MTMS.01.099 Mathematical Statistics

Lecture 15. Introduction to linear regression

Tõnu Kollo

tonu.kollo@ut.ee

Fall 2016

Introduction to Linear Regression

We have studied two important areas of inferential statistics: **confidence intervals** and **hypothesis testing**. In the last lecture we had a brief overview about possible dependences between 2 variables. Different correlation coefficients could be calculated for different type of variables which measure dependence of variables in different meaning: linear dependence, monotone dependence, functional dependence in general.

Correlation analysis is a statistical method used to determine whether a relationship between variables exists.

Both variables are considered as variable in this case.

Inferential statistics involves also determining whether a random variable depends on one or more numerical or quantitative variables. Then we are interested in dependence of one variable from the other(s), which are considered as fixed.

For example,

- a businessperson may want to know whether the volume of sales for a given month depends on the amount of advertising the firm does that month.
- Educators are interested in determining whether the student's score on a particular exam depends on the number of hours a student studied and how often he/she attended lectures and seminars.
- Medical researchers are interested does blood pressure depends on person's age and sex.

These questions can be answered by using the techniques of **regression analysis**.

Short introduction to Linear Regression (2)

Regression is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or non-linear.
There are two types of relationships: simple and multiple.

- In a simple relationship, there are two variables - an independent variable, also called an **explanatory variable** or a predictor variable, and a dependent variable, also called a **response variable**.
- This simple relationship analysis is called **simple regression**, and there is one independent variable that is used to predict the dependent variable.
  - A manager may wish to see if the number of years the sales people have been working for the company has anything to do with the amount of sales they make. This type of study involves a simple relationship, since there are only two variables - years of experience and amount of sales.

Short introduction to Linear Regression (3)

- In a multiple relationship, called **multiple regression**, two or more independent variables are used to predict one dependent variable.
  - For example, an educator may wish to investigate the relationship between a student's success in university and factors such as the number of hours devoted to studying, the student's grades in gymnasium etc. This type of study involves several variables.

Short introduction to Linear Regression (4)

Simple relationships can be **positive** or **negative**.

- A positive relationship exists when both variables increase or decrease at the same time.
  - For example, a person's height and weight are related; and the relationship is positive, since the taller a person is, generally, the bigger is the person weight.
- In a negative relationship, as one variable increases, the other variable decreases, and vice versa.
  - For example, if you measure the strength of people over 60 years of age, you will find that as age increases, strength generally decreases. The word 'generally' is used here because there are exceptions.

Example

In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables. For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of students, determine the hours each studied, and obtain their grades on the exam. A table can be made for the data, as shown here:

Student	**Hours of study** x	**Grade** y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Example (2)

As stated previously, the two variables for this study are called the independent variable and the dependent variable.

