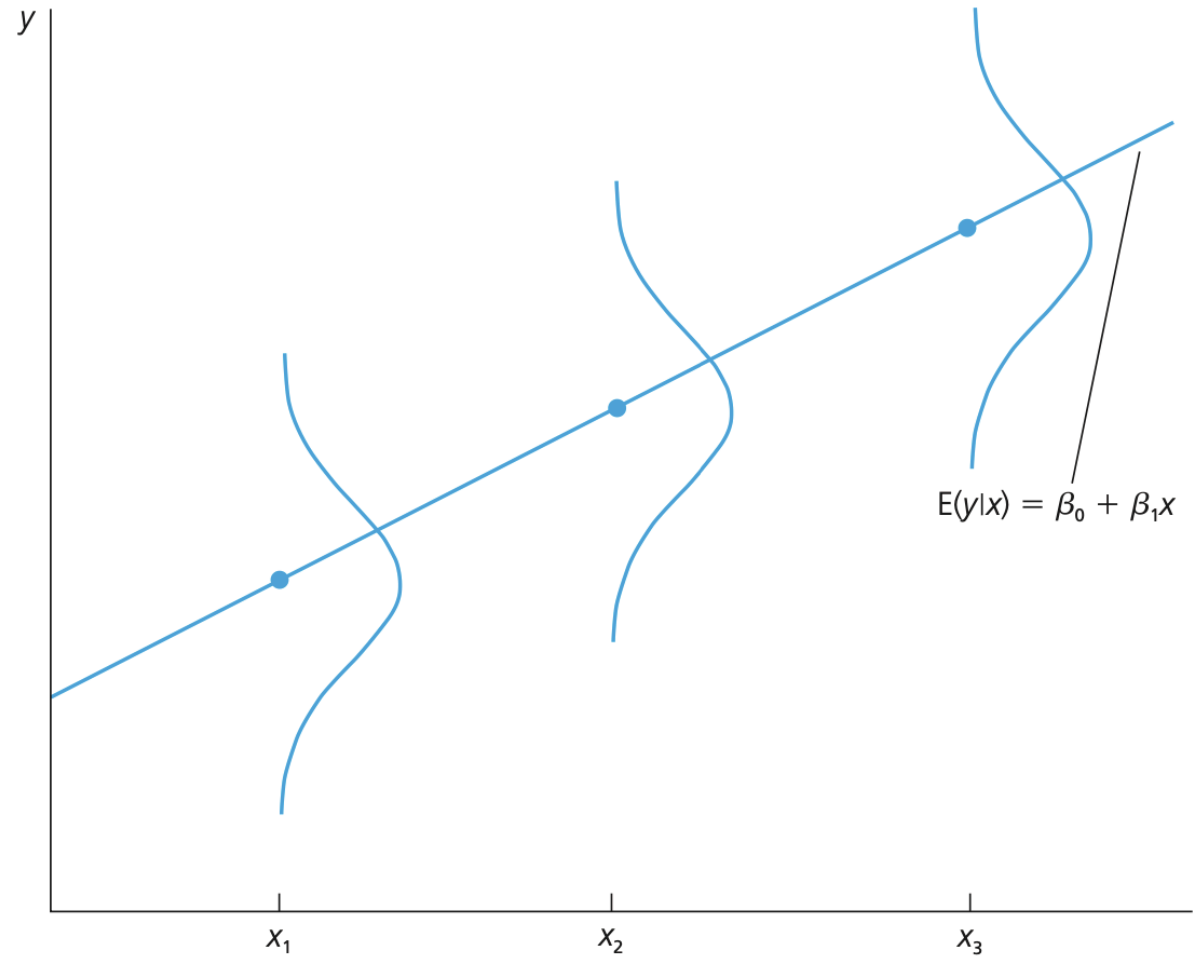
- Called the zero conditional mean assumption.

- Because the expected value is a linear operator, $E(u|x) = 0$ implies

$$E(y \mid x) = \beta_0 + \beta_1 x + E(u \mid x) = \beta_0 + \beta_1 x \tag{12}$$

- which shows the population regression function (PRF) is a linear function of x.
- A different approach to simple regression ignores the causality issue and just starts with a linear model for $E(y|x)$ as a descriptive device.

- Equation (12) shows that the population regression function (PRF), $E(y|x)$, is a linear function of x. The linearity means that a one-unit increase in x changes the expected value of y by the amount $\beta_1$.



