



Topic 1: Simple Linear Regression



Simple Linear Regression Model

Suppose we have reason to believe that one economic indicator is causally associated with another one.

For example,

- *inflation and an unemployment rate in a country;*
- *demand for a product and its price;*
- *the gross domestic product (GDP) growth rate and profitability of securities.*

There is a set of statistical data on indicators that are of our interest.

The task:

Using the available empirical data set, it is necessary to choose (if possible) a function that links these two economic indicators.

Simple Regression Model

Let X be a random variable that describes the 1st economic indicator – *independent variable (regressor)*,

Y be a random variable that describes the 2nd economic indicator – *dependent variable*.

A pair (X, Y) is a two-dimensional random variable, and the ordered set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ forms a two-dimensional random sample of (X, Y) .

This tie in a general has the following form $Y = f(X) + \varepsilon$.

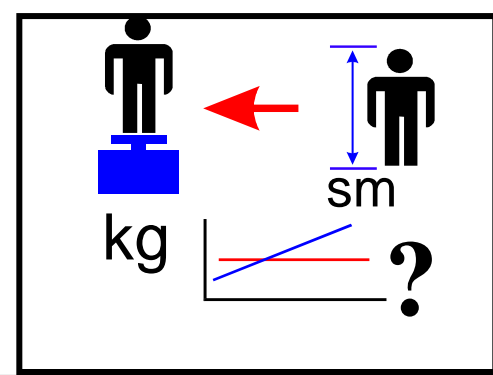
We look for a function $f(X)$. The explained part of Y is its conditional expectation $E(Y|X = x)$ for a given value x of X .

Simple Regression Model takes the form

$$Y = \bar{y}_x + \varepsilon,$$

where $\bar{y}_x = E(Y | X = x)$ is conditional expectation of Y for a given value x of X .

Simple Linear Regression Model



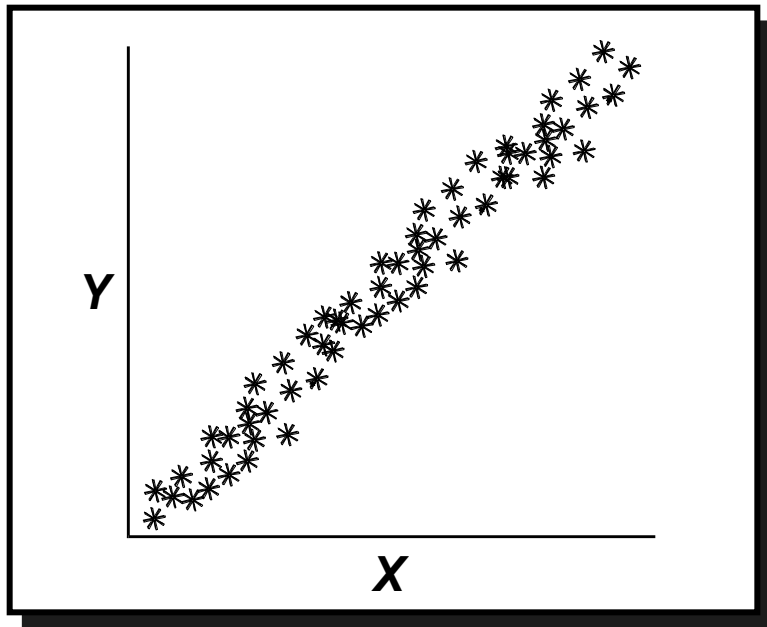
Assume that a relationship between Y and X is closed to linear, so, we have

$$y = \alpha + \beta x + \varepsilon$$

or

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

(1)



Correlation field

or

Scatter plot

In order to check the linearity hypothesis between two variables we need to look at their correlation field.



Linear Regression Function

Definition. The function $\hat{y}_x = E(Y | X = x)$ that determines the mean value of variable Y , given that the independent variable takes a fixed value x , is called *the regression function*.

Consider a linear regression model (1):

$$y = \alpha + \beta x + \varepsilon.$$

For a linear case the conditional expectation of r.v. Y is

$$\bar{y}_x = \alpha + \beta x.$$

This is the theoretic *linear regression function*.