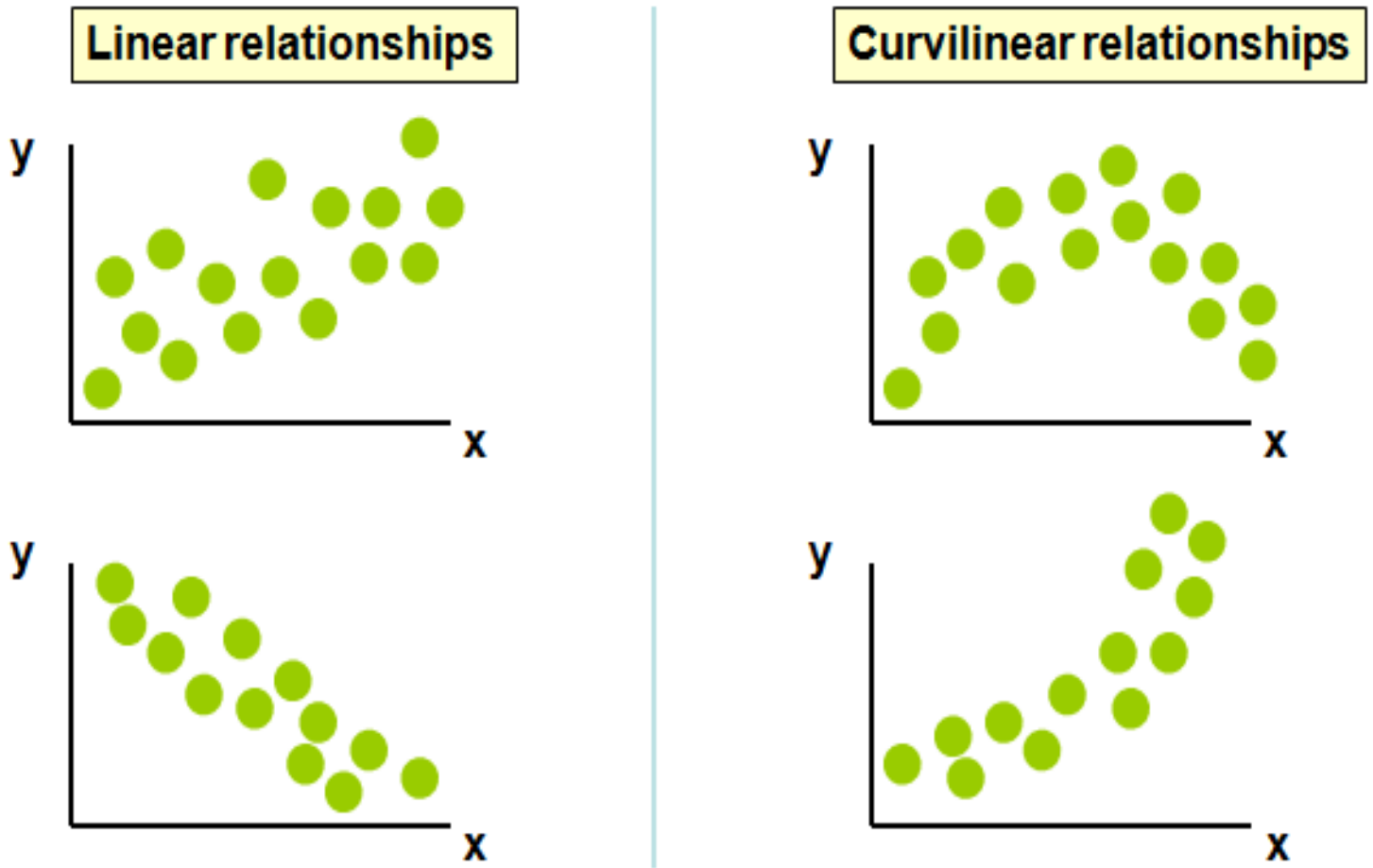
- The **independent variable** is the variable in regression that can be controlled or manipulated. In this case, the number of hours of study is the independent variable and is designated as the X variable.
- The **dependent variable** is the variable in regression that cannot be controlled or manipulated. The grade the student received on the exam is the dependent variable, designated as the Y variable.
- The reason for this distinction between the variables is that you assume that the grade the student earns depends on the number of hours the student studied. Also, you assume that, to some extent, the student can regulate or control the number of hours he or she studies for the exam.