# EXAMPLE: College versus High School GPA.

- Suppose for the population of students attending a university, we (somehow) know

$$E(\text{col } GPA \mid hsGPA) = 1.5 + 0.5 hsGPA$$

- y is college GPA, and x is high school GPA
- If $hsGPA = 3.6$ then the average of colGP A among students with this particular high school GPA is

$$1.5 + 0.5(3.6) = 3.3$$

## Note

*Regression analysis is essentially about explaining effects of explanatory variables on average outcomes of y.*

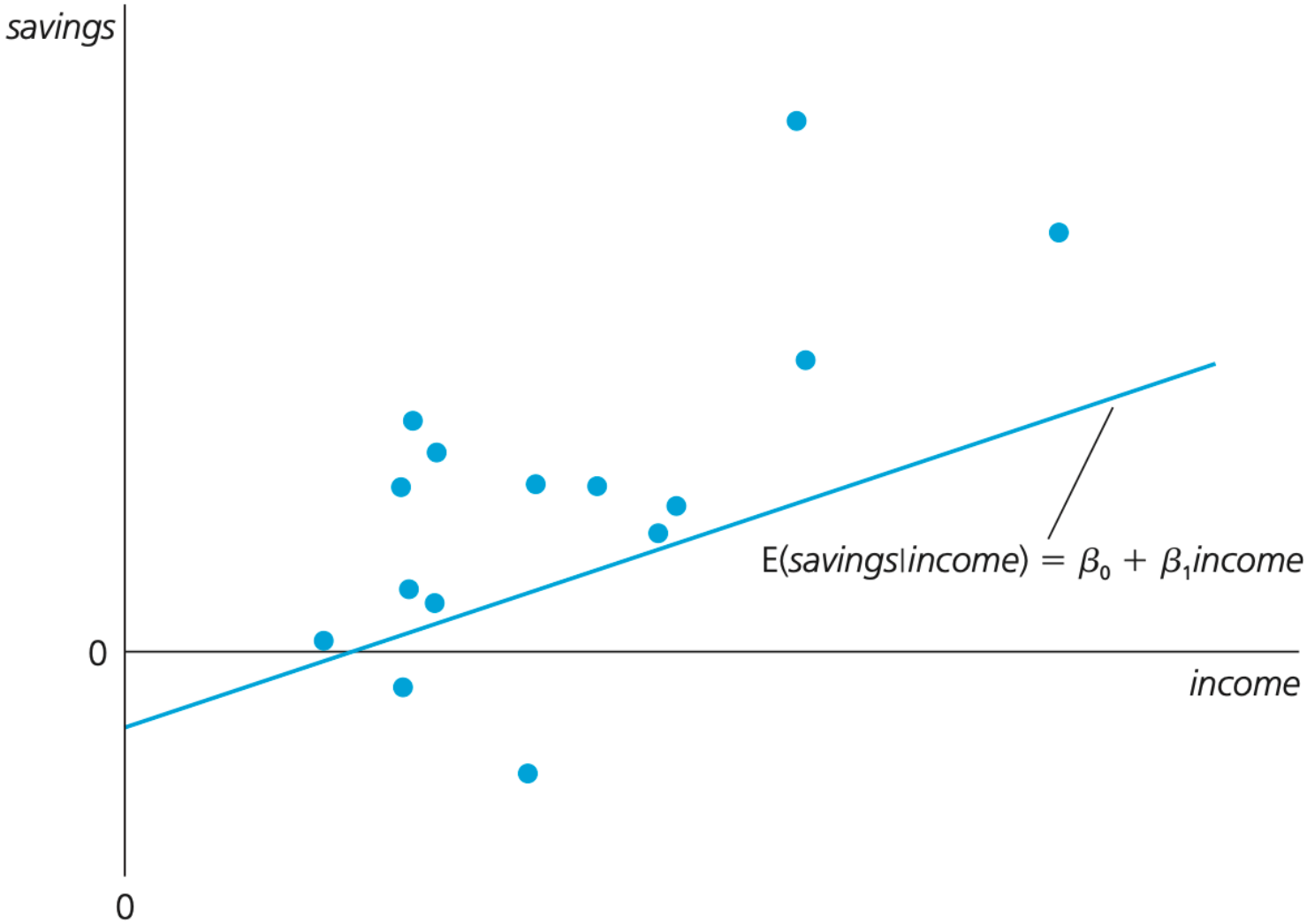
# What we'll learn in the next lecture.

- Multiple Regression Analysis: Estimation & Inference & OLS Asymptotics
- Heteroskedasticity

# Deriving the ordinary Least Squares Estimates

- Given data on x and y, how can we estimate the population parameters, $\beta_0$ and $\beta_1$?
- Let $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ be a sample of size n (the number of observations) from the population. Think of this as a random sample.
- The next graph shows $n = 15$ families and the population regression of saving on income.

