

**The Federal State-Funded Educational Institution of Higher  
Professional Education "Financial University under the  
Government of the Russian Federation"**

**Department:  
Systems Analysis and Modeling of Economic Processes**

# Econometrics

Ph.D. in Technics,  
Sc.D. in Economics  
Prof.

Ilona V. Tregub  
E-mail: ilonavl\_fa@mail.ru

## References

### Theory

- Dougherty. Introduction to Econometrics. Oxford: Oxford University Press, 2011 fourth edition [ISBN 9780199567089].

### Practice

- Tregub I.V. Mathematical Model of Dynamics the Economic Systems. M.: Finacademia, 2009. -118 p.  
Трегуб И.В. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ДИНАМИКИ ЭКОНОМИЧЕСКИХ СИСТЕМ. – М.: Финакадемия, 2009. –118 с.
- Tregub I.V. PREDICTION OF ECONOMIC INDICATORS. M.: PSTM, 2009. -195p.  
Трегуб И.В. ПРОГНОЗИРОВАНИЕ ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ НА РЫНКЕ ДОПОЛНИТЕЛЬНЫХ УСЛУГ СОТОВОЙ СВЯЗИ. М.: ПСТМ, 2007. – 195с.

### Additionally

- Tregub I.V. Simulation. M.: Finacademia, 2007. -44p.  
Трегуб И.В. ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ. М.: Финакадемия, 2007. – 44с.

ScD., Prof. Ilona V. Tregub

## Content of Creative Research

1. **Describe** the economic system. **Formulate** the problem. **Determine** the endogenous and exogenous variables
2. **Collect the statistical data**
  - 2.1. Describe the statistical data related to the model
  - 2.2. Construct the scatter diagram
  - 2.3. Do a correlation analysis
3. **Construct the econometric model**
  - 3.1 Model specification
  - 3.2 Estimate the coefficients of the model in Excel
  - 3.3 Estimated model
  - 3.4 Interpret the coefficients of the model
  - 3.5 Tests:  $R^2$ -test; F-test; t-test; GQ-test; DW-test
  - 3.6 Construct the confidence interval
  - 3.7 Check the adequacy of the model
  - 3.8 Model forecasting
4. **Conclusions, recommendations**

© Tregub Ilona Vladimirovna

3

## 1. ECONOMIC DATA AND ECONOMETRIC ANALYSIS

Questions:

- 1.1. What is Econometrics?
- 1.2. Role of Econometrics.
- 1.3. Main Application of Econometrics.
- 1.4. Types of Economic Data.

© Tregub Ilona Vladimirovna

4

## 1.1. What is Econometrics?

In the first issue of *Econometrics*, the *Econometric Society* stated that

**«Econometrics is the field of economics that concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory.»**

© Tregub Ilona Vladimirovna

5

## 1.2. Role of Econometrics

**Develop statistical methods for**

- (i) Estimating economic relationship,
- (ii) Testing economic theories,
- (iii) Evaluating and implementing government and business policy.

© Tregub Ilona Vladimirovna

6

## 1.3. Main Application of Econometrics.

### **Forecasting:**

- Macroeconomic and Financial variables;
- The effect of specific policy;
- Economic time series.

Data used in econometrics are typically NON EXPERIMENTAL DATA, i.e. data are NOT obtained via controlled laboratory experiments.

Basically, the econometrician simply collects data, typically from official statistics.

© Tregub Ilona Vladimirovna

7

## **Internet Resources**

- <http://www.gks.ru>
- <http://www.cbr.ru>
- <http://www.finam.ru>
- <http://data.worldbank.org/data-catalog>
- <http://stats.uis.unesco.org/unesco/ReportFolders/ReportFolders.aspx>
- <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/>
- <http://e3.prime-tass.ru/macro/>
- <http://research.stlouisfed.org/fred2/categories>
- <http://hdr.undp.org/en/data>
- <http://www.tns-global.ru/rus/index.wbp>
- <http://www.numbeo.com/cost-of-living/>

© Tregub Ilona Vladimirovna

8

## 1.4.Types of Economic Data

- Cross Section Data.
- Time Series Data.
- Panel Data.

© Tregub Ilona Vladimirovna

9

### *Cross Sectional Data*

Dispersed data relating to one period, or without respect to variance due to time.

We observe  $n$  individuals (consumers, firms, households, etc) at a given moment of time.

*Cross sectional data* often used in microeconomics.

© Tregub Ilona Vladimirovna

10

## *Time Series Data*

A sequence of data points, measured typically at successive times spaced at uniform time intervals.

We observe one variable (or several variables) over  $T$  periods.

Generally, macroeconomic and financial data are time series data.

© Tregub Ilona Vladimirovna

11

## *Panel Data*

Panel data contains observations on multiple phenomena observed over multiple time periods for the same objects (firms or individuals)

We observe a cross section of  $n$  individuals over  $T$  period of time.

*Example.* We observe the consumption expenditure on non-durables of  $n$  household over  $T$  periods.

© Tregub Ilona Vladimirovna

12

## 2. FORMULATION OF THE ECONOMETRIC MODEL

Questions:

2.1. Basic concept and definitions

2.2. Specification of the Econometrics Model. First Principle of the Specification

2.3. General Scheme of the Econometric Model

© Tregub Ilona Vladimirovna

13

### 2.1. Basic concept and definitions

The Basic concept of econometrics are "the object", "the variable" and "the model"

• **Object** is an economic, business or finance unit

- the National or Global Economy
- Households
- Firms
- Farms
- Financial Markets

ScD., Prof. Ilona V. Tregub

## 2.1. Basic concept and definitions

The Basic concept of econometrics are "the object", "the variable" and "the model"



- **Variable** is a quantitative characteristic of the object, which can take different values in the process of the economic activity

There are two types of econometric variables

**Endogeneous variables** are internal variables of a model. Their values should be explain by an econometrics model

**Exogeneous variables** are external variables of a model. They should explain the value of the internal variable.

ScD., Prof. Ilona V. Tregub


## Some alternative names of variables

$y$	$x_1, x_2, \dots, x_k$
<b>Endogenous</b>	<b>Exogenous</b>
Dependent	Independent
Regressand	Regressors
Effect variable	Causal variables
Explained variable	Explanatory variables



## 2.1. Basic concept and definitions

The Basic concept of econometrics are "the object", "the variable" and "the model"

- 
- **Model** is a mathematical expression between the variables of the object

The model can be represented as a set of graphs or tables, or a system of mathematical equations and inequalities which connect together all the object variables

ScD., Prof. Ilona V. Tregub

## 2.2. Specification of a model First Principle of the specification

Model specification is a detailed description of the object behavior by the mathematical language

First Principle of the specification

Specification of a model is a result of a translation of the economics laws into mathematical language. Linear mathematical equations used to build the model whenever possible

© Tregub Ilona Vladimirovna

18

## 2.2. Specification of a model First Principle of the specification

### Example 1

Consider the competitive product market.

The task is to get the model specification, links the levels of demand ( $Y^d$ ) and supply ( $Y^s$ ) and the equilibrium price ( $p$ )

ScD., Prof. Ilona V. Tregub

## 2.2. Specification of a model First Principle of the specification Solution of Example 1

### **Economic theory states**

1. The demand for goods depends on the price of the goods. Increasing the price of goods, leads to a decrease of demand for this product
2. Supply increases with increasing price
3. The equilibrium price corresponds to the equality between supply and demand

ScD., Prof. Ilona V. Tregub

## 2.2. Specification of a model First Principle of the specification

### Solution of Example 1

Write the statement (1-3) in mathematical language.