The theoretical and the empirical regression functions (*population and sample*)

The theoretical regression model has the form:

$$y = \bar{y}_x + \varepsilon \quad (2)$$

where

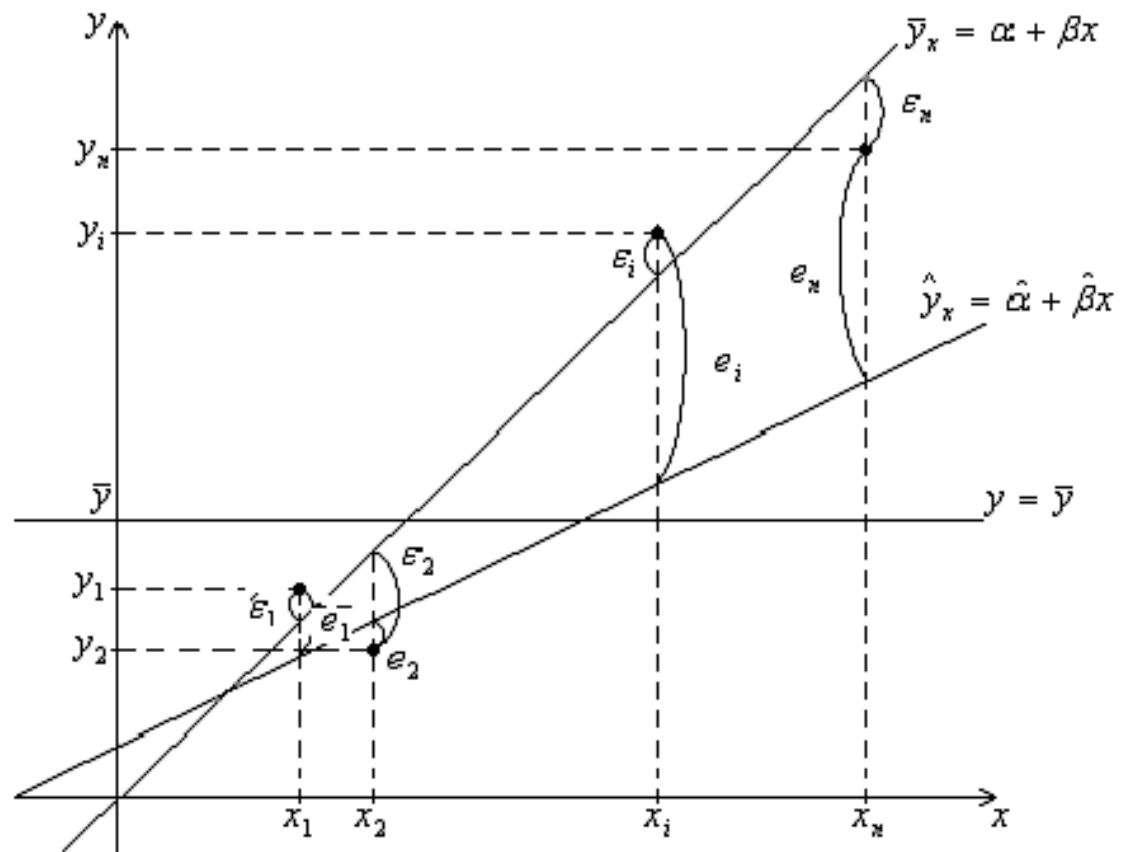
$$\bar{y}_x = \alpha + \beta x$$

The empirical regression model:

$$y = \hat{y}_x + e \quad (3)$$

where

$$\hat{y}_x = \hat{\alpha} + \hat{\beta}x.$$

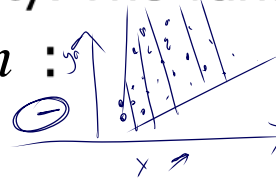


Gauss-Markov assumptions

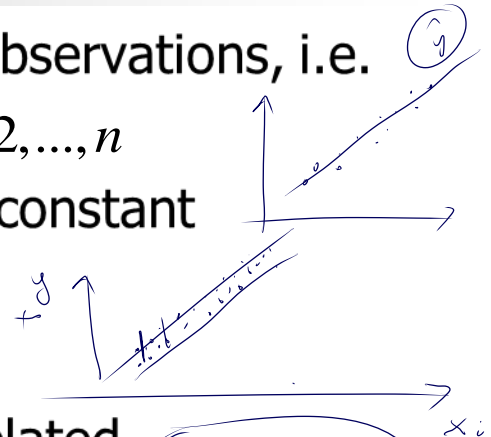
1. The expected value of the error term is zero for all observations, i.e.