Scatter plot

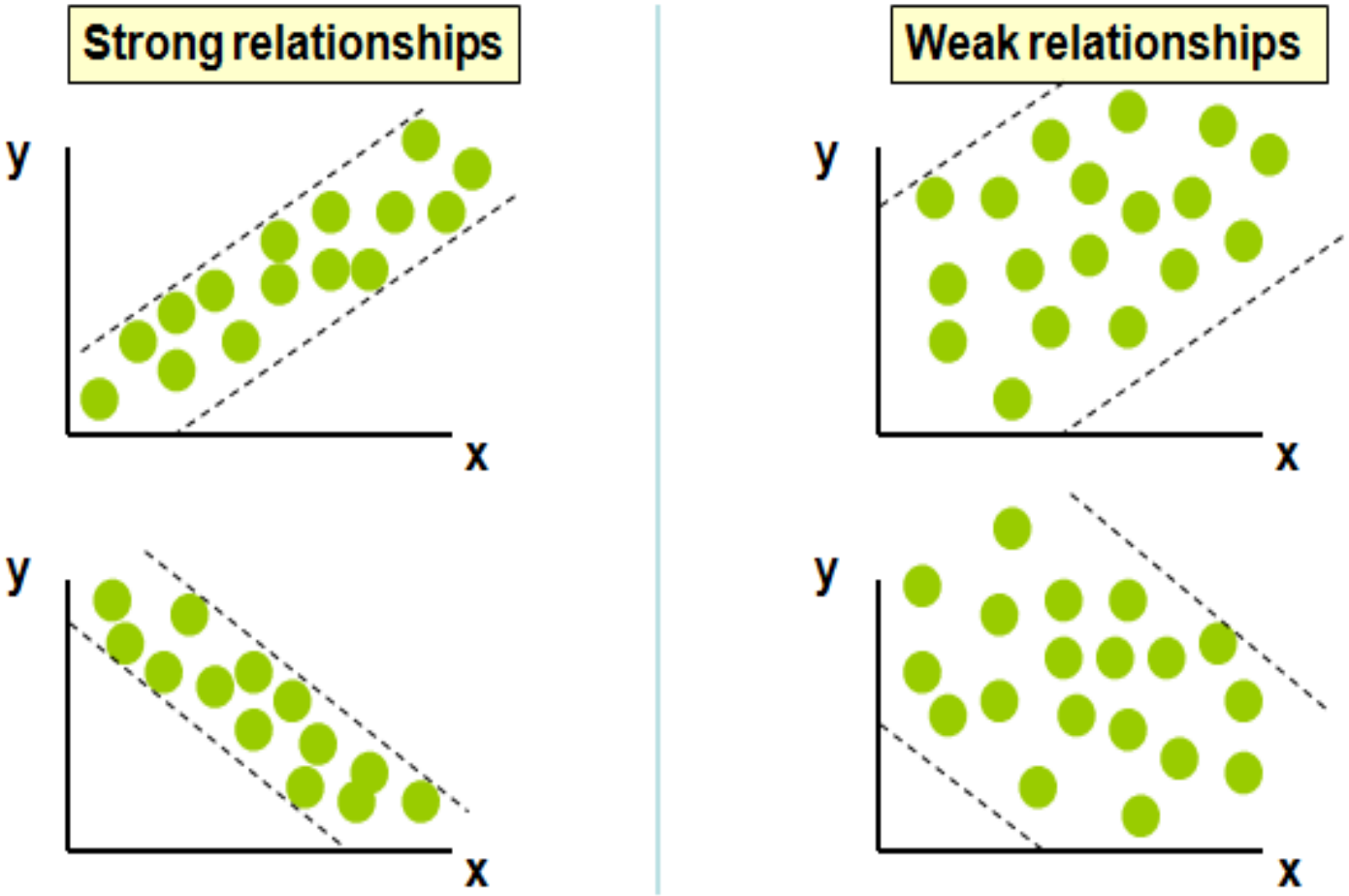
A **scatter plot** (or scatter diagram) is used to show the relationship between the independent and dependent variables.

- The independent variable X is plotted on the horizontal axis, and the dependent variable Y is plotted on the vertical axis.
- The scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables. The scales of the variables can be different, and the coordinates of the axes are determined by the smallest and largest data values of the variables.