According to the first principle of specification, we will translate the economic laws in the language of mathematics, and will use linear algebraic functions

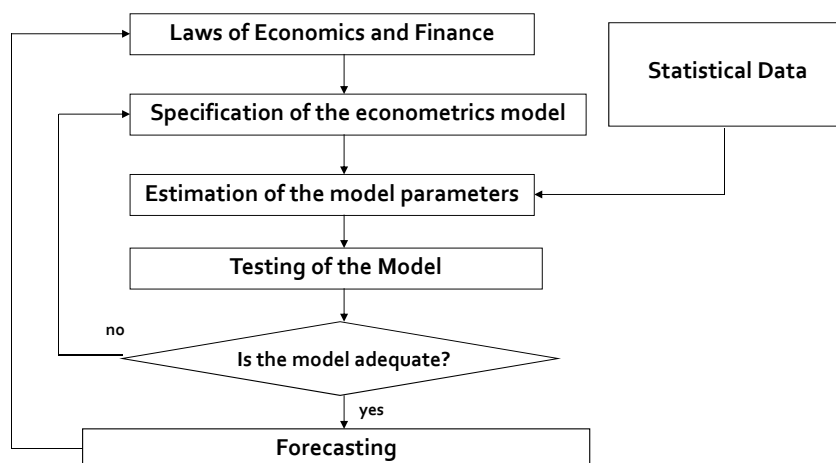
The model takes the form:

$$\begin{cases} Y^d = a_0 + a_1 p \\ Y^s = b_0 + b_1 p \\ Y^d = Y^s \\ a_1 < 0; (a_0, b_0, b_1) > 0 \end{cases}$$

- $Y^d$ ,  $Y^s$ ,  $p$  – variables,
- $a_0$ ,  $a_1$ ,  $b_0$ ,  $b_1$  – unknown parameters

ScD., Prof. Ilona V. Tregub

## General scheme of the econometric model



## 2. Formulation of the Econometric Model

Questions .

2.4. Types of Models

2.5. 2-d, 3-d Principles of Specification

2.6. Current and Lagged Variables. Predefined variables

2.7. Webby Model

2.8. 4-th Principle of Specification

### Types of models

#### Static and Dynamic Models

Models can be classified as

- Static model
  - Variables, properties of the objects or phenomenon are measured at the same moment of time
  - Model describes the objects or phenomenon as a photo
- Dynamic models
  - Variables can be measured at different times
  - Model describes the changes in the behavior of the objects or phenomenon

## 2-d, 3-d Principles of Specification of Econometric Models

- Number of equations in the econometric model should be equal to the number of endogenous variables, included into a model
- Each variable in the model should be dated, i.e. the moment of time when the variable is measured, should be clearly defined

## Current and Lagged Variables. Predefined variables

- Current / Modern  
measured at current moment of time,  
e.g.  $y_t, x_t$
- Lagged  
measured at the previous moment of time,  
e.g.  $x_{t-1}, y_{t-2} \dots$
- Predefined variables include lagged and exogenous variables

## Webby model of market equilibrium

Task.

Required to make the specification of the model, which explains levels of demand ( $Y^d$ ) and supply ( $Y^s$ ) of goods and its market price ( $p$ ) by means of the income per capita ( $x$ )

Use the following economic laws:

1. Levels of demand depends on the price and the income per capita.  
The level of demand decreases with increasing prices.  
The level of demand increases when incomes rise.
2. Levels of supply at the current moment of time can be explained by the price at the previous moment.  
The supply increases with increasing prices.
3. The market price is the result of the equality of supply and demand.

## Webby model of market equilibrium Structural (Initial) Form

$$\begin{cases} Y_t^d = a_0 + a_1 p_t + a_2 x_t \\ Y_t^s = b_0 + b_1 p_{t-1} \\ Y_t^d = Y_t^s \\ a_0, b_0, b_1 \geq 0; \quad a_1 \leq 0 \end{cases}$$

Here  $p_{t-1}$  means the price of the good in previous moment of time (one period ago)

Substitute the first and the second equation in the third equation, and write the formula for the price, demand and supply

## Webby model. Reduced Form.

$$\begin{cases} p_t = \frac{b_0 - a_0}{a_1} - \frac{a_2}{a_1} x_t + \frac{b_1}{a_1} p_{t-1} \\ Y_t^d = b_0 + b_1 p_{t-1} \\ Y_t^s = b_0 + b_1 p_{t-1} \end{cases}$$

## 4-th Principles of Specification of Econometric Models

- In the behavioral equation of the model should include disturbance term

Webby model

$$\begin{cases} Y_t^d = a_0 + a_1 p_t + a_2 x_t + \varepsilon_t \\ Y_t^s = b_0 + b_1 p_{t-1} + \nu_t \\ Y_t^d = Y_t^s \\ a_0, b_0, b_1 \geq 0; a_1 \leq 0 \\ E(\varepsilon_t) = 0; E(\nu_t) = 0 \\ \sigma(\varepsilon_t) = \text{const}; \sigma(\nu_t) = \text{const} \end{cases}$$

## 3. Correlation Analysis

### 3.1. Correlation

### 3.2. Scatter diagram

### 3.1. Correlation

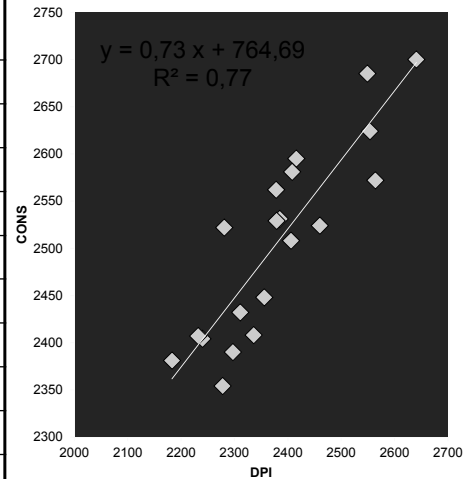
- The correlation coefficient may indicate that two variables are associated with one another, but it does not give any idea of the kind of relationship involved.
- If we say  $y$  and  $x$  are correlated, it means that we are treating  $y$  and  $x$  in a completely symmetrical way.

ScD., Prof. Ilona V. Tregub



## 3.2. Scatter diagram

#	Revenue, \$ DPI	Consumption, \$ CONS	#	Revenue, \$ DPI	Consumption, \$ CONS
1	2508	2406	11	2432	2311
2	2572	2564	12	2354	2278
3	2408	2336	13	2404	2240
4	2522	2281	14	2381	2183
5	2700	2641	15	2581	2408
6	2531	2385	16	2529	2379
7	2390	2297	17	2562	2378
8	2595	2416	18	2624	2554
9	2524	2460	19	2407	2232
10	2685	2549	20	2448	2356



## 4. SIMPLE REGRESSION ANALYSIS

Questions:

- 4.1. The Simple Linear Model;
- 4.2. Least Squares Regression;
- 4.3. Interpretation of a Regression Equation;
- 4.4. Goodness of Fit:  $R^2$ .

## 4.1. The Simple Linear Model

$$Y_i = \underbrace{\beta_1 + \beta_2 X_i}_1 + \underbrace{u_i}_2$$

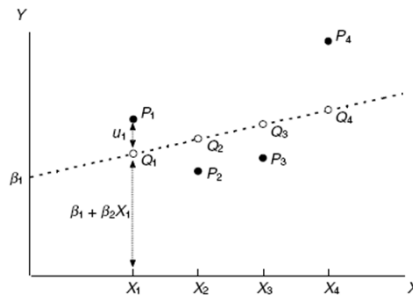


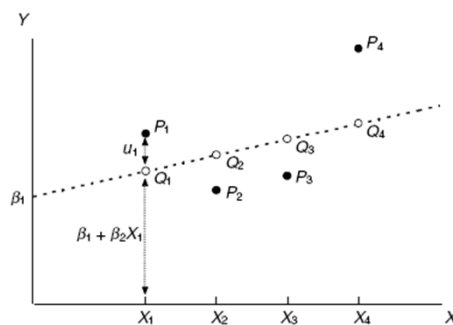
Figure illustrates how two components (1 and 2) combine to determine Y.

$X_1$ ,  $X_2$ ,  $X_3$ , and  $X_4$  are four hypothetical values of the explanatory variable X.

If the relationship between Y and X were exact, the corresponding values of Y would be represented by the points  $Q_1 - Q_4$  on the line.

## 4.1. The Simple Linear Model

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$



The disturbance term causes the actual values of Y to be different.

In the diagram, the disturbance term has been assumed to be positive in the first and fourth observations and negative in the other two, with the result that, if one plots the actual values of Y against the values of X, one obtains the points  $P_1 - P_4$ .

#### 4.1. The Simple Linear Model

- The actual values of  $\beta_1$  and  $\beta_2$ , and hence the location of the Q points, are unknown, as are the values of the disturbance term in the observations.
- The task of regression analysis is to obtain estimates of  $\beta_1$  and  $\beta_2$ , and hence an estimate of the location of the line, given the P points.

#### Why does the disturbance term exist?

There are several reasons.

**1. Omission of explanatory variables:**

The relationship between  $Y$  and  $X$  is almost certain to be a simplification. In reality there will be other factors affecting  $Y$  that have been left out of equation , and their influence will cause the points to lie off the line. It often happens that there are variables that you would like to include in the regression equation but cannot because you are unable to measure them.