$$E(\varepsilon_i) = 0, \quad i = 1, 2, \dots, n$$

2. Homoskedasticity. The variance of the error term is constant for all $i = 1, 2, \dots, n$:



$$Var(\varepsilon_i) = \sigma^2.$$



3. Error term is independently distributed and not correlated

x_i

x_j

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j.$$

$$\frac{\sum \varepsilon_i \varepsilon_j - \bar{\varepsilon}_i \bar{\varepsilon}_j}{\sqrt{\sum \varepsilon_i^2 \sum \varepsilon_j^2}}$$

4. x_i is deterministic: independent variable x is uncorrelated with the error term.

Additional assumption (!)

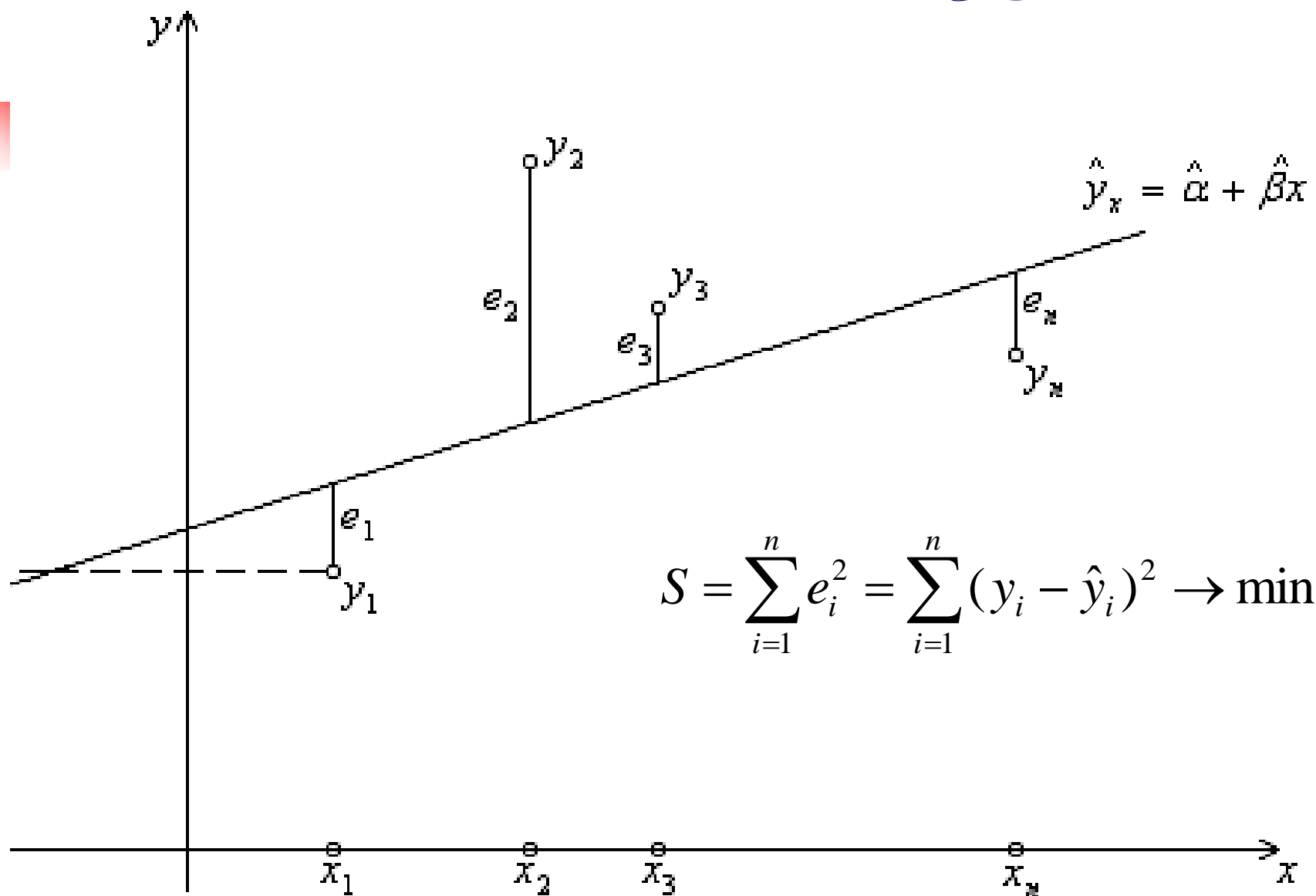
$$\varepsilon_i \in N(0, \sigma^2)$$

5. Observation errors are normally distributed random variables


The error variance is a measure of model uncertainty.