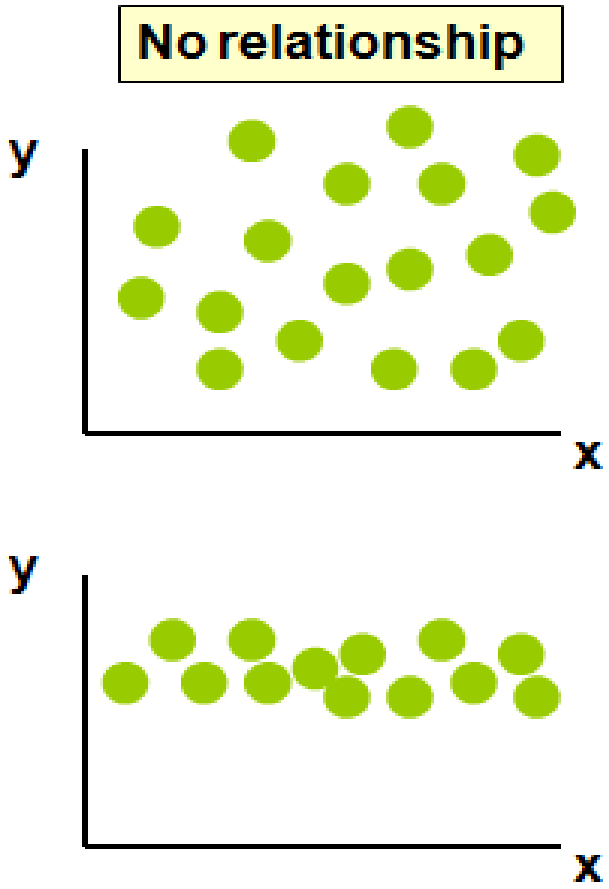
Scatter plot. Relationships graphically



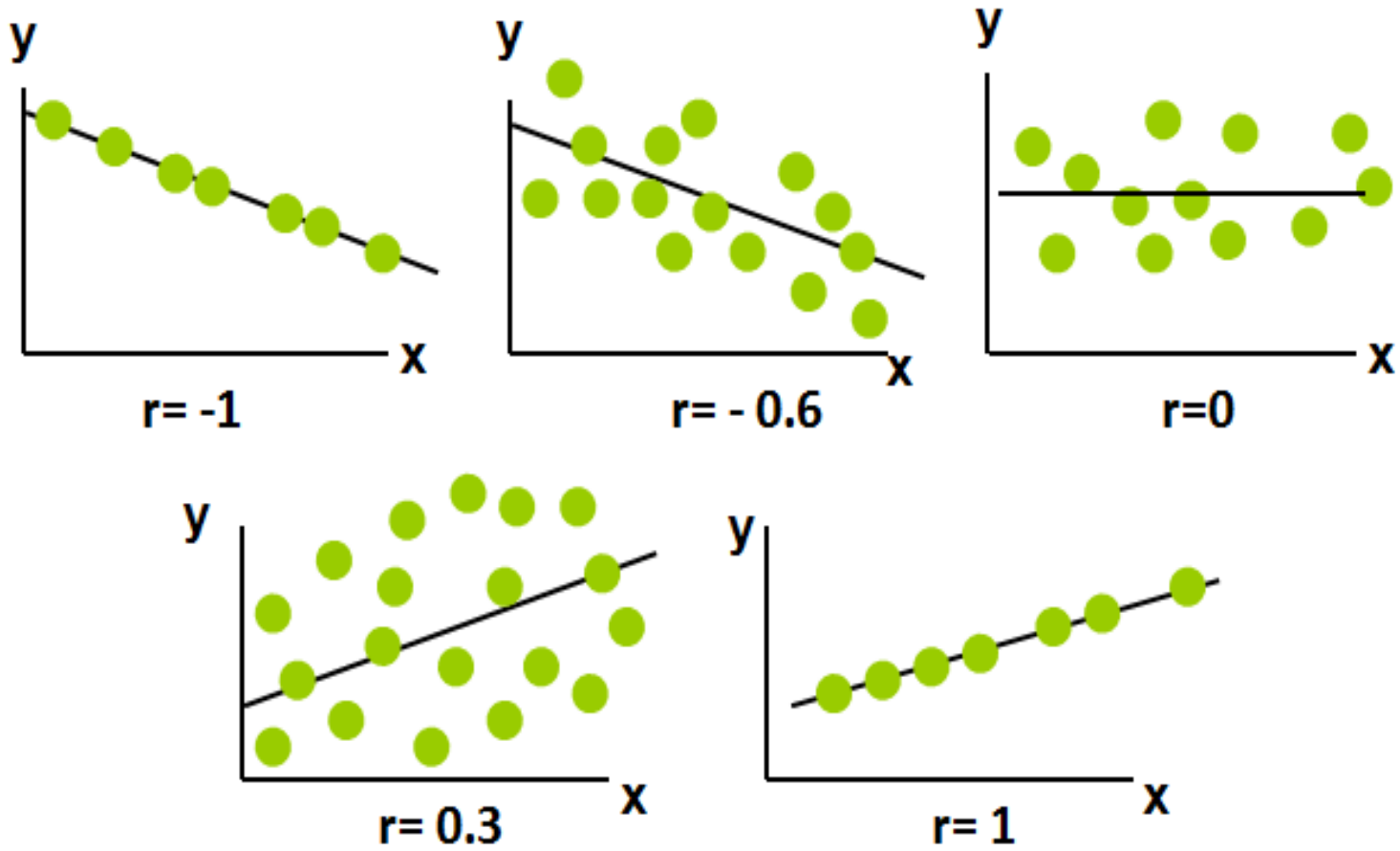
Scatter plot. Relationships graphically (2)