## Why does the disturbance term exist?

### 2. Aggregation of variables:

In many cases the relationship is an attempt to summarize in aggregate a number of microeconomic relationships.

### 3. Model misspecification:

The model may be misspecified in terms of its structure.

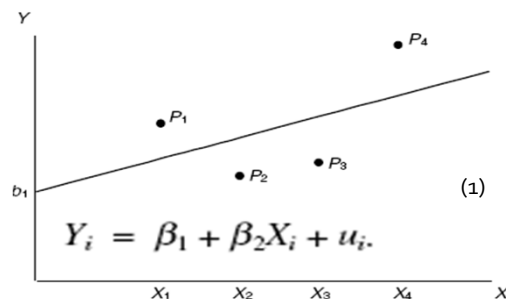
### 4. Functional misspecification:

The functional relationship between  $Y$  and  $X$  may be misspecified mathematically.

### 5. Measurement error:

If the measurement of one or more of the variables in the relationship is subject to error, the observed values will not appear to conform to an exact relationship, and the discrepancy contributes to the disturbance term.

## • 4.2. Least Squares Regression



Suppose that we are given the four observations on  $X$  and  $Y$  represented in Figure and we are asked to obtain estimates of the values of  $\beta_1$  and  $\beta_2$  in equation (1). As a rough approximation, you could do this by plotting the four  $P$  points and drawing a line to fit them as best you can.

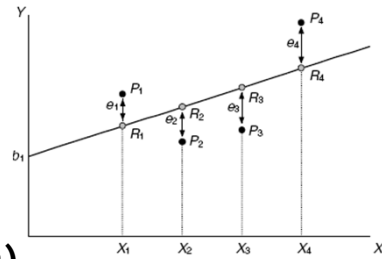
*This has been done in Figure*

The intersection of the line with the  $Y$ -axis provides an estimate of the intercept  $\beta_1$ , which will be denoted  $b_1$ , and the slope provides an estimate of the slope coefficient  $\beta_2$ , which will be denoted  $b_2$ .

## • 4.2. Least Squares Regression

The fitted line will be written

$$\hat{Y}_i = b_1 + b_2 X_i \quad (2)$$



the caret mark over Y indicating that it is the fitted value of Y corresponding to X, not the actual value.

In Figure, the fitted points are represented by the points  $R_1 - R_4$ .

$b_1$  and  $b_2$  are only estimates, and they may be good or bad.

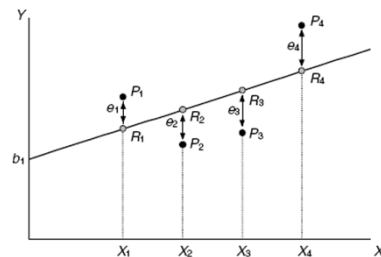
Is there a way of calculating good estimates of  $\beta_1$  and  $\beta_2$  algebraically?

The first step is to define what is known as a residual for each observation.

This is the difference between the actual value of Y in any observation and the fitted value given by the regression line, that is, the vertical distance between  $P_i$  and  $R_i$  in observation  $i$ .

It will be denoted  $e_i$ :  $e_i = Y_i - \hat{Y}_i$

(3)



Substituting (2) into (3), we obtain

$$e_i = Y_i - b_1 - b_2 X_i \quad (4)$$

and hence the residual in each observation depends on our choice of  $b_1$  and  $b_2$ .

Obviously, we wish to fit the regression line, that is, choose  $b_1$  and  $b_2$ , in such a way as to make the residuals as small as possible.

One way of overcoming the problem is to minimize *the sum of the squares of the residuals - RSS*.

In our case

$$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2$$

If one could reduce  $RSS$  to  $0$ , one would have a perfect fit, for this would imply that all the residuals are equal to  $0$ .

*The* line would go through all the points, but of course in general the disturbance term makes this impossible.

The form used here is usually referred to as **ordinary least squares** and abbreviated **OLS**.

## Least Squares Regression with One Explanatory Variable

We shall now consider the general case where there are  $n$  observations on two variables  $X$  and  $Y$  and, supposing  $Y$  to depend on  $X$ , we will fit the equation

$$\hat{Y}_i = b_1 + b_2 X_i.$$

The fitted value of the dependent variable in observation  $i$

$$\hat{Y}_i, \text{ will be } (b_1 + b_2 X_i),$$

and the residual  $e_i$  will be  $(Y_i - b_1 - b_2 X_i)$ .

We wish to choose  $b_1$  and  $b_2$  so as to minimize the residual sum of the squares, RSS, given by  $RSS = e_1^2 + \dots + e_n^2 = \sum_{i=1}^n e_i^2$

We shall assume that the true model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

and we shall estimate the coefficients  $b_1$  and  $b_2$  of the equation

$$\hat{Y}_i = b_1 + b_2 X_i.$$

Now we want to choose  $b_1$  and  $b_2$  to minimize RSS.

To do this, we calculate the partial derivative and find the values of  $b_1$  and  $b_2$  that satisfy

$$\frac{\partial RSS}{\partial b_1} = 0 \quad \text{and} \quad \frac{\partial RSS}{\partial b_2} = 0$$

We will begin by expressing the square of the residual in observation  $i$  in terms of  $b_1$ ,  $b_2$ , and the data on  $X$  and  $Y$ :

$$\begin{aligned} e_i^2 &= (Y_i - \hat{Y}_i)^2 = (Y_i - b_1 - b_2 X_i)^2 \\ &= Y_i^2 + b_1^2 + b_2^2 X_i^2 - 2b_1 Y_i - 2b_2 X_i Y_i + 2b_1 b_2 X_i \end{aligned}$$

$$RSS = (Y_1 - b_1 - b_2 X_1)^2 + \dots + (Y_n - b_1 - b_2 X_n)^2 =$$

Summing over all the  $n$  observations, we can write RSS as

$$\begin{aligned} &= Y_1^2 + b_1^2 + b_2^2 X_1^2 - 2b_1 Y_1 - 2b_2 X_1 Y_1 + 2b_1 b_2 X_1 \\ &\quad + \dots \\ &\quad + Y_n^2 + b_1^2 + b_2^2 X_n^2 - 2b_1 Y_n - 2b_2 X_n Y_n + 2b_1 b_2 X_n = \\ &= \sum_{i=1}^n Y_i^2 + n b_1^2 + b_2^2 \sum_{i=1}^n X_i^2 - 2b_1 \sum_{i=1}^n Y_i - 2b_2 \sum_{i=1}^n X_i Y_i + 2b_1 b_2 \sum_{i=1}^n X_i \end{aligned}$$

Note that RSS is effectively a quadratic expression in  $b_1$  and  $b_2$ , with numerical coefficients determined by the data on  $X$  and  $Y$  in the sample.

We can influence the size of RSS only through our choice of  $b_1$  and  $b_2$ .

The data on  $X$  and  $Y$ , which determine the locations of the observations in the scatter diagram, are fixed once we have taken the sample.



The first order conditions for a minimum,

$$\frac{\partial RSS}{\partial b_1} = 0 \quad \text{and} \quad \frac{\partial RSS}{\partial b_2} = 0$$

yield the following equations:

$$2b_2 \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i Y_i + 2b_1 \sum_{i=1}^n X_i = 0 \qquad 2nb_1 - 2 \sum_{i=1}^n Y_i + 2b_2 \sum_{i=1}^n X_i = 0$$

These equations are known as the normal equations for the regression coefficients

$$2nb_1 - 2 \sum_{i=1}^n Y_i + 2b_2 \sum_{i=1}^n X_i = 0 \qquad 2b_2 \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i Y_i + 2b_1 \sum_{i=1}^n X_i = 0$$

Left Equation allows us to write  $b_1$  in terms of  $\bar{Y}$ ,  $\bar{X}$ , and the as yet unknown  $b_2$

Noting that  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  and  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$

may be rewritten

$$2nb_1 - 2n\bar{Y} + 2b_2 n\bar{X} = 0$$

and hence  $2b_2 \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i Y_i + 2(\bar{Y} - b_2 \bar{X})n\bar{X} = 0$

Substituting for  $b_1$  in (5), and again noting that

we obtain

$$b_1 = \bar{Y} - b_2 \bar{X}$$

$$(5) \quad \sum_{i=1}^n X_i = n\bar{X}$$

Separating the terms involving  $b_2$  and not involving  $b_2$  on opposite sides of the equation, we have

$$2b_2 \left[ \left( \sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right] = 2 \sum_{i=1}^n X_i Y_i - 2n\bar{X}\bar{Y}$$

Dividing both sides by  $2n$  we have

$$\left[ \frac{1}{n} \left( \sum_{i=1}^n X_i^2 \right) - \bar{X}^2 \right] b_2 = \frac{1}{n} \left( \sum_{i=1}^n X_i Y_i \right) - \bar{X}\bar{Y}$$

Using the alternative expressions for sample variance and covariance, this may be rewritten

$$b_2 \text{Var}(X) = \text{Cov}(X, Y)$$

and so

$$b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (6)$$

Having found  $b_2$  from (6), you find  $b_1$  from (5).