Homoskedasticity implies that the model uncertainty is identical across observations.

Ordinary Least Squares (OLS)



Gauss-Markov theorem


$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

Theorem. *Under the assumptions 1-4 for model (1), estimators obtained by the ordinary least squares method (OLS estimators), have the lowest variance in the class of linear unbiased estimators, and they equal:*

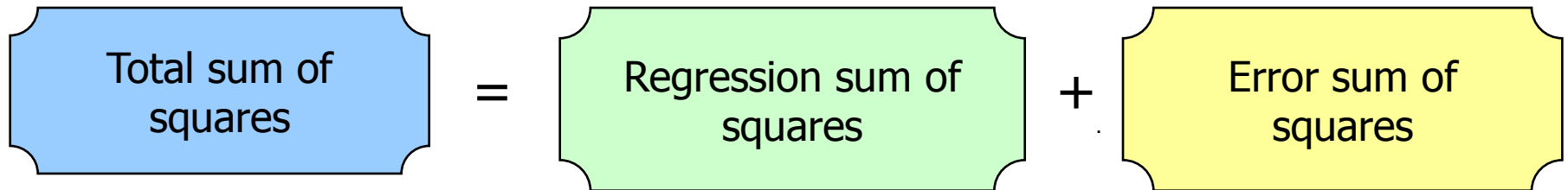
$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \frac{\hat{\mu}_{xy}}{\hat{\sigma}_x^2} \end{cases} \quad (2)$$

Proof: *see* YouTube. Ben Lambert 's video courses by «Deriving Least Squares Estimators» (Parts 1-4), «Gauss-Markov Assumptions» (Parts 1-3).

*The theorem states that in a linear regression model in which the errors have expectation zero and are uncorrelated and have equal finite variances, the best linear unbiased estimator (**BLUE**) of the coefficient is given by the ordinary least squares (**OLS**) estimator.*

The main equation of regression analysis

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{(***)} = \underbrace{\sum_{i=1}^n (\hat{y}_{x_i} - \bar{y})^2}_{(*)} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_{x_i})^2}_{(**)}$$



$$TSS = RSS + ESS$$

- If X does not have any effect on Y $\rightarrow TSS = ESS$
- If the dependent variable Y is not influenced by anything other than X $\rightarrow TSS = RSS$

Response (Y)