Scatterplot of savings and income for 15 families, and the population regression $E(savings|income) = \beta 0 + \beta_1 income$.



- Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

  where the i subscript indicates a particular observation.
- We observe $y_i$ and $x_i$, but not $u_i$. (However, we know $u_i$ is there.)

- We use the two restrictions

$$E(u) = 0$$

$$Cov(x, u) = 0$$

  to obtain estimating equations for $\beta_0$ and $\beta_1$.
- Remember, the first condition essentially defines the intercept.
- The second condition, stated in terms of the covariance, means that x and u are uncorrelated.
- Both conditions are implied by the zero conditional mean assumption

$$E(u|x) = 0$$

# Deriving the ordinary Least Squares Estimates

- With $E(u) = 0$, $Cov(x, u) = 0$ is the same as $E(xu) = 0$ because $Cov(x, u) = E(xu) - E(x)E(u)$.
- Next we plug in for u into the two equations:

$$E(y - \beta_0 - \beta_1 x) = 0$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0$$

- These are the two conditions in the population that determine $\beta_0$ and $\beta_1$. So we use their sample analogs, which is a method of moments approach to estimation.

- In other words, we use

$$n^{-1}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0$$

$$n^{-1}\sum_{i=1}^{n}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0$$

to determine $\hat{\beta}_0$ and $\hat{\beta}_1$ , the estimates from the data.

# Deriving the ordinary Least Squares Estimates

- To solve the equations, pass the summation operator through the first equation:

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right) &= n^{-1}\sum_{i=1}^{n} y_i - n^{-1}\sum_{i=1}^{n}\hat{\beta}_0 - n^{-1}\sum_{i=1}^{n}\hat{\beta}_1 x_i \\
&= n^{-1}\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1\left(n^{-1}\sum_{i=1}^{n} x_i\right) \\
&= \bar{y}-\hat{\beta}_0-\hat{\beta}_1\bar{x}
\end{aligned}
\tag{13}
$$

- We use the standard notation $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ for the average of the n numbers $\{y_i : i = 1, 2, \ldots, n\}$. For emphasis, we call $\bar{y}$ a sample
- We have shown that the first equation,

$$n^{-1} \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

implies

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

- Rewrite this equation so that the intercept is terms of the slope (and the sample averages on y and x):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and plug this into the second equation (and drop the division by n):

$$\sum_{i=1}^{n} x_i \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

so

$$\sum_{i=1}^{n} x_i \left[ y_i - \left( \bar{y} - \hat{\beta}_1 \bar{x} \right) - \hat{\beta}_1 x_i \right] = 0$$

# Deriving the ordinary Least Squares Estimates

- Simple algebra gives

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \left[ \sum_{i=1}^{n} x_i (x_i - \bar{x}) \right]$$

and so we have one linear equation in the one unknown $\hat{\beta}_1$ .

- Showing the solution for $\hat{\beta}_1$ uses three useful facts about the summation operator:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$\sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