Home Task: Use the second order conditions to confirm that we have minimized  $RSS$ .

We have found that RSS is minimized when

$$b_2 = \frac{\frac{1}{n} \left( \sum_{i=1}^n X_i Y_i \right) - \bar{X} \bar{Y}}{\left[ \frac{1}{n} \left( \sum_{i=1}^n X_i^2 \right) - \bar{X}^2 \right]}$$

$$\sum_{i=1}^n X_i = n\bar{X}$$

$$b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$b_1 = \bar{Y} - b_2 \bar{X}$$

### 4.3. Interpretation of a Linear Regression Equation

A foolproof way of interpreting the coefficients of a linear regression

$$\hat{Y}_i = b_1 + b_2 X_i$$

when Y and X are variables with straightforward natural units (not logarithms or other functions).

The first step is to say that a one-unit increase in X (measured in units of X) will cause a  $b_2$  unit increase in Y (measured in units of Y).

The second step is to check to see what the units of X and Y actually are, and to replace the word "unit" with the actual unit of measurement.

The third step is to see whether the result could be expressed in a better way, without altering its substance.

The constant,  $b_1$ , gives the predicted value of Y (in units of Y) for X equal to 0.

It may or may not have a plausible meaning, depending on the context.

## 4.4. Goodness of Fit: $R^2$

### Three Useful Results Relating to OLS Regressions

*Proof of (1)*

so

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i$$

$$\sum_{i=1}^n e_i = \sum_{i=1}^n Y_i - nb_1 - b_2 \sum_{i=1}^n X_i$$

Dividing by  $n$ ,

$$\begin{aligned} \bar{e} &= \bar{Y} - b_1 - b_2 \bar{X} \\ &= \bar{Y} - (\bar{Y} - b_2 \bar{X}) - b_2 \bar{X} = 0 \end{aligned}$$

*Proof of (2)*

$$e_i = Y_i - \hat{Y}_i$$

so

$$\sum_{i=1}^n e_i = \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i$$

Dividing by  $n$ ,

$$\bar{e} = \bar{Y} - \bar{\hat{Y}}$$

But  $\bar{e} = 0$ , so  $\bar{\hat{Y}} = \bar{Y}$ .

$$\bar{e}_i$$

*Proof of (3)*

$$\text{Cov}(\hat{Y}, e) = \text{Cov}([b_1 + b_2 X], e) = \text{Cov}(b_1, e) + \text{Cov}(b_2 X, e)$$

$$= b_2 \text{Cov}(X, e) = b_2 \text{Cov}(X, [Y - b_1 - b_2 X])$$

$$= b_2 [\text{Cov}(X, Y) - \text{Cov}(X, b_1) - b_2 \text{Var}(X)]$$

$$= b_2 [\text{Cov}(X, Y) - b_2 \text{Var}(X)]$$

$$= b_2 \left[ \text{Cov}(X, Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \text{Var}(X) \right] = 0$$

We can split the value of  $Y_i$  in each observation into two components,

$$\hat{Y}_i \quad e_i,$$

after running a regression.

$$Y_i = \hat{Y}_i + e_i$$

We can use this to decompose the variance of  $Y$ :

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(\hat{Y} + e) \\ &= \text{Var}(\hat{Y}) + \text{Var}(e) + 2\text{Cov}(\hat{Y}, e) \end{aligned}$$

According to proof of (3)

$$\text{Cov}(\hat{Y}, e) \text{ must be equal to } 0$$

Hence we obtain

$$\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e) \quad (4)$$

This means that we can decompose the variance of  $Y$  into two parts,

$$\text{Var}(\hat{Y}), \quad \text{and} \quad \text{Var}(e),$$

the part "explained" by the regression line

$$\text{Var}(\hat{Y}),$$

the "unexplained" part

$$\text{Var}(e),$$

In view of (4)

$$\text{Var}(\hat{Y})/\text{Var}(Y)$$

*is the proportion of the variance explained by the regression line.*

This proportion is known as the coefficient of determination or, more usually,  $R^2$ :

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}$$

The maximum value of  $R^2$  is 1.

*This occurs when the regression line fits the observations exactly, so that*

$$\hat{Y}_i = Y_i$$

in all observations and all the residuals are 0.

Then  $\text{Var}(\hat{Y}) = \text{Var}(Y)$ ,  $\text{Var}(e)$  is 0,

and one has a perfect fit.

If there is no apparent relationship between the values of  $Y$  and  $X$  in the sample,  $R^2$  will be close to 0.

Often it is convenient to decompose the variance as "sums of squares".



From (4) one has

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2 + \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2$$

multiplying through by  $n$  and using *Proof of (2)*

$$\bar{e} = 0 \text{ and } \bar{\hat{Y}} = \bar{Y}$$

and so

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

Thus

$$TSS = ESS + RSS$$

$$TSS = ESS + RSS$$

where

$\sum_{i=1}^n (Y_i - \bar{Y})^2$  - TSS, the total sum of squares, is given by the left side of the equation

$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$  - ESS, the explained sum of squares,

$\sum_{i=1}^n e_i^2$  - RSS, the residual sum of squares.

ESS and RSS are the two terms on the right side.

One would like  $R^2$  to be as high as possible.

*In particular, we would like the coefficients  $b_1$  and  $b_2$  to be chosen in such a way as to maximize  $R^2$ .*

Does this conflict with our criterion that  $b_1$  and  $b_2$  should be chosen to minimize the sum of the squares of the residuals?

In view of (4) we can rewrite  $R^2$  as

$$R^2 = 1 - \frac{\text{Var}(e)}{\text{Var}(Y)}$$

Thus

$$R^2 = 1 - \frac{\frac{1}{n} \sum_{i=1}^n e_i^2}{\text{Var}(Y)} = 1 - \frac{\frac{1}{n} \text{RSS}}{\text{Var}(Y)}$$

and so the values of  $b_1$  and  $b_2$  that minimize the residual sum of squares automatically maximize  $R^2$ .

## 4. SIMPLE REGRESSION ANALYSIS

Questions:

- 4.1. The Simple Linear Model;
- 4.2. Least Squares Regression;
- 4.3. Interpretation of a Regression Equation;
- 4.4. Goodness of Fit:  $R^2$ .
- 4.5. The F-Test of Goodness of Fit;
- 4.6. The Random Components of the Regression Coefficients

### 4.5. The F-Test of Goodness of Fit

Even if there is no relationship between  $Y$  and  $X$ , in any given sample of observations there may appear to be one, if only a faint one.

Only by coincidence will the sample covariance be exactly equal to 0.

Accordingly, only by coincidence will the correlation coefficient and  $R^2$  be exactly equal to 0.

Suppose that the regression model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

We take as our null hypothesis that there is no relationship between  $Y$  and  $X$ , *that is*,

$$H_0: \beta_2 = 0.$$

We calculate the value that would be exceeded by  $R^2$  as a matter of chance, 5 percent of the time.

We then take this figure as the critical level of  $R^2$  *for a 5 percent significance test*.

If it is exceeded, we reject the null hypothesis in favor of

$$H_1: \beta_2 \neq 0.$$

Suppose that, as in this case, you can decompose the variance of the dependent variable into "explained" and "unexplained" components using (formula (4) previous lecture)

$$\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e)$$

Using the definition of sample variance, and multiplying through by  $n$ , we can rewrite the decomposition as

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

(Remember that  $e$  is 0 and that the sample mean of  $\hat{Y}$  is equal to the sample mean of  $Y$ .)

The left side is TSS, the total sum of squares of the values of the dependent variable about its sample mean.

The first term on the right side is ESS, the explained sum of squares, and the second term is RSS, the unexplained, residual sum of squares:

$$TSS = ESS + RSS$$

The F statistic for the goodness of fit of a regression is written as the explained sum of squares, per explanatory variable, divided by the residual sum of squares, per degree of freedom remaining:

$$F = \frac{ESS / (k - 1)}{RSS / (n - k)}$$

where k is the number of parameters in the regression equation (intercept and k – 1 slope coefficients).

