# ECON 3740: INTRODUCTION TO ECONOMETRICS

Instructor: Chaoyi Chen
Department of Economics and Finance, University of Guelph

Lecture 3

# Lecture outline

Last lecture, we reviewed how to measure the shape of a probability distribution, the joint distribution, and some often used probability distributions. Today, we will go through

- Steps in empirical economic analysis
- Data
  - Cross-sectional data
  - Time-series data
  - Pooled cross-sectional data
  - Panel data
- Linear regression with 1 regressor
  - Model Terminologies
  - Graphical illustration
  - The assumptions
  - Ordinary least square estimators derivation

# Steps in Empirical Economic Analysis

- What is econometrics?
  - Econometrics = use of statistical methods to analyze economic data
  - Econometricians typically analyze nonexperimental data
- Typical goals of econometric analysis
  - Estimating relationships between economic variables
  - Testing economic theories and hypotheses
  - Forecasting economic variables
  - Evaluating government and business policy
- What is an empirical analysis?
  - An empirical analysis uses data to test a theory or to estimate a relationship.
- Steps for an empirical analysis
  - Step 1: Economic model (this step is often skipped). Examples are Utility Maximization, Profit Maximization, Government Objectives, Political Objectivs.
  - Step 2: Econometric model

# Steps in empirical economic analysis: an example from Becker (1968)

- Economic model of criminal activity
  - In a seminal article, Nobel Prize winner Gary Becker postulated a utility maximization framework to describe an individuals participation in crime.

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

  - where $y$ is hours spent in criminal activities. Check your textbook or the article to identify $x_1, ..x_7$.
  - However, economic theory does not specify the functional form $(f(.))$.

- Econometric model of criminal activity
  - We can assume a linear relationship to specify the functional form.

$$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \mu$$

  - where $\mu$ is unobserved determinants of criminal activity

# Causality

- Causality: The most important thing you will learn in this class
- Very hard to determine in a non-experimental setting. Examples:
  - Considering maternal smoking and infant birth weight, will the smoking mother give the same weight as the non-smoking mothers?
  - Considering the temperature and $CO_2$, does more $CO_2$ lead to higher temperatures or higher temperatures lead to more $CO_2$?
  - Considering public debt and GDP, does public debt have a positive/negative effect on GDP only or GDP have a positive/negative effect on public debt only or the relationship is bidirectional?

# Data - cross-sectional

- Cross-sectional Data
  - Sample of agents taken at one point in time.
  - Ideally, the data is a random sample and observations are independent.
  - We mainly focus on cross-sectional data in this course!

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

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

- Question are cross-sectional observation independent?

# Data - time-Series

- Time-series Data
  - Repeat observations on specific agents over time. Examples include stock prices, money supply, consumer price index, GDP...

**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 |

- Question are time-series observation independent?

# Data - pooled cross-sectional

- Pooled Cross Sections
  - have both cross-sectional and time series features.
  - the point of a pooled cross-sectional analysis is often to see how a key relationship has changed over time.

**TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices**

| obsno | year | hprice | proptax | sqrft | bdrms | bthrms |
|---|---|---|---|---|---|---|
| 1 | 1993 | 85,500 | 42 | 1600 | 3 | 2.0 |
| 2 | 1993 | 67,300 | 36 | 1440 | 3 | 2.5 |
| 3 | 1993 | 134,000 | 38 | 2000 | 4 | 2.5 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 250 | 1993 | 243,600 | 41 | 2600 | 4 | 3.0 |
| 251 | 1995 | 65,000 | 16 | 1250 | 2 | 1.0 |
| 252 | 1995 | 182,400 | 20 | 2200 | 4 | 2.0 |
| 253 | 1995 | 97,500 | 15 | 1540 | 3 | 2.0 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 520 | 1995 | 57,200 | 16 | 1100 | 2 | 1.5 |

# Data - panel

- Panel-Data
  - Have repeat observations for the same agents in different time periods.
  - The key feature of panel data that distinguishes them from a pooled cross section is that the *same* cross-sectional units are followed over a given time period.

**TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics**

| obsno | city | year | murders | population | unem | police |
|---|---|---|---|---|---|---|
| 1 | 1 | 1986 | 5 | 350,000 | 8.7 | 440 |
| 2 | 1 | 1990 | 8 | 359,200 | 7.2 | 471 |
| 3 | 2 | 1986 | 2 | 64,300 | 5.4 | 75 |
| 4 | 2 | 1990 | 1 | 65,100 | 5.5 | 75 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 297 | 149 | 1986 | 10 | 260,700 | 9.6 | 286 |
| 298 | 149 | 1990 | 6 | 245,000 | 9.8 | 334 |
| 299 | 150 | 1986 | 25 | 543,000 | 4.3 | 520 |
| 300 | 150 | 1990 | 32 | 546,200 | 5.2 | 493 |