# Deriving the ordinary Least Squares Estimates

- So, we can write the equation to solve is

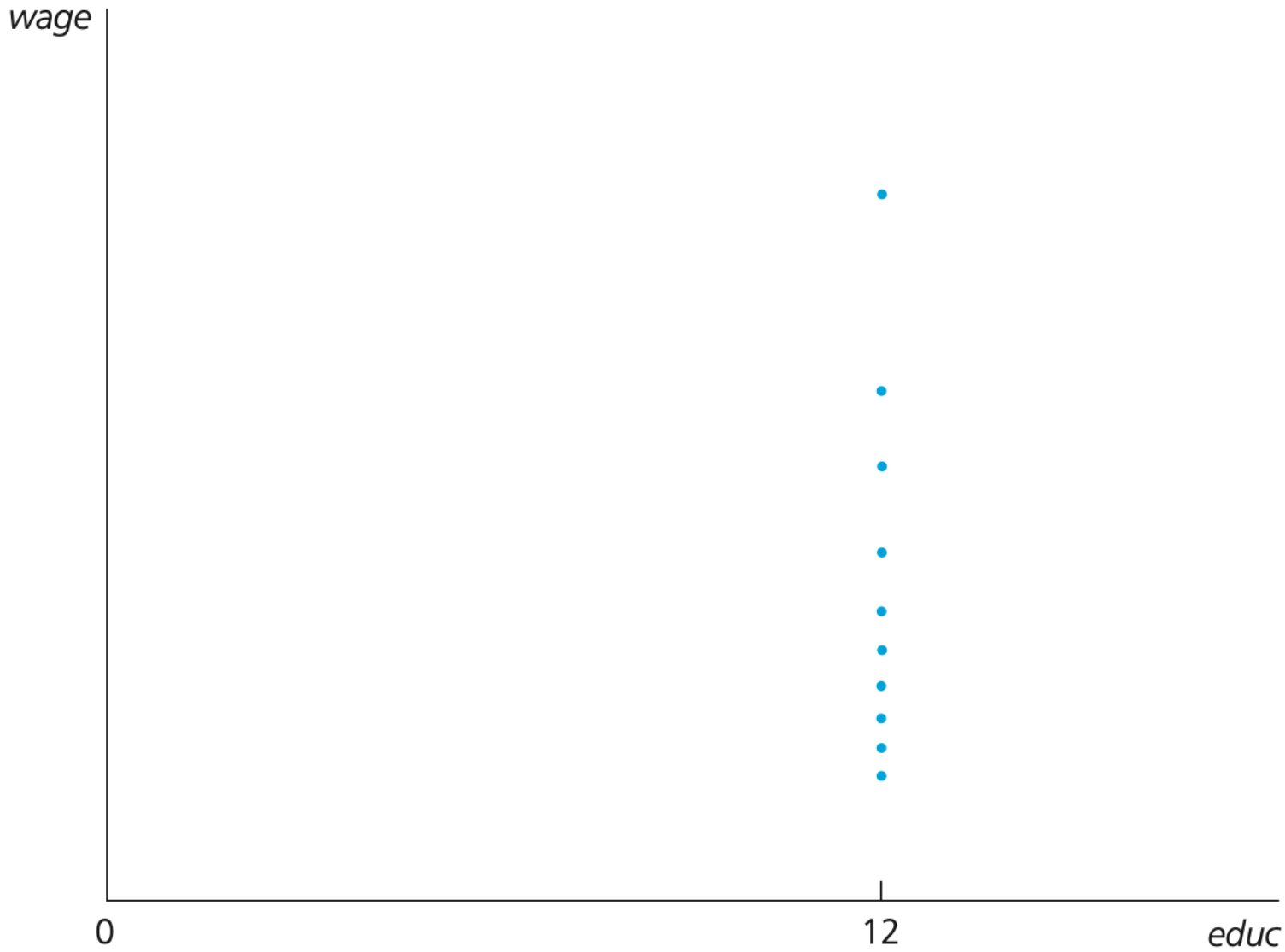
$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)=\hat{\beta}_1\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]$$

- If $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2>0$, we can write

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{\text{Sample Covariance }\left(x_i,y_i\right)}{\text{Sample Variance }\left(x_i\right)}$$

- The previous formula for $\hat{\beta}_1$ is important. It shows us how to take the data we have and compute the slope estimate. For reasons we will see, $\hat{\beta}_1$ is called the ordinary least squares (OLS) slope estimate. We often refer to it as the slope estimate.
- It can be computed whenever the sample variance of the $x_i$ is not zero, which only rules out the case where each $x_i$ is the same value.
- The following graph shows we have no way to determine the slope in a relationship between wage and educ if we observe a sample where everyone has 12 years of schooling.

A scatterplot of wage against education when $educ_i = 12$ for all i.



- Situations like those in the previous graph are very rare. Except with very small sample sizes we will be able to compute a slope estimate.
- Once we have $\hat{\beta}_1$,we compute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.This is the OLS intercept estimate.
- These days, one lets a computer do the calculations, which can be tedious even if n is small.

# Deriving the ordinary Least Squares Estimates

- Where does the name "ordinary least squares" come from?
- For any candidates $\hat{\beta}_0$ and $\hat{\beta}_1$ , define a fitted value for each data point i as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  We have n of these. It is the value we predict for $y_i$ given that x has taken on the value $x_i$.
- The mistake we make is the residual:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

  and we have n residuals.

- Suppose we measure the size of the mistake, for each i, by squaring the residual: $\hat{u}_i^2$ . Then we add them all up:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- This quantity is called the sum of squared residuals.
- If we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of squared residuals it can be shown (using calculus or other arguments) that the solutions are the slope and intercept estimates we obtained before.

- Once we have the numbers $\hat{\beta}_0$ and $\hat{\beta}_1$ for a given data set, we write the OLS regression line as a function of x:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The OLS regression line allows us to predict y for any (sensible) value of x. It is also called the sample regression function.
- The intercept, $\hat{\beta}_0$ , is the predicted y when $x = 0$. (The prediction is usually meaningless if $x = 0$ is not possible.)