By dividing both the numerator and the denominator of the ratio by TSS, this F statistic may equivalently be expressed in terms of  $R^2$ :

$$F = \frac{(ESS / TSS) / (k - 1)}{(RSS / TSS) / (n - k)} = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

In the present context, k is 2,

$$F = \frac{R^2}{(1 - R^2) / (n - 2)}$$

Having calculated  $F$  from your value of  $R^2$ , you look up  $F_{\text{crit}}$ , the critical level of  $F$ , in the appropriate table.

**If**

$F$  is greater than  $F_{\text{crit}}$ ,

you conclude that the "explanation" of  $Y$  is better than is likely to have arisen by chance.

## 4.6. The Random Components of the Regression Coefficients

A least squares regression coefficient is a special form of random variable whose properties depend on those of the disturbance term in the equation.

Suppose that  $Y$  depends on  $X$  according to the relationship

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

and we are fitting the regression equation given a sample of  $n$  observations

$$\hat{Y}_i = b_1 + b_2 X_i$$

We shall also continue to assume that  $X$  is a *nonstochastic* exogenous variable.

Its value in each observation may be considered to be predetermined by factors unconnected with the present relationship.

Note that  $Y_i$  has two components.

It has a nonrandom component  $(\beta_1 + \beta_2 X_i)$ , which owes nothing to the laws of chance ( $\beta_1$  and  $\beta_2$  may be unknown, but they are fixed constants), and it has the random component  $u_i$ .



We can calculate  $b_2$  according to the usual formula

$$b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$b_2$  also has a random component.

$\text{Cov}(X, Y)$  depends on the values of  $Y$ , and the values of  $Y$  depend on the values of  $u$ .

If the values of the disturbance term had been different in the  $n$  observations, we would have obtained different values of  $Y$ ,

hence different values of  $\text{Cov}(X, Y)$ , and hence different values of  $b_2$ .

We can in theory decompose  $b_2$  into its nonrandom and random components

$$\text{Cov}(X, Y) = \text{Cov}(X, [\beta_1 + \beta_2 X + u])$$

$$= \text{Cov}(X, \beta_1) + \text{Cov}(X, \beta_2 X) + \text{Cov}(X, u)$$

$$\beta_1 = \text{const and } \beta_2 = \text{const,}$$

By Covariance Rule,  $\text{Cov}(X, \beta_1)$  must be equal to 0

$\text{Cov}(X, \beta_2 X)$  is equal to  $\beta_2 \text{Cov}(X, X)$ .

$\text{Cov}(X, X)$  is the same as  $\text{Var}(X)$ .

Hence we can write

$$\text{Cov}(X, Y) = \beta_2 \text{Var}(X) + \text{Cov}(X, u)$$

and so

$$b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)} \quad (*)$$

Thus we have shown that the regression coefficient  $b_2$  obtained from any sample consists of

- (1) a fixed component, equal to the true value,  $\beta_2$ , and
- (2) a random component dependent on  $\text{Cov}(X, u)$ , which is responsible for its variations around this central tendency.

One may easily show that  $b_1$  has a fixed component equal to the true value,  $\beta_1$ , plus a random component that depends on the random factor  $u$ .

Note that you are not able to make these decompositions in practice because you do not know the true values of  $\beta_1$  and  $\beta_2$  or the actual values of  $u$  in the sample.

We are interested in them because they enable us to say something about the theoretical properties of  $b_1$  and  $b_2$ , given certain assumptions.

## 5. The Gauss–Markov Theorem

We shall continue to work with the simple regression model where Y depends on X according to the relationship

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

and we are fitting the regression equation given a sample of n observations

$$\hat{Y}_i = b_1 + b_2 X_i$$

The properties of the regression coefficients depend critically on the properties of the disturbance term.

Indeed the latter has to satisfy four conditions, known as the Gauss–Markov conditions, if ordinary least squares regression analysis is to give the best possible results.

**Gauss–Markov Condition 1:  $E(u_i) = 0$  for All Observations**

The first condition is that the expected value of the disturbance term in any observation should be 0.

Sometimes it will be positive, sometimes negative, but it should not have a systematic tendency in either direction.

**Gauss–Markov Condition 2: Population Variance of  $u_i$  Constant for All Observations**

$$\sigma_{u_i}^2 = \sigma_u^2 \text{ for all } i$$

The second condition is that the population variance of the disturbance term should be constant for all observations.

Sometimes the disturbance term will be greater, sometimes smaller, but there should not be any a priori reason for it to be more irregular in some observations than in others.

The constant is usually denoted  $\sigma_u^2$

Since  $E(u_i)$  is 0, the population variance of  $u_i$  is equal to  $E(u_i^2)$ ,

so the condition can also be written

$$E(u_i^2) = \sigma_u^2 \text{ for all } i$$

$\sigma_u$  of course, is unknown.

One of the tasks of regression analysis is to estimate the standard deviation of the disturbance term.

**Gauss–Markov Condition 3:**

**$u_i$  Distributed Independently of  $u_j$  ( $i \neq j$ )**

This condition states that there should be no systematic association between the values of the disturbance term in any two observations.

The condition implies that  $\sigma_{u_i u_j}$  the population covariance between  $u_i$  and  $u_j$ , is 0, because

$$\sigma_{u_i u_j} = E[(u_i - \mu_u)(u_j - \mu_u)] = E(u_i u_j) = E(u_i)E(u_j) = 0$$

**Gauss–Markov Condition 4:  $u$  Distributed Independently of the Explanatory Variables**

The population covariance between the explanatory variable and the disturbance term is 0.

Since  $E(u_i)$  is 0, and the term involving  $X$  is nonstochastic

$$\sigma_{X_i u_i} = E[\{X_i - E(X_i)\}\{u_i - \mu_u\}] = (X_i - X_i) E(u_i) = 0$$

## **The Normality Assumption**

In addition to the Gauss–Markov conditions, one usually assumes that the disturbance term is normally distributed.

## 6. Unbiasedness of the Regression Coefficients

$$b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)} \quad (*)$$

From (\*) we can show that  $b_2$  must be an unbiased estimator of  $\beta_2$  if the fourth Gauss–Markov condition is satisfied:

$$E(b_2) = E\left[\beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}\right] = \beta_2 + E\left[\frac{\text{Cov}(X, u)}{\text{Var}(X)}\right]$$

If we adopt the strong version of the fourth Gauss–Markov condition and assume that  $X$  is nonrandom, we may also take  $\text{Var}(X)$  as a given constant, and so

$$E(b_2) = \beta_2 + \frac{1}{\text{Var}(X)} E[\text{Cov}(X, u)]$$



We will demonstrate that  $E[\text{Cov}(X, u)]$  is 0:

$$\begin{aligned}
 E[\text{Cov}(X, u)] &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})\right] = \\
 &= \frac{1}{n} (E[(X_1 - \bar{X})(u_1 - \bar{u})] + \dots + E[(X_n - \bar{X})(u_n - \bar{u})]) = \\
 &= \frac{1}{n} \sum_{i=1}^n E[(X_i - \bar{X})(u_i - \bar{u})] = \\
 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) E(u_i - \bar{u}) = 0
 \end{aligned}$$

In the second line on the previous slide, the *second expected value rule*<sup>2</sup> has been used to bring out  $(1/n)$  of the expression as a common factor, and the *first rule*<sup>1</sup> has been used to break up the expectation of the sum into the sum of the expectations.

---

<sup>1</sup>  $E(X + Y) = E(X) + E(Y)$

<sup>2</sup>  $E(cX) = cE(X)$

In the third line, the term involving  $X$  has been brought out because  $X$  is nonstochastic.

By virtue of the first Gauss–Markov condition,  $E(u_i)$  is 0, and hence  $E(\bar{u})$  is also 0.

Therefore  $E[\text{Cov}(X, u)]$  is 0 and  $E(b_2) = \beta_2$

In other words,  $b_2$  is an unbiased estimator of  $\beta_2$ .

One may easily show that  $b_1$  is an unbiased estimator of  $\beta_1$ .

## 7. Precision of the Regression Coefficients

Now we shall consider  $\sigma_{b_1}^2$  and  $\sigma_{b_2}^2$   
the population variances of  $b_1$  and  $b_2$  about  
their population means.