Scatter plot. Relationships graphically (3)



Different correlation coefficients graphically



Assumptions for the Correlation Coefficient

Before starting with regression analysis we first calculate the Linear Correlation Coefficient. The necessary assumptions are the following.

- The sample is a random sample.
- The variables have a joint normal distribution. (This means that given any specific value of X, the Y values are normally distributed; and given any specific value of Y, the X values are normally distributed.)

There are no units associated with $r(X, Y)$, and the value of $r(X, Y)$ will remain unchanged if the X and Y values are switched.

Tests for Correlation

How to test whether our calculated correlation coefficient is different from zero? We have to test

$$H_0 : r = 0,$$

$$H_1 : r \neq 0.$$

Even for the normal population exact distribution of the estimator of the correlation coefficient is complicated. But at least two approximations can be used. Denote an estimator of the $r(X, Y)$ by

$$R(X, Y) = \frac{1/n \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \overline{X})(Y_j - \overline{Y})}{S_X S_Y}$$

where S_X and S_Y are estimators of the standard deviations of X and Y respectively.

Tests for Correlation

For big values of the sample size n

$$T_R = \frac{R(X,Y)\sqrt{n-2}}{\sqrt{1-R(X,Y)^2}} \approx t(n-2)$$

and for testing we apply the usual t-test for a two-sided hypothesis.
Convergence to the t-distribution is not fast and for smaller sample sizes so-called Fisher's z-test is recommended.
Consider the following statistic:

$$Z = \frac{1}{2} ln\Big(\frac{1+R(X,Y)}{1-R(X,Y)}\Big).$$

Then under H_0

$$Z \approx N(0, \frac{1}{n-3})$$

After normalization we can use standard normal ditribution:

$$Z\sqrt{n-3} \approx N(0,1).$$

Steps of Regression Analysis

- In studying relationships between two variables, collect the data and then construct a scatter plot.
- The purpose of the scatter plot is to determine the nature of the relationship.
- After the scatter plot is drawn, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship.
- If the value of the correlation coefficient is significant, the next step is to determine the equation of the regression line, which is the data's line of best fit.
- The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

Regression Analysis (2)

Regression analysis is used to:

1. Predict the value of a dependent variable based on the value of at least one independent variable.
2. Explain the impact of changes in an independent variable on the dependent variable.

Once again – dependent variable is the variable we wish to explain and independent variable is the variable used to explain the dependent variable.