- The slope, $\hat{\beta}_1$ , allows us to predict changes in y for any (reasonable) change in x:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x$$

- If $\Delta x = 1$, so that x increases by one unit, then $\Delta\hat{y} = \hat{\beta}_1$

# Deriving the ordinary Least Squares Estimates

## Example

*Effects of Education on Hourly Wage (WAGE2.DTA). Data are from 1991 on men only. wage is reported in dollars per hour, educ is highest grade completed.*

- The estimated equation is

$$\widehat{wage} = 146.95 + 60.21\text{educ}$$
$$n = 759$$

- Each additional year of schooling is estimated to be worth \$60.21.

- In the Stata output, $\hat{\beta}_0 = 146.95$ is the Coef. labeled "_cons." $\hat{\beta}_1 = 60.21$ is the Coef. labeled "educ."
- We will learn about the other numbers as we go.
- General form of the Stata command:

```
reg  y  x
```

The order of y and x is critical!

```
  obs:           935
 vars:            17                          14 Apr 1999 13:41
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
wage            int     %9.0g                 monthly earnings
hours           byte    %9.0g                 average weekly hours
IQ              int     %9.0g                 IQ score
KWW             byte    %9.0g                 knowledge of world work score
educ            byte    %9.0g                 years of education
exper           byte    %9.0g                 years of work experience
tenure          byte    %9.0g                 years with current employer
age             byte    %9.0g                 age in years
married         byte    %9.0g                 =1 if married
black           byte    %9.0g                 =1 if black
south           byte    %9.0g                 =1 if live in south
urban           byte    %9.0g                 =1 if live in SMSA
sibs            byte    %9.0g                 number of siblings
brthord         byte    %9.0g                 birth order
meduc           byte    %9.0g                 mother's education
feduc           byte    %9.0g                 father's education
lwage           float   %9.0g                 natural log of wage
-------------------------------------------------------------------------------
Sorted by:
```

```
. sum wage educ IQ
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| wage | 935 | 957.9455 | 404.3608 | 115 | 3078 |
| educ | 935 | 13.46845 | 2.196654 | 9 | 18 |
| IQ | 935 | 101.2824 | 15.05264 | 50 | 145 |

```
. tab educ
```

| years of education | Freq. | Percent | Cum. |
|---|---|---|---|
| 9 | 10 | 1.07 | 1.07 |
| 10 | 35 | 3.74 | 4.81 |
| 11 | 43 | 4.60 | 9.41 |
| 12 | 393 | 42.03 | 51.44 |
| 13 | 85 | 9.09 | 60.53 |
| 14 | 77 | 8.24 | 68.77 |
| 15 | 45 | 4.81 | 73.58 |
| 16 | 150 | 16.04 | 89.63 |
| 17 | 40 | 4.28 | 93.90 |
| 18 | 57 | 6.10 | 100.00 |
| Total | 935 | 100.00 | |

```
. reg wage educ
```

| Source | SS | df | MS | Number of obs | = | 935 |
|---|---|---|---|---|---|---|
| | | | | F(1, 933) | = | 111.79 |
| Model | 16340644.5 | 1 | 16340644.5 | Prob > F | = | 0.0000 |
| Residual | 136375524 | 933 | 146168.836 | R-squared | = | 0.1070 |
| | | | | Adj R-squared | = | 0.1060 |
| Total | 152716168 | 934 | 163507.675 | Root MSE | = | 382.32 |

| wage | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| educ | 60.21428 | 5.694982 | 10.57 | 0.000 | 49.03783 | 71.39074 |
| _cons | 146.9524 | 77.71496 | 1.89 | 0.059 | -5.56393 | 299.4688 |

- Reminder: When we write the population regression line,

$$\text{wage} = \beta_0 + \beta_1 educ + u$$

we do not know $\beta_0$ and $\beta_1$. Rather, $\hat{\beta}_0 = 146.95$ and $\hat{\beta}_1 = 60.21$ are our estimates from this particular sample of 759 men. These estimates may or may not be close to the population values. If we obtain another sample of 759 men the estimates would almost certainly change.

- The function

$$\widehat{\text{wage}} = 146.95 + 60.21 \text{ educ}$$

  is the OLS (or sample) regression line.
- Plugging in $educ = 0$ gives the silly prediction $\widehat{\text{wage}} = 146.95$. Extrapolating outside the range of the data can produce strange predictions. There are no men in the sample with $educ < 8$.

- When $educ = 9$,

$$\widehat{\text{wage}} = 146.95 + 60.21(9) = 688.84$$

- The predicted hourly wage at eight years of education is \$688.84, which we can think of as our estimate of the average wage in the population when $educ = 9$. But no one in the sample earns exactly \$688.84: some earn more, some earn less. One worker earns \$705, which is close.