These are given by the following expressions  
(Thomas, 1983)

$$\sigma_{b_1}^2 = \frac{\sigma_u^2}{n} \left[ 1 + \frac{\bar{X}^2}{\text{Var}(X)} \right] \text{ and } \sigma_{b_2}^2 = \frac{\sigma_u^2}{n \text{Var}(X)}$$

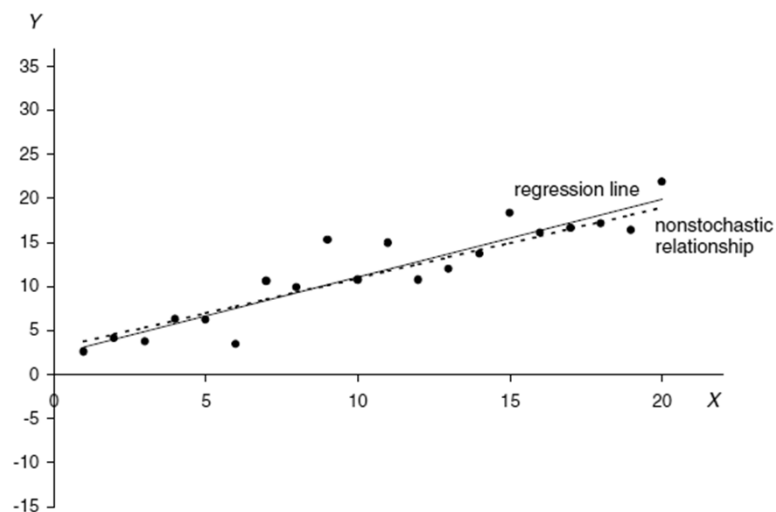


Figure 3.3a. Disturbance term with relatively small variance

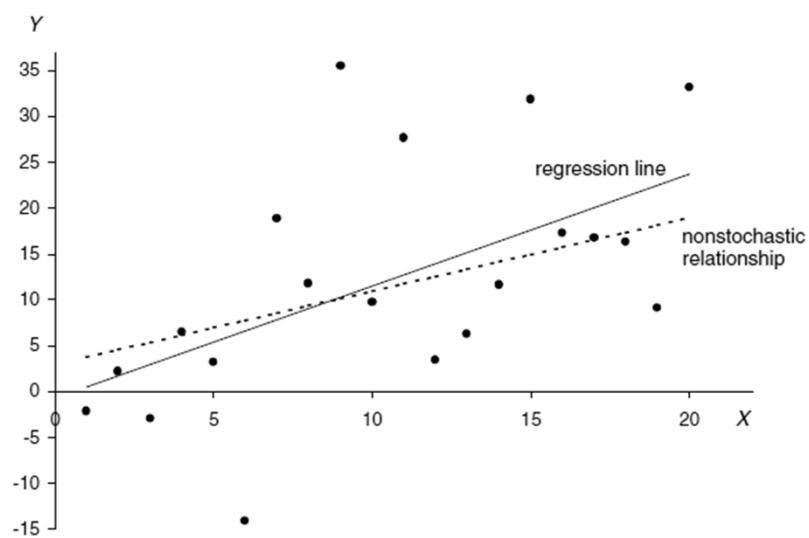


Figure 3.3b. Disturbance term with relatively large variance

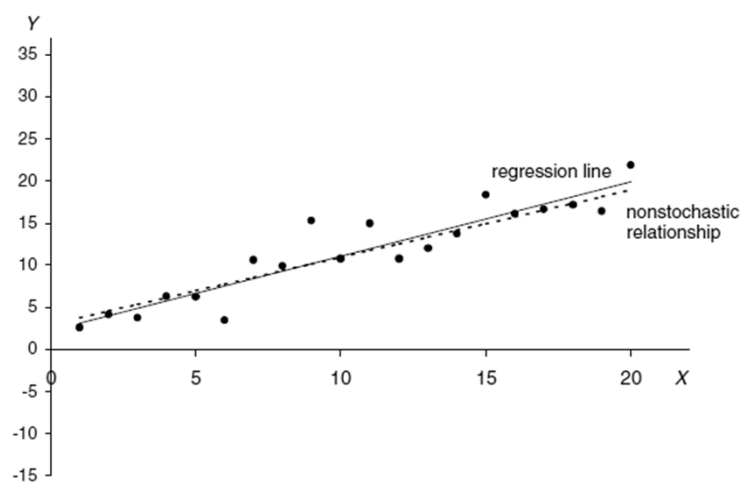


Figure 3.4a.  $X$  with relatively large variance

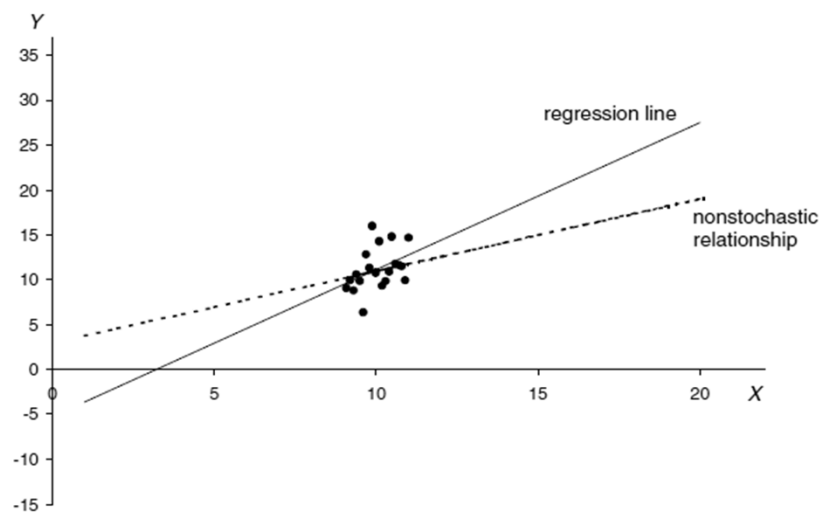


Figure 3.4b.  $X$  with relatively small variance

The standard errors of the regressions coefficient will be calculated

$$\text{s.e.}(b_1) = \sqrt{\frac{s_u^2}{n} \left[ 1 + \frac{\bar{X}^2}{\text{Var}(X)} \right]} \quad \text{and} \quad \text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{n \text{Var}(X)}}$$

(\*)

The higher the variance of the disturbance term, the higher the sample variance of the residuals is likely to be,

and hence the higher will be the standard errors of the coefficients in the regression equation, reflecting the risk that the coefficients are inaccurate.

However, it is only a risk. It is possible that in any particular sample the effects of the disturbance term in the different observations will cancel each other out and the regression coefficients will be accurate after all.

## 8. Testing Hypotheses Relating to the Regression Coefficients

Suppose you have a theoretical relationship

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

and your null and alternative hypotheses are

$$H_0: \beta_2 = \beta_2^0, \quad H_1: \beta_2 \neq \beta_2^0.$$

We have assumed that the standard deviation of  $b_2$  is known, which is most unlikely in practice.

It has to be estimated by the standard error of  $b_2$ , given by (\*).

This causes two modifications to the test procedure.

---

(\*)

$$b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}$$

First,  $z$  is now defined using  $\text{s.e.}(b_2)$  instead of  $\text{s.d.}(b_2)$ , and it is referred to as the  $t$  statistic

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)}$$

Second, the critical levels of  $t$  depend upon what is known as a  $t$ -distribution

The critical value of  $t$  denote as  $t_{\text{crit}}$

The condition that a regression estimate should not lead to the rejection of a null hypothesis

$$H_0: \beta_2 = \beta_2^0$$

is

$$-t_{\text{crit}} \leq \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} \leq t_{\text{crit}}$$



Hence we have the decision rule: reject  $H_0$

$$\text{if } \left| \frac{b_2 - \beta_2^0}{s.e.(b_2)} \right| > t_{\text{crit}}$$

do not reject if

$$\left| \frac{b_2 - \beta_2^0}{s.e.(b_2)} \right| \leq t_{\text{crit}}$$

Where  $\left| \frac{b_2 - \beta_2^0}{s.e.(b_2)} \right|$  is the absolute value (numerical value, neglecting the sign) of  $t$ .