Simple Linear Regression Analysis

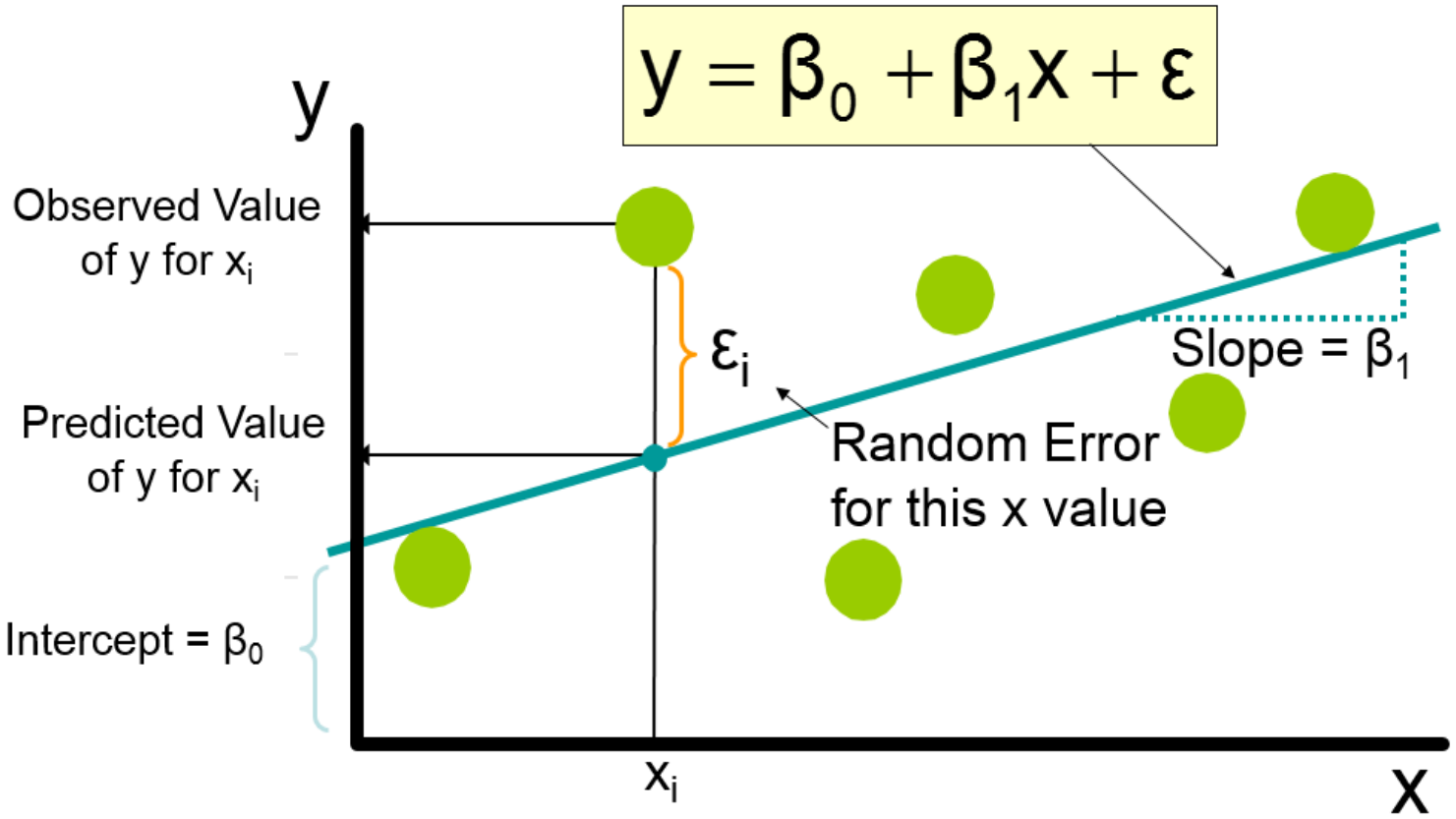
- Only one independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X

Population linear regression mathematically:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where β_0 and β_1 are unknown parameters and ε is random error or so-called residual.

Simple Linear Regression Analysis (2)



Assumptions of Linear Regression

- Error values (ε) are statistically independent.
- Error values are normally distributed for any given value of X.
- The probability distribution of the errors is normal.
- The underlying relationship between the X variable and the Y variable is linear.

Estimated Regression Model

The unknown parameters β_0 and β_1 have to be estimated from the sample data:

$$y_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

Denote

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i).$$

The parameters β_i are estimated by the Least Squares Method:

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

The individual random error terms ε_i have a zero mean.

For finding minimum we take derivatives by both parameters and put equal to zero

$$\frac{d\left(\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2\right)}{d\beta_0}=0,$$

$$\frac{d\left(\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2\right)}{d\beta_1}=0.$$

After expressing β_0 and β_1 from the obtained equations we get least squares estimates

$$\widehat{\beta}_0=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i^2\overline{y}-\frac{1}{n}\sum_{i=1}^{n}x_i y_i\overline{x}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2},$$

$$\widehat{\beta}_1=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\overline{x}\cdot\overline{y}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2}.$$