LECTURE 2

Introduction to Econometrics

INTRODUCTION TO LINEAR REGRESSION ANALYSIS I

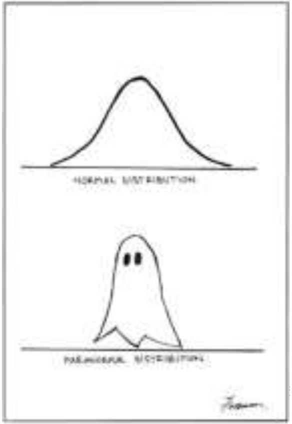
Hieu Nguyen

Fall Semester, 2024

PREVIOUS LECTURE...

Introduction, organization, review of statistical background

- random variables
- mean, variance, standard deviation
- covariance, correlation, independence
- normal distribution
- standardized random variables



WARM-UP EXERCISE

- What is the correlation between X and Y?

$$\begin{array}{cc} X & Y \\ \end{array}$$
$$\begin{pmatrix} 5 & 10 \\ 3 & 6 \\ -1 & -4 \\ 6 & 8 \\ 2 & 5 \end{pmatrix}$$

- **Correlation:** $Corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$
- **Covariance:**

  $$Cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right] = E\left[XY\right] - E[X]E[Y]$$
- **Standard deviation :** $\sigma_X = \sqrt{Var[X]}$
- **Variance:** $Var[X] = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$

LECTURE 2.

- **Introduction to simple linear regression analysis**

  Sampling and estimation

  OLS principle

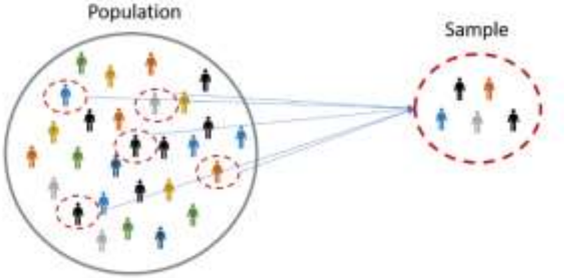
- Readings:

  Studenmund, A. H., Using Econometrics: A Practical Guide, Chapters 1, 2.1, 16.1, 16.2

  Wooldridge, J. M., Introductory Econometrics: A Modern Approach, Chapters 2.1, 2.2

SAMPLING

- **Population**: the entire group of items that interests us
- **Sample**: the part of the population that we actually observe
- **Statistical inference**: use of the sample to draw conclusion about the characteristics of the population from which the sample came
- Examples: medical experiments, opinion polls



Random sampling vs selection bias

- Correct statistical inference can be performed only on a **random sample** - a sample that reflects the true distribution of the population
- **Biased sample**: any sample that differs systematically from the population that it is intended to represent
- **Selection bias**: occurs when the selection of the sample systematically excludes or under represents certain groups
  - Example: opinion poll about tuition payments among undergraduate students vs all citizens
- **Self-selection bias**: occurs when we examine data for a group of people who have chosen to be in that group
  - Example: accident records of people who buy collision insurance

EXERCISE 1

- American Express and the French tourist office sponsored a survey that found that most visitors to France do not consider the French to be especially unfriendly.
- The sample consisted of 1,000 Americans who have visited France more than once for pleasure over the past two years.
- Is this survey unbiased?

ESTIMATION

- **Parameter**: a true characteristic of the distribution of a variable, whose value is unknown, but can be estimated

  Example: population mean $E[X]$

- **Estimator**: a sample statistic that is used to estimate the value of the parameter

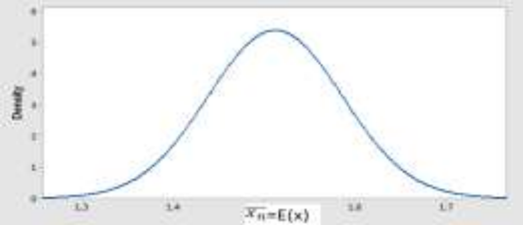
  Example: sample mean $\overline{X}_n$

  Note that the estimator is a random variable (it has a probability distribution, mean, variance,...)