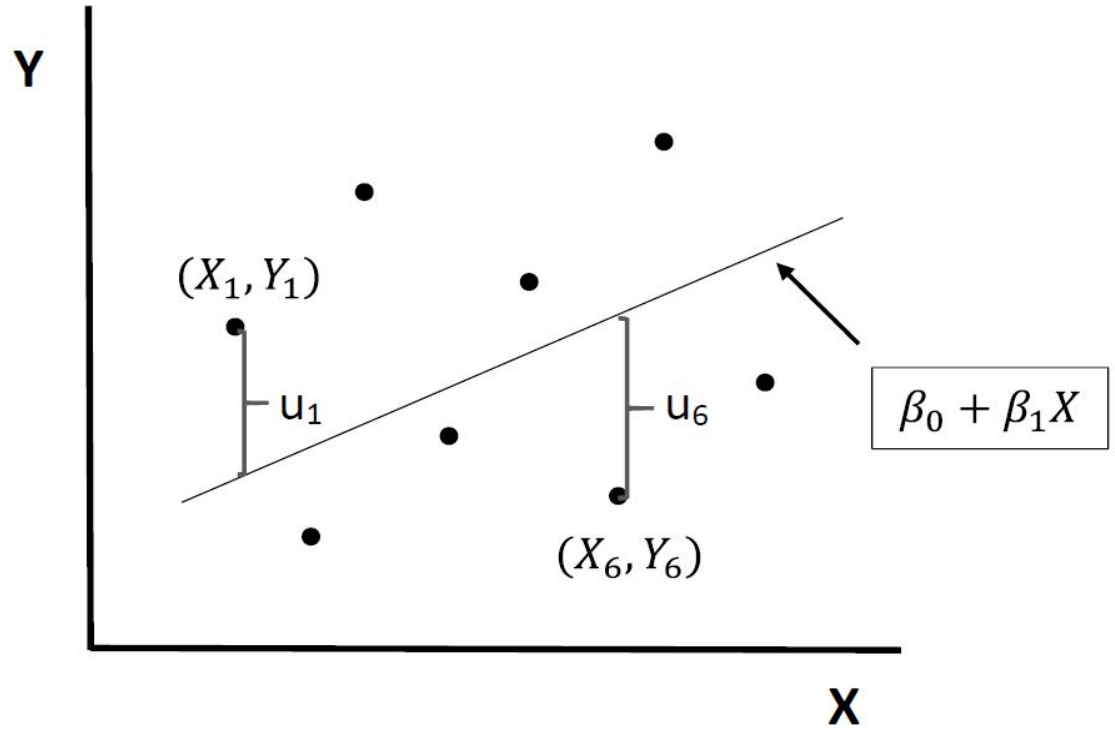
# Linear regression with one regressor: terminology

- The linear regression model with one regressor is

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

- where
  - $Y_i$ is the dependent variable/ regressand/ explained variable / LHS variable.
  - $X_i$ is the independent variable / regressor / explanatory variable / RHS variable.
  - $\mu_i$ is the error term, or "disturbance" term, which contains all other factors determining $_i$)
  - $\beta_0$ is the intercept of the *population* regression line (expected value of $Y$ when $X = 0$)
  - $\beta_1$ is the slope of the *population* regression line
  - $\beta_0 + \beta_1 X_i$ is the *population* regression line

# Linear regression with one regressor: a graphical example



# Linear regression with one regressor: assumptions

- The model

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

- Assumption 1:

$$E(\mu) = 0$$

- Remark 1: This is innocuous as long as we have an intercept in the model.

- Assumption 2:

$$E(\mu|X) = E(\mu)$$

- Combined with assumption 1 this gives us

$$E(\mu|X) = 0$$

- This means that given *any* x, the value of $\mu$ we expect will be 0. In other word, $\mu$ is mean independent of $X$
- Remark 2: This is not necessarily realistic and the hard assumption to satisfy.

# Linear regression with one regressor: an example to understand assumptions

- To see why $E(\mu|X) = 0$ is hard to satisfy, let's consider an example
- A simple wage model relating a persons wage to observed education and other unobserved factors can be characterized as

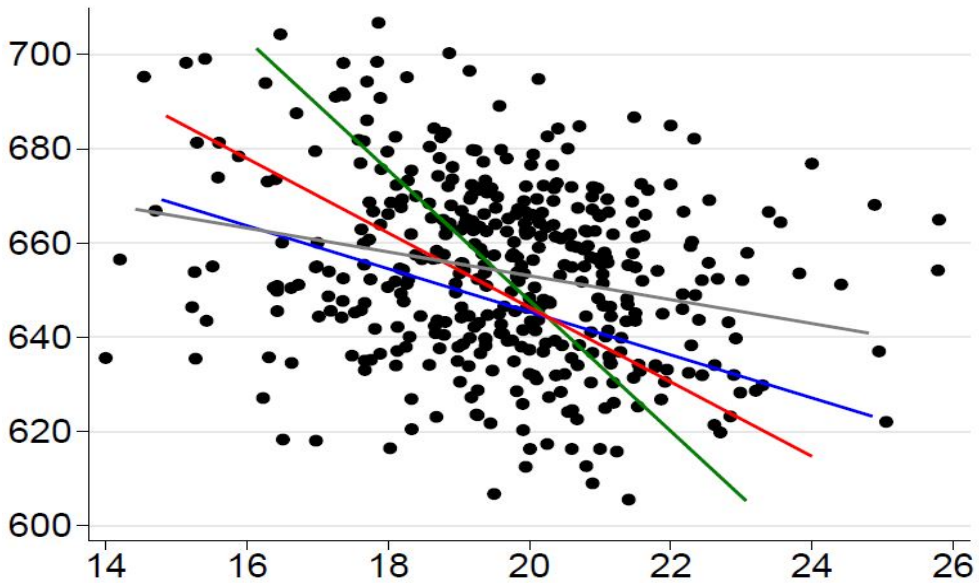
$$\text{wage} = \beta_0 + \beta_1 \times \textit{educ} \ + \mu$$