## 9. Confidence Intervals

We have shown that

$$-t_{\text{crit}} \leq \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} \leq t_{\text{crit}}$$

we can see that regression coefficient  $b_2$   
and hypothetical value  $\beta_2$  are  
incompatible if either

$$\frac{b_2 - \beta_2}{\text{s.e.}(b_2)} > t_{\text{crit}} \quad \text{or} \quad \frac{b_2 - \beta_2}{\text{s.e.}(b_2)} < -t_{\text{crit}}$$

that is, if either

$$b_2 - \beta_2 > \text{s.e.}(b_2) \times t_{\text{crit}} \quad \text{or} \quad b_2 - \beta_2 < -\text{s.e.}(b_2) \times t_{\text{crit}}$$

that is, if either

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} > \beta_2 \quad \text{or} \quad b_2 + \text{s.e.}(b_2) \times t_{\text{crit}} < \beta_2$$

It therefore follows that a hypothetical  $\beta_2$  is compatible with the regression result if both

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} \leq \beta_2 \quad \text{or} \quad b_2 + \text{s.e.}(b_2) \times t_{\text{crit}} \geq \beta_2$$

that is, if  $\beta_2$  satisfies the double inequality

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} \leq \beta_2 \leq b_2 + \text{s.e.}(b_2) \times t_{\text{crit}}$$

Any hypothetical value of  $\beta_2$  that satisfies

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} \leq \beta_2 \leq b_2 + \text{s.e.}(b_2) \times t_{\text{crit}}$$

will therefore automatically be compatible with the estimate  $b_2$ , that is, will not be rejected by it.

The set of all such values, given by the interval between the lower and upper limits of the inequality, is known as the **confidence interval** for  $\beta_2$

Note that the center of the confidence interval is  $b_2$  itself.

The limits are equidistant on either side.

Note also that, since the value of  $t_{\text{crit}}$  depends upon the choice of significance level, the limits will also depend on this choice.

If the 5 percent significance level is adopted, the corresponding confidence interval is known as the 95 percent confidence interval.

If the 1 percent level is chosen, one obtains the 99 percent confidence interval, and so on.

# **HETEROSCEDASTICITY**

**and**

# **AUTOCORRELATION**



## Heteroscedasticity and Its Implications



- The second of the Gauss–Markov conditions states that the variance of the disturbance term in each observation should be constant.
- The disturbance term in each observation has only one value, so what can be meant *by its* "variance"?




## Heteroscedasticity and Its Implications



- What we are talking about is its potential behavior before the sample is generated.
- *When we write the model*

$$Y = \beta_1 + \beta_2 X + u,$$

the first two Gauss–Markov conditions state that the disturbance terms  $u_1, \dots, u_n$  in the  $n$  observations are drawn from probability distributions that have 0 mean and the same variance.



## Heteroscedasticity and Its Implications

Their *actual values* in the sample will sometimes be positive, sometimes negative, sometimes relatively far from 0, sometimes relatively close, but there will be no a priori reason to anticipate a particularly erratic value in any given observation.

To put it another way, the probability of  $u$  reaching a given positive (or negative) value will be the same in all observations.

This condition is known as homoscedasticity, which means "same dispersion".

## Heteroscedasticity and Its Implications

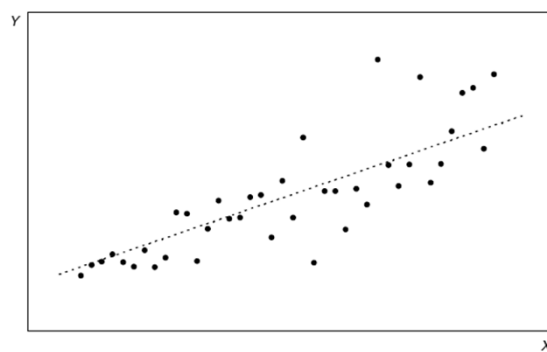


Figure 8.3 Model with a heteroscedastic disturbance term

Figure 8.3 illustrates how a typical scatter diagram would look if  $Y$  were an increasing function of  $X$  and the heteroscedasticity were present



## Heteroscedasticity and Its Implications



Why does heteroscedasticity matter?

If heteroscedasticity is present

1. the OLS estimators are inefficient.
2. the standard errors of the regression coefficients will be wrong




## Heteroscedasticity and Its Implications



It is quite likely that the standard errors will be underestimated, so the **t** statistics will be overestimated and you will have a misleading impression of the precision of your regression coefficients.

You may be led to believe that a coefficient is significantly different from 0, at a given significance level, when in fact it is not.



# Possible Causes of Heteroscedasticity




## Possible Causes of Heteroscedasticity



If the true relationship is given by

$$Y = \beta_1 + \beta_2 X + u,$$

it may well be the case that the variations in the omitted variables and the measurement errors that are jointly responsible for the disturbance term will be relatively small when *Y and X are small and large when they are large*, economic variables tending to move in size together.





# Detection of Heteroscedasticity: The Goldfeld–Quandt Test




## Detection of Heteroscedasticity: The Goldfeld–Quandt Test




The most common formal test for heteroscedasticity is that of Goldfeld and Quandt (1965).

It assumes that  $\sigma_{ui}$ , *the standard deviation of the probability distribution of the disturbance term in observation  $i$ , is proportional to the size of  $X_i$ .* It also assumes that the disturbance term is normally distributed and satisfies the other Gauss–Markov conditions.




### **Detection of Heteroscedasticity: The Goldfeld–Quandt Test**

The  $n$  observations in the sample are ordered by the magnitude of  $X$  and separate regressions are run for the first  $n'$  and for the last  $n'$  observations, the middle  $(n - 2n')$  observations being dropped entirely.

If heteroscedasticity is present, and if the assumption concerning its nature is true, the variance of  $u$  in the last  $n'$  observations will be greater than that in the first  $n'$ , and this will be reflected in the residual sums of squares in the two subregressions. 

### **Detection of Heteroscedasticity: The Goldfeld–Quandt Test**

Denoting these by  $RSS_1$  and  $RSS_2$  for the subregressions with the first  $n'$  and the last  $n'$  observations, respectively, the ratio  $RSS_2/RSS_1$  will be distributed as an  $F$  statistic with  $(n' - k)$  and  $(n' - k)$  degrees of freedom, where  $k$  is the number of parameters in the equation, under the null hypothesis of homoscedasticity. 



## Detection of Heteroscedasticity: The Goldfeld–Quandt Test



The power of the test depends on the choice of  $n'$  in relation to  $n$ .

As a result of some experiments undertaken by them, Goldfeld and Quandt suggest that  $n'$  should be about 11 when  $n$  is 30 and about 22 when  $n$  is 60, suggesting that  $n'$  should be about three-eighths of  $n$ .

If there is more than one explanatory variable in the model, the observations should be ordered by that which is hypothesized to be associated with  $\sigma_i$ .



## Detection of Heteroscedasticity: The Goldfeld–Quandt Test



The null hypothesis for the test is that  $RSS_2$  is not significantly greater than  $RSS_1$ , and the alternative hypothesis is that it is significantly greater.

If  $RSS_2$  turns out to be smaller than  $RSS_1$ , you are not going to reject the null hypothesis and there is no point in computing the test statistic  $RSS_2/RSS_1$ .






## Detection of Heteroscedasticity: The Goldfeld–Quandt Test



However, the Goldfeld–Quandt test can also be used for the case where the standard deviation of the disturbance term is hypothesized to be inversely proportional to  $X_i$ .

*The procedure is the same as before,  
but the test statistic is now  $RSS_1/RSS_2$ ,  
and it will again be distributed as an  $F$  statistic  
with  $(n' - k)$  and  $(n' - k)$  degrees of freedom  
under the null hypothesis of homoscedasticity.*



# What Can You Do about Heteroscedasticity

## What Can You Do about Heteroscedasticity

Suppose that the true relationship is

- $$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Let the standard deviation of the disturbance term in observation  $i$  be  $\sigma_{ui}$

If you happened to know  $\sigma_{ui}$  for each observation,

you could eliminate the heteroscedasticity by dividing each observation by its value of  $\sigma$ .

## What Can You Do about Heteroscedasticity

The model becomes

$$\frac{Y_i}{\sigma_{u_i}} = \beta_1 \frac{1}{\sigma_{u_i}} + \beta_2 \frac{X_i}{\sigma_{u_i}} + \frac{u_i}{\sigma_{u_i}}$$

The disturbance term  $u_i / \sigma_{ui}$  is homoscedastic because the population variance of  $\frac{u_i}{\sigma_i}$  is

$$E\left\{\left(\frac{u_i}{\sigma_{u_i}}\right)^2\right\} = \frac{1}{\sigma_{u_i}^2} E(u_i^2) = \frac{1}{\sigma_{u_i}^2} \sigma_{u_i}^2 = 1$$

## What Can You Do about Heteroscedasticity

The model becomes

$$\frac{Y_i}{\sigma_{u_i}} = \beta_1 \frac{1}{\sigma_{u_i}} + \beta_2 \frac{X_i}{\sigma_{u_i}} + \frac{u_i}{\sigma_{u_i}}$$

The disturbance term  $u_i / \sigma_{u_i}$  is homoscedastic because the population variance of  $\frac{u_i}{\sigma_{u_i}}$  is

$$E\left\{\left(\frac{u_i}{\sigma_{u_i}}\right)^2\right\} = \frac{1}{\sigma_{u_i}^2} E(u_i^2) = \frac{1}{\sigma_{u_i}^2} \sigma_{u_i}^2 = 1$$

## What Can You Do about Heteroscedasticity

Therefore, every observation will have a disturbance term drawn from a distribution with population variance 1, and the model will be homoscedastic.