- **Estimate**: the specific value of the estimator that is obtained on a specific sample

Properties of an estimator

- An estimator is **unbiased** if the mean of its distribution is equal to the value of the parameter it is estimating
- An estimator is **consistent** if it converges to the value of the true parameter as the sample size increases
- An estimator is **efficient** if the variance of its sampling distribution is the smallest possible



EXERCISE 2

- A young econometrician wants to estimate the relationship between foreign direct investments (FDI) in her country and firm profitability.
- Her reasoning is that better managerial skills introduced by foreign owners increases firms' profitability.
- She collects a random sample of 8,750 firms and finds that one sixth of the firms were entered within last few years by foreign investors. The rest of the firms are owned domestically.
- When she compares indicators of profitability, such as ROA and ROE, between the domestic and foreign-owned firms, she finds significantly better outcomes for foreign-owned firms.
- She concludes that FDI increases firms' profitability. Is this conclusion correct?

ECONOMETRIC MODELS

- Econometric model is an estimable formulation of a theoretical relationship
- Theory says: $Q = f(P, P_s, Y)$

  $Q \ldots$ quantity demanded

  $P \ldots$ commodity's price

  $P_s \ldots$ price of substitute good

  $Y \ldots$ disposable income
- We simplify: $Q = \beta_0 + \beta_1 P + \beta_2 P_s + \beta_3 Y$
- We estimate: $Q = 31.50 - 0.73P + 0.11P_s + 0.23Y$

Econometric models

- Today's econometrics deals with different, even very general models
- During this course we will cover just linear regression models
- We will see how these models are estimated by

  Ordinary Least Squares (OLS)

  Generalized Least Squares (GLS)

  Instrumental Variables (IV)
- We will perform estimation on different types of data

DATA USED IN ECONOMETRICS

cross-section	**repeated cross-section**
sample of units (eg. firms, individuals) taken at a given point in time	several independent samples of units (eg. firms, individuals) taken at different points in time

time-series	**panel data**
observations of variable(s) in different points in time (eg. GDP)	time series for each cross-sectional unit in the data set (eg. GDP of various countries)

Data used in econometrics - Examples

- Country's macroeconomic indicators (GDP, inflation rate, net exports, etc.) month by month
- Data about firms' employees or financial indicators as of the end of the year
- Records of bank clients who were given a loan
- Annual social security or tax records of individual workers

STEPS OF AN ECONOMETRIC ANALYSIS

1. Formulation of an economic model (rigorous or intuitive)
2. Formulation of an econometric model based on the economic model
3. Collection of data
4. Estimation of the econometric model
5. Interpretation of results

EXAMPLE - ECONOMIC MODEL

- Denote:

  p … price of the good
  c … firm's average cost per one unit of output
  $q(p)$ … demand for firm's output

Firm profit:

$$\pi = q(p) \cdot (p - c)$$

Demand for good:

$$q(p) = a - b \cdot p$$

- Derive:

$$q = \frac{a}{2} - \frac{b}{2} \cdot c$$

- We call q dependent variable and c explanatory variable

EXAMPLE - ECONOMETRIC MODEL

- Write the relationship in a simple linear form

$$q = \beta_0 + \beta_1 c$$

(have in mind that $\beta_0 = \frac{a}{2}$ and $\beta_1 = -\frac{b}{2}$

- There are other (unpredictable) things that influence firms' sales ⇒ add disturbance term

$$q = \beta_0 + \beta_1 c + \varepsilon$$

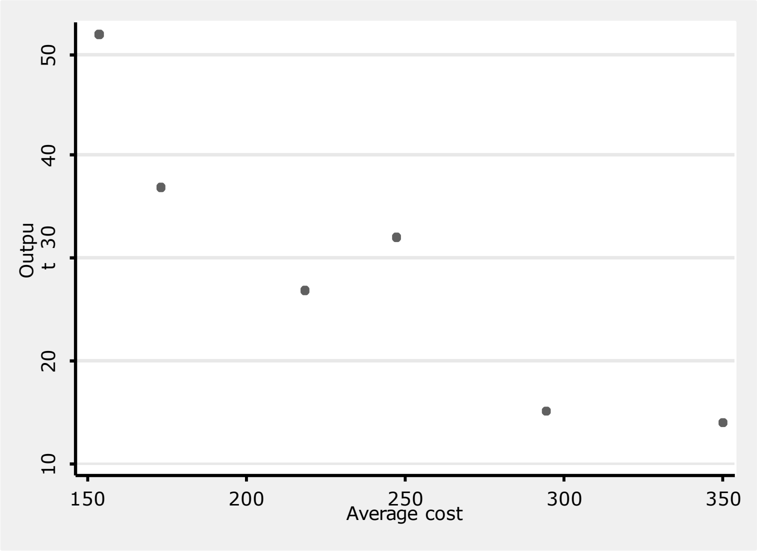
- Find the value of parameters β_1 (slope) and β_0 (intercept)