- where *wage* is measured in dollars per hour and *educ* is years of education.

- To simplify the discussion, we can interpret $\mu$ as innate ability (all other innate factors affect wage).
- Then, $E(\mu|X) = 0$ requires that the average level of ability is the same $=0$, regardless of years of education.
- However, average ability may increase with years of education (Intuitively, people with more ability choose to become more educated), which violates the assumption.

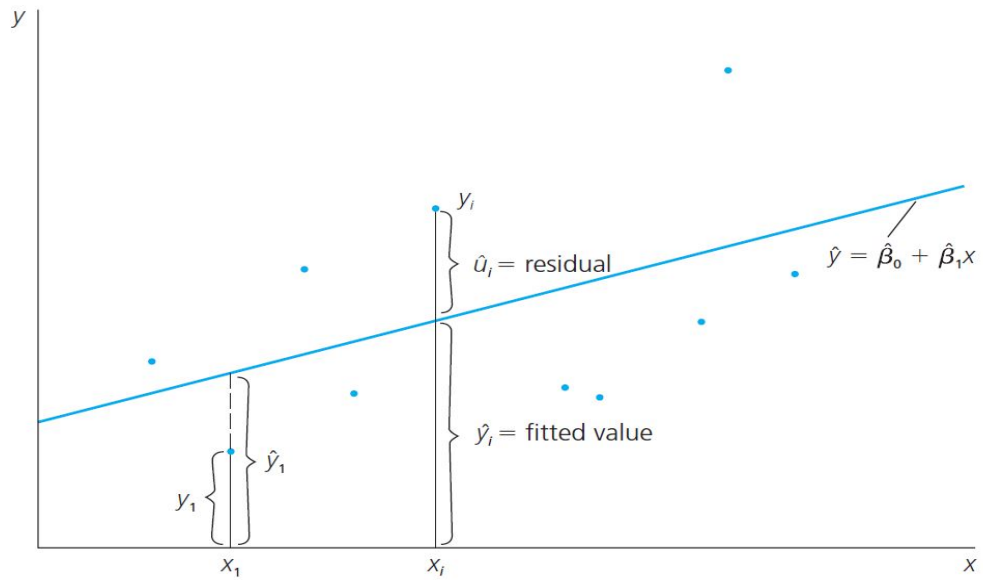
# The ordinary least squares estimator (OLS)

- In general we don't know $\beta_0$ and $\beta_1$. But, we can estimate them using a random sample of data.
- We just need to find the line that fits the data best. However, the problem is how we define the "best fits" and how we estimate the model based on the definition



# The ordinary least squares estimator (OLS)

- Definition for the best fit line: the best fit line should minimize the sum of squared residuals ($\hat{\mu}_i = y_i - \hat{y}_i$).




The estimators of $\beta_0$ and $\beta_1$ obtained from above definition are called as *ordinary least square* (OLS) estimator.

# The ordinary least squares estimator (OLS)

- The OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared mistakes made in predicting $Y$ given $X$
- Let $\hat{\beta}_0$ and $\beta_1$ be estimators of $\beta_0$ and $\beta_1$
- Predicted Value

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

- Residual

$$\hat{\mu}_i = y_i - \hat{y}_i$$

- Thus, the estimators of the slope and intercept can be obtained by minimizing

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{\mu}^2$$

# The ordinary least squares estimator (OLS): derivation

- We can rewrite previous objective function as

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

- How we solve for the $\hat{\beta}_0$ and $\hat{\beta}_1$? Take derivatives!
- Differentiate $\sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$ with respect to $\beta_0$, we have

$$\sum_{i=1}^{n} -2\left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

- Divide by -2, then divide by $n$, we have

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

- Question: to which assumption does this equation correspond?

# The ordinary least squares estimator (OLS): derivation

- Similarly, we can Differentiate $\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2$ with respect to $\beta_1$, we have

$$\sum_{i=1}^{n}(-2)x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0$$

- Divide by -2, then divide by $n$, we have

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

- Question: to which assumption does this equation correspond?