```
. list wage if educ == 9

     +------+
     | wage |
     |------|
112. |  895 |
349. |  800 |
457. |  477 |
469. |  995 |
573. |  705 |
     |------|
679. |  500 |
691. |  975 |
781. | 1200 |
814. |  625 |
856. |  571 |
     +------+
```

# Properties of OLS on Any Sample of data

- One we have

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

we get the OLS fitted values by plugging the $x_i$ into the equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \ldots, n$$

- The OLS residuals are

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1, 2, \ldots, n$$

```
. list wage if educ == 8

. reg wage educ

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(1, 933)       =    111.79
       Model |  16340644.5         1  16340644.5   Prob > F        =    0.0000
    Residual |   136375524       933  146168.836   R-squared       =    0.1070
-------------+----------------------------------   Adj R-squared   =    0.1060
       Total |   152716168       934  163507.675   Root MSE        =    382.32

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   60.21428   5.694982    10.57   0.000     49.03783    71.39074
       _cons |   146.9524   77.71496     1.89   0.059    -5.56393    299.4688
------------------------------------------------------------------------------

. predict wagehat
(option xb assumed; fitted values)

. predict uhat, resid
```

```
. list wage educ wagehat uhat in 1/15
```

| | wage | educ | wagehat | uhat |
|---|---|---|---|---|
| 1. | 769 | 12 | 869.5239 | -100.5238 |
| 2. | 808 | 18 | 1230.81 | -422.8095 |
| 3. | 825 | 14 | 989.9524 | -164.9524 |
| 4. | 650 | 12 | 869.5239 | -219.5238 |
| 5. | 562 | 11 | 809.3096 | -247.3096 |
| 6. | 1400 | 16 | 1110.381 | 289.619 |
| 7. | 600 | 10 | 749.0953 | -149.0953 |
| 8. | 1081 | 18 | 1230.81 | -149.8095 |
| 9. | 1154 | 15 | 1050.167 | 103.8333 |
| 10. | 1000 | 12 | 869.5239 | 130.4762 |
| 11. | 930 | 18 | 1230.81 | -300.8095 |
| 12. | 921 | 14 | 989.9524 | -68.95241 |
| 13. | 900 | 15 | 1050.167 | -150.1667 |
| 14. | 1318 | 16 | 1110.381 | 207.619 |
| 15. | 1792 | 16 | 1110.381 | 681.619 |

# Properties of OLS on Any Sample of data

- Some residuals are positive, others are negative. None in the first 15 is especially close to zero. Years of schooling, by itself, need not be a very good predictor of wage. We need to formalize how good it is.

# Algebraic Properties of OLS Statistics

(1) The OLS residuals always add up to zero:

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

- Because $y_i = \hat{y}_i + \hat{u}_i$ by definition,

$$n^{-1} \sum_{i=1}^{n} y_i = n^{-1} \sum_{i=1}^{n} \hat{y}_i + n^{-1} \sum_{i=1}^{n} \hat{u}_i$$

and so $\bar{y} = \bar{\hat{y}}$. In other words, the sample average of the actual $y_i$ is the same as the sample average of the fitted values.

# Algebraic Properties of OLS Statistics

(2) The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero:

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

- Because the $\hat{y}_i$ are linear functions of the $x_i$, the fitted values and residuals are uncorrelated, too:

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$$

# Algebraic Properties of OLS Statistics

(3) The point $\bar{x}$, $\bar{y}$ is always on the OLS regression line.

- That is, if we plug in the average for x, we predict the sample average for y:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

# Goodness-of-Fit

- For each observation, write

$$y_i = \hat{y}_i + \hat{u}_i$$

- Define the total sum of squares (SST), explained sum of squares (SSE) –Stata calls this the "model sum of squares"–and residual sum of squares (or sum of squared residuals) as

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

- By writing

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST = \sum_{i=1}^{n} [(y_i - \hat{y}_i) - (\hat{y}_i - \bar{y})]^2$$
$$= \sum_{i=1}^{n} [\hat{u}_i - (\hat{y}_i - \bar{y})]^2$$

and using that the fitted values and residuals are uncorrelated, can show

$$SST = SSE + SSR$$