Example - Data

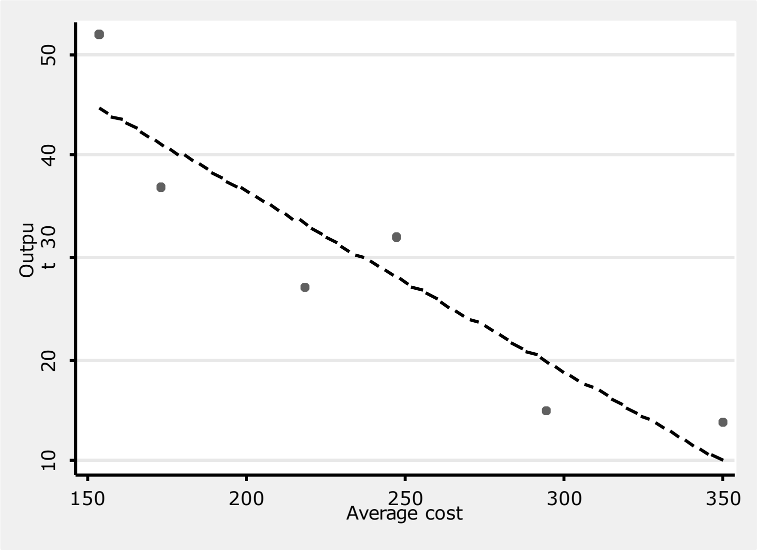
- Ideally: investigate all firms in the economy
- Reality: investigate a sample of firms
  We need a random (unbiased) sample of firms
- Collect data:

Firm	1	2	3	4	5	6
q	15	32	52	14	37	27
c	294	247	153	350	173	218

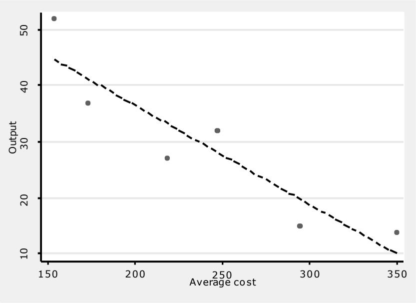
EXAMPLE - DATA



EXAMPLE - ESTIMATION



Example - Estimation



OLS method:

Make the fit as good as possible
$\Downarrow$
Make the misfit as low as possible

Minimize the (vertical) distance between data points and regression line

Minimize the sum of squared deviations

Terminology

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ldots \text{ regression line}$$

$y_i \ldots$ dependent/explained variable (i-th observation)

$x_i \ldots$ independent/explanatory variable (i-th observation)

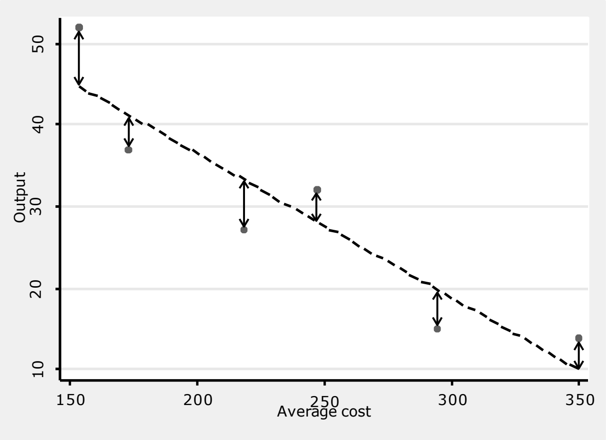
$\varepsilon_i \ldots$ random error term/disturbance (of i-th observation)

$\beta_0 \ldots$ intercept parameter ($\hat{\beta}_0 \ldots$ estimate of this parameter)

$\beta_1 \ldots$ slope parameter ($\hat{\beta}_1 \ldots$ estimate of this parameter)

ORDINARY LEAST SQUARES

- OLS = fitting the regression line by minimizing the sum of vertical distance between the regression line and the observed points



Ordinary Least Squares - principle

- Take the squared differences between observed point y_i and regression line $\beta_0 + \beta_1 x_i$:

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

- Sum them over all n observations:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize this sum

$$\left[\widehat{\beta}_0, \widehat{\beta}_1\right] = \operatorname*{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Ordinary Least Squares - derivation

$$\left[\widehat{\beta}_0, \widehat{\beta}_1\right] = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- FOC:

$$\frac{\partial}{\partial \beta_0} : \qquad -2\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0$$

$$\frac{\partial}{\partial \beta_1} : \qquad -2\sum_{i=1}^{n} x_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0$$