The revised model may be rewritten

$$Y_i' = \beta_1 h_i + \beta_2 X_i' + u_i',$$

where  $Y_i' = Y_i / \sigma_{u_i}$ ,  $X_i' = X_i / \sigma_{u_i}$

$h$  is a new variable whose value in observation  $i$  is  $1/\sigma_{u_i}$  and  $u_i' = u_i / \sigma_{u_i}$




## What Can You Do about Heteroscedasticity



Note that there should not be a constant term in the equation. By regressing  $Y'$  on  $h$  and  $X'$ , you will obtain efficient estimates of  $\beta_1$  and  $\beta_2$  with unbiased standard errors.

In practice it may be a good idea to try several variables for scaling the observations and to compare the results.

If the results are roughly similar each time, and tests fail to reject the null hypothesis of homoscedasticity, your problem should be at an end.



## Autocorrelation

## Possible Causes of Autocorrelation

Autocorrelation normally occurs only in regression analysis using time series data.

The disturbance term in a regression equation picks up the influence of those variables affecting the dependent variable that have not been included in the regression equation.

If the value of  $u$  in *any observation is to be* independent of its value in the previous one, the value of any variable hidden in  $u$  *must be* uncorrelated with its value at the time of the previous observation.

## Possible Causes of Autocorrelation

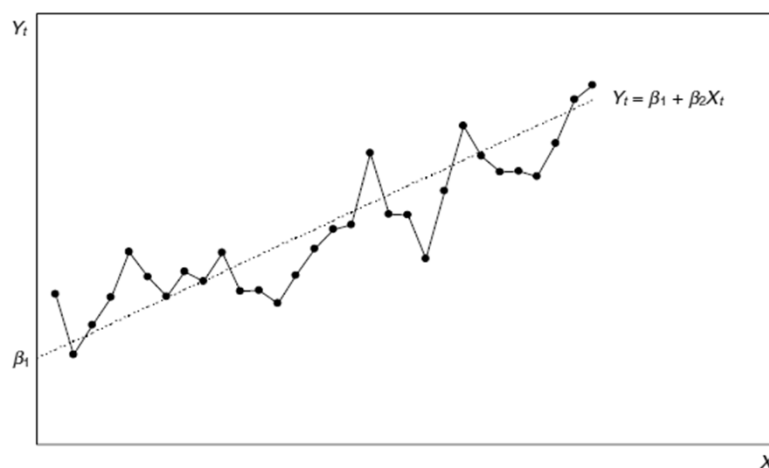
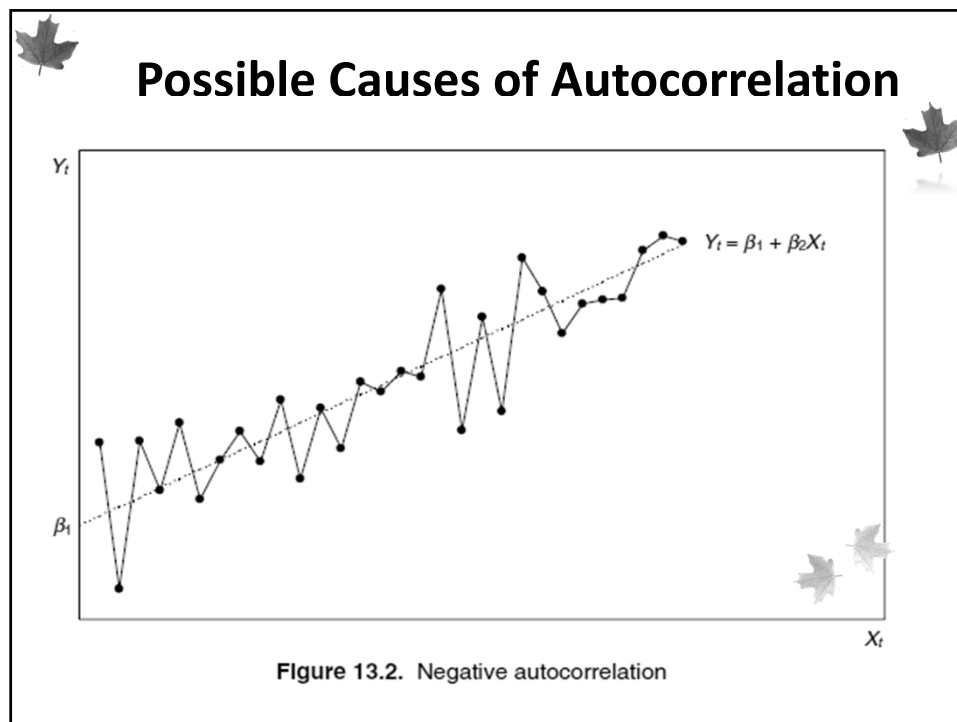


Figure 13.1. Positive autocorrelation





## Detection of First-Order Autocorrelation: the Durbin–Watson Test



## The Durbin–Watson Test

We will mostly be concerned with first-order autoregressive autocorrelation, often denoted AR(1),

where the disturbance term  $u$  in the model

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

is generated by the process

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

where  $\varepsilon_t$  is a random variable whose value in any observation is independent of its value in all the other observations.



## The Durbin–Watson Test

Because AR(1) is such a common form of autocorrelation, the standard test statistic for it, the Durbin–Watson  $d$  statistic, is usually included in the basic diagnostic statistics.

It is calculated from the residuals using the expression

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$






## The Durbin–Watson Test



If there is no autocorrelation present,  
*d should be close to 2.*

If there is positive autocorrelation,  
*d will tend to be less than 2.*

*If there is negative autocorrelation,*  
*it will tend to be greater than 2.*





## The Durbin–Watson Test



The test assumes that *d lies between 4 and 0.*  
The critical value of *d at any significance level depends on the number of explanatory variables in the regression equation and the number of observations in the sample.*  
It is possible to calculate upper and lower *limits for the critical value of d.*

*Those for positive autocorrelation are usually denoted  $d_U$  and  $d_L$ .*

*We know, that  $d_{crit}$  lies somewhere between  $d_L$  and  $d_U$ .*





## The Durbin–Watson Test

This leaves you with three possible outcomes for the test

1.  $d$  is less than  $d_L$ . In this case, it must be lower than  $d_{crit}$  so you would reject the null hypothesis and conclude that positive autocorrelation is present.
2.  $d$  is greater than  $d_U$ . In this case,  $d$  must be greater than  $d_{crit}$ , so you would fail to reject the null hypothesis.
3.  $d$  lies between  $d_L$  and  $d_U$ . In this case,  $d$  might be greater or less than  $d_{crit}$ . You do not know which, so you cannot tell whether you should reject or not reject the null hypothesis.



## The Durbin–Watson Test

In cases (1) and (2), the Durbin–Watson test gives you a definite answer,

but in case (3) you are left in a zone of indecision, and there is nothing that you can do about it.

Level 4 –  $d_U$  gives the lower limit,

below which you fail to reject the null hypothesis of no autocorrelation,

and  $4 - d_L$  gives the upper one, above which you conclude that there is evidence of negative autocorrelation



# What Can You Do about Autocorrelation

## What Can You Do about Autocorrelation

We will consider only the case of AR(1) autocorrelation.

Suppose that the model is

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (1)$$

with  $u_t$  generated by the process

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

If we lag first equation by one time period and multiply by  $\rho$ , we have

$$(2) \quad \rho Y_{t-1} = \beta_1 \rho + \beta_2 \rho X_{t-1} + \rho u_{t-1}$$



## What Can You Do about Autocorrelation



Now subtract (2) from (1):

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2 X_t - \beta_2 \rho X_{t-1} + u_t - \rho u_{t-1}$$

Hence  $Y_t = \beta_1(1-\rho) + \rho Y_{t-1} + \beta_2 X_t - \beta_2 \rho X_{t-1} + \varepsilon_t,$

The model is now free from autocorrelation because the disturbance term has been reduced to the innovation  $\varepsilon_t$ .