- Assuming $SST > 0$, we can define the fraction of the total variation in $y_i$ that is explained by $x_i$ (or the OLS regression line) as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Called the R-squared of the regression.
- It can be shown to equal the square of the correlation between $y_i$ and $\hat{y}_i$. Therefore,

$$0 \leq R^2 \leq 1$$

- $R^2 = 0$ means no linear relationship between $y_i$ and $x_i$. $R^2 = 1$ means a perfect linear relationship.
- As $R^2 = 0$ increases, the $y_i$ are closer and closer to falling on the OLS regression line.
- Do not want to fixate on $R^2$. It is a useful summary measure but tells us nothing about causality. Having a "high" R-squared is neither necessary nor sufficient to infer causality.

# Heteroskedasticity

## Defintion

*Heteroskedasticity: the variance of the unobserved factors changes across different segments of the population*

The five Gauss-Markov Assumptions for OLS regression:

- MLR.1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- MLR.2: random sampling from the population
- MLR.3: no perfect collinearity in the sample
- MLR.4: $E(u \mid x_1, \ldots, x_k) = E(u) = 0$ (exogenous explanatory variables)
- MLR.5: $\text{Var}(u \mid x_1, \ldots, x_k) = \text{Var}(u) = \sigma^2$ (homoskedasticity)

- But what if we drop MLR.5 and act as if we know nothing about

$$\text{Var}\,(u \mid x_1, \ldots, x_k) = \text{Var}(u \mid \mathbf{x})?$$

- OLS is still unbiased and consistent under MLR.1 to MLR.4. (We did not use MLR.5 to obtain either of these properties.) This is an important conclusion: Heteroskedasticity does not cause bias or inconsistency in the $\hat{\beta}_j$.

# Consequences of Heteroskedasticity for OLS

- However, if $Var(u \mid \mathbf{x})$ depends on x –that is, heteroskedasticity is present –then OLS is no longer BLUE (best linear unbiased estimator). In principle, it is possible to find unbiased estimators that have smaller variances than the OLS estimators. As similar comment holds for asymptotic efficiency.
- Practically, a more important point is that the usual standard errors are no longer valid, which means the t statistics and confidence intervals that use these standard errors cannot be trusted. This is true even in large samples.
- Joint hypotheses tests using the usual F statistic are no longer valid in the presence of heteroskedasticity.

- Witout MLR.5, there are still good reasons to use OLS, but we need to modify the usual test statistics to make them valid in the presence of heteroskedasticity.
- We are not talking about a new estimation method. It is still OLS estimation to obtain the $\hat{\beta}_j$ . But we need to use heteroskedasticity-robust inference after OLS estimation. We will see how to do that.

# Heteroskedasticity-Robust Inference after OLS Estimation

- Fortunately, standard errors and all test statistics can be modified to be valid in the presence of heteroskedasticity of unknown form. This includes the possibility of homoskedasticity, that is, MLR.5 actually holds. So we can compute CIs and conduct statistical inference without worrying about whether MLR.5 holds.
- Most regression packages include an option with OLS estimation that computes heteroskedasticity-robust standard errors, which then produces heteroskedasticity-robust t statistics and heteroskedasticity-robust confidence intervals.

- In Stata, the general command is: reg y x1 x2 ... xk, robust, where "robust" means "robust to heteroskedasticity of any form."
- Using heteroskedasticity-robust statistics for multiple exclusion restrictions is straightforward because many econometrics packages now compute such statistics routinely.

## Example

*Using COLLEGE.DTA (with an uneccesary fourth digit to emphasize that the robust standard errors are different)*

In this example, the robust standard errors are all slightly larger than the usual standard errors, but this has little consequence. (CIs are slightly wider, t statistics slightly lower.)