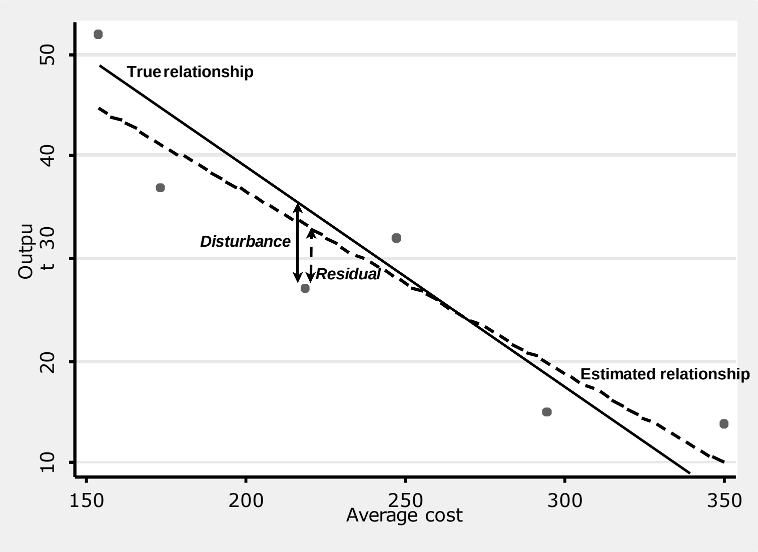
- We express:

$$\widehat{\beta}_0 = \overline{y}_n - \widehat{\beta}_1 \overline{x}_n \qquad\qquad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}$$

RESIDUAL

- Residual is the vertical difference between the estimated regression line and the observation points
- OLS minimizes the sum of squares of all residuals
- It is the difference between the true value y_i and the estimated value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- We define: $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- Residual e_i (observed) is not the same as the disturbance ε_i (unobserved)!!!
- Residual is an estimate of the disturbance: $e_i = \hat{\varepsilon}_i$

RESIDUAL VS. DISTURBANCE



Getting back to the example

We have the economic model

$$q = \frac{a}{2} - \frac{b}{2} \cdot c$$

We estimate

$$q_i = \beta_0 + \beta_1 c_i + \varepsilon_i$$

(having in mind that $\beta_0 = \frac{a}{2}$ and $\beta_1 = -\frac{b}{2}$)

Our data:

Firm	1	2	3	4	5	6
q	15	32	52	14	37	27
c	294	247	153	350	173	218

Getting back to the example

- When we plug in the formula:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{6}(c_i - \overline{c})(q_i - \overline{q})}{\sum_{i=1}^{6}(c_i - \overline{c})^2} = -0.177$$

GETTING BACK TO THE EXAMPLE

- When we plug in the formula:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{6} (c_i - \overline{c})(q_i - \overline{q})}{\sum_{i=1}^{6} (c_i - \overline{c})^2} = -0.177$$

$$\widehat{\beta}_0 = \overline{q} - \widehat{\beta}_1 \overline{c} = 71.74$$

- The estimated equation is

$$\widehat{q} = 71.74 - 0.177c$$

and so

$$\widehat{a} = 2\widehat{\beta}_0 = 143.48 \quad \text{and} \quad \widehat{b} = -2\widehat{\beta}_1 = 0.353$$

MEANING OF REGRESSION COEFFICIENT

- Consider the model

$$q = \beta_0 + \beta_1 c$$

estimated as $\hat{q} = 71.74 - 0.177c$

q . . . demand for firm's output

c . . . firm's average cost per unit of output

- Meaning of β_1 is the impact of a one unit increase in c on the dependent variable q
- When average costs increase by 1 unit, quantity demanded decreases by 0.177 units

BEHIND THE ERROR TERM

- The stochastic error term must be present in a regression equation because of:

  1. omission of many minor influences (unavailable data)
  2. measurement error
  3. possibly incorrect functional form
  4. stochastic character of unpredictable human behavior

- Remember that all of these factors are included in the error term and may alter its properties

- The properties of the error term determine the properties of the estimates

SUMMARY

- We have learned that an econometric analysis consistsof
  1. definition of the model
  2. estimation
  3. interpretation
- We have explained the principle of OLS: minimizing the sum of squared differences between the observations and the regression line
- We have derived the formulas of theestimates:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \qquad \widehat{\beta}_0 = \overline{y}_n - \widehat{\beta}_1 \overline{x}_n$$

What's next

- In the next lectures, we will
  - derive estimation formulas for multivariate models
  - specify properties of the OLS estimator
  - start using Gretl for data description and estimation