```
. reg lwage female exper coll

      Source |       SS       df       MS              Number of obs =     750
-------------+------------------------------           F(  3,   746) =  107.44
       Model |   54.396801     3   18.132267           Prob > F      =  0.0000
    Residual |  125.899028   746  .168765454           R-squared     =  0.3017
-------------+------------------------------           Adj R-squared =  0.2989
       Total |  180.295829   749  .240715393           Root MSE      =  .41081

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2202328   .0317963    -6.93   0.000    -.2826537   -.1578118
       exper |   .0521043   .0058418     8.92   0.000     .0406361    .0635725
        coll |   .0761594   .0065847    11.57   0.000     .0632326    .0890862
       _cons |   1.649161   .0720279    22.90   0.000     1.507759    1.790562
------------------------------------------------------------------------------

. reg lwage female exper coll, robust

Linear regression                               Number of obs =     750
                                                F(  3,   746) =  115.21
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3017
                                                Root MSE      =  .41081

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2202328   .0324685    -6.78   0.000    -.2839732   -.1564923
       exper |   .0521043   .0059534     8.75   0.000     .0404168    .0637917
        coll |   .0761594   .0067958    11.21   0.000     .0628183    .0895005
       _cons |   1.649161    .075371    21.88   0.000     1.501196    1.797125
------------------------------------------------------------------------------
```

# Testing for Heteroskedasticity

- Before the discovery of heteroskedasticity-robust inference, the common approach to heteroskedasticity was to first test for it and then, if it was found (at a sufficiently small significance level), to abandon OLS for weighted least squares.
- But with simple adjustments to the usual OLS test statistics, there is less of a case for even testing for heteroskedasticity.

A few reasons to test for heteroskedasticity include:

- (1) We may want to know whether we need to report robust standard errors.
- (2) We may want to know whether we can improve over OLS, which is possible if there is heteroskedasticity.
- (3) We may actually want to determine whether the variability in y about its mean changes with the values of the $x_j$.

# Testing for Heteroskedasticity

- For example, are wages (or log wages) more or less variable for women than men, holding other factors fixed? Are average test scores (or pass rates) really less variable in larger schools –after controlling for certain factors?
- When our data have a time dimension, we might want to know what has happened to the error variance over time.

# Steps in Computing the Breusch-Pagan Test

- Estimate the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$ by OLS, saving the OLS residuals, $\hat{u}_i$. Compute the squared residuals, $\hat{u}_i^2$ . (There is a squared residual for each of the n observations.)
- Regress $\hat{u}_i^2$ on all explanatory variables and compute the usual F test of joint significance of the explanatory variables.
- If the p-value of the test in step 2 is sufficiently small, reject the null of homoskedasticity and conclude Assumption MLR.5 fails.

# Generic Stata commands:

## Note

*reg y x1 x2 ... xk*
*predict uh, resid*
*gen uhsq* $= uh^2$
*reg uhsq x1 x2 ... xk*

- After the final regression, we look at the F statistic in the upper right hand corner.
- A statistic that allows the condtional fourth moment, $E(u^4|x)$, to depend on x under $H_0$, uses reg uhsq x1 x2 ... xk, robust.

## Example

*Wage Equations Using WAGE2.DTA*

Start with the equation where wage is the dependent variable.

```
. reg wage female exper coll

      Source |       SS       df       MS              Number of obs =     750
-------------+------------------------------           F(  3,   746) =   85.47
       Model |  5384.33651     3  1794.77884            Prob > F      =  0.0000
    Residual |  15664.6139   746   20.998142           R-squared     =  0.2558
-------------+------------------------------           Adj R-squared =  0.2528
       Total |  21048.9504   749  28.1027376           Root MSE      =  4.5824

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.457225   .3546709    -6.93   0.000    -3.153497   -1.760954
       exper |   .4158217   .0651616     6.38   0.000     .2878998    .5437436
        coll |   .8004933   .0734492    10.90   0.000     .6563015     .944685
       _cons |   5.785301    .803433     7.20   0.000     4.208042     7.36256
------------------------------------------------------------------------------

. predict uh, resid
```

- If we forget to square the residuals, we get nonsense: the F statistic is zero because, by construction, the OLS residuals are uncorrelated with all explanatory variables.

```
. reg uh female exper coll

      Source |       SS       df       MS              Number of obs =     750
-------------+------------------------------           F(  3,   746) =    0.00
       Model |  3.6380e-12     3  1.2127e-12            Prob > F      =  1.0000
    Residual |  15664.6138   746  20.9981419           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0040
       Total |  15664.6138   749  20.9140372           Root MSE      =  4.5824

------------------------------------------------------------------------------
          uh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -3.64e-09   .3546709    -0.00   1.000    -.6962719    .6962719
       exper |   4.39e-10   .0651616     0.00   1.000    -.1279219    .1279219
        coll |  -1.57e-09   .0734492    -0.00   1.000    -.1441917    .1441917
       _cons |   8.76e-10    .803433     0.00   1.000    -1.577259    1.577259
------------------------------------------------------------------------------
```

```
. gen uhsq = uh^2

. reg uhsq female exper coll

      Source |       SS       df       MS              Number of obs =     750
-------------+------------------------------           F(  3,   746) =    7.73
       Model |  90564.9383     3  30188.3128            Prob > F      =  0.0000
    Residual |  2914948.07   746  3907.43709            R-squared     =  0.0301
-------------+------------------------------           Adj R-squared =  0.0262
       Total |  3005513.01   749  4012.70094            Root MSE      =  62.509

------------------------------------------------------------------------------
        uhsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -13.51787   4.838171    -2.79   0.005    -23.01592   -4.019825
       exper |  -.0778488   .8888885    -0.09   0.930    -1.822869    1.667172
        coll |   3.472537   1.001942     3.47   0.001     1.505575    5.439498
       _cons |   21.24324   10.95987     1.94   0.053    -.2726092    42.75909
------------------------------------------------------------------------------
```

- The Breusch-Pagan F statistic is 7. 73 giving a p-value of zero to four decimal places. This is a very strong statistical rejection of the null of homoskedasticity.
- There are other variations of tests for heteroskedasticity. The White test for heteroskedasticy includes the explanatory variables, as with the B-P test, but also the nonredundant squares and interactions of all explanatory variables.