\bar{Y}

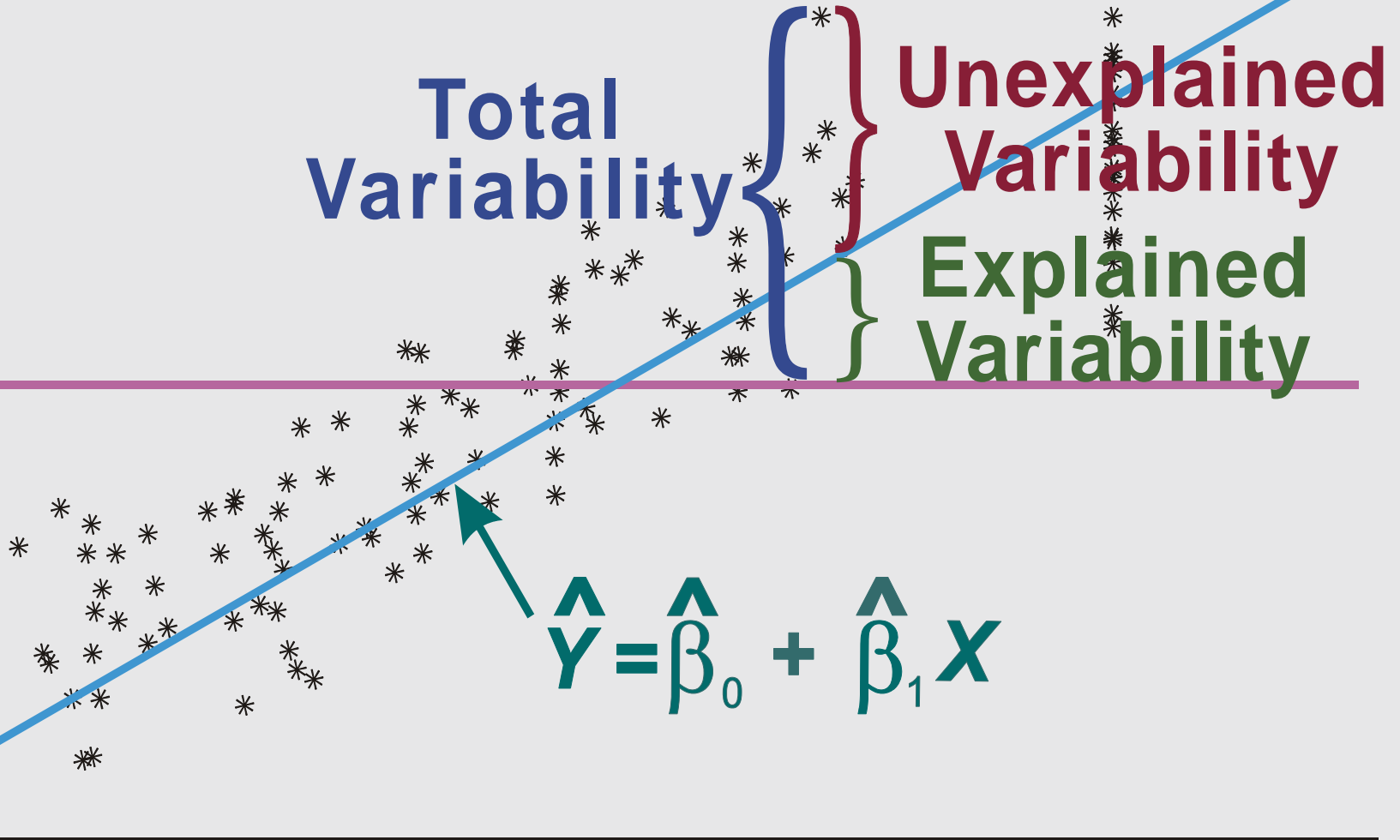
Total
Variability

Unexplained
Variability

Explained
Variability

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Predictor (X)



Coefficient of determination and its application in the regression analysis

The coefficient of determination is a number that indicates the proportion of the variance in the dependent variable that is defined by the independent variable.

The coefficient of determination is a statistical measure of how well the regression line approximates the real data. R^2 of 1 indicates that the regression line perfectly fits the data.

Definition. The coefficient of determination is defined by the formula

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}.$$

Only if α (constant) is in a model

$$R^2 \in [0,1]$$

It follows from Definition that

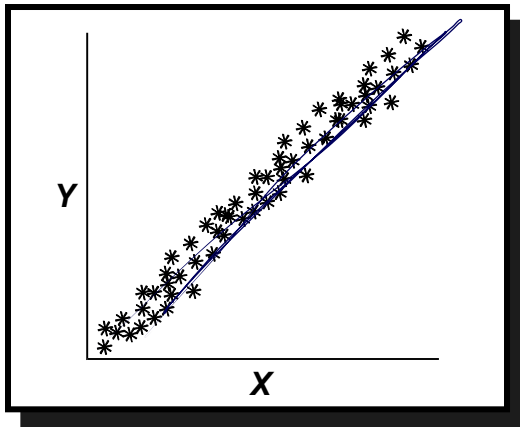
- If $R^2 = 0$ \Rightarrow model does nothing;
- If $R^2 = 1$ \Rightarrow there is an exact fit: all observations lie on the regression line (all the residuals = 0);
- The closer R^2 to 1, the better the quality of the considered model.

The coefficient of determination and the correlation coefficient

Consider how R^2 relates to the correlation coefficient $\hat{\rho}_{xy}$.
In case of simple linear regression, we have $R^2 = \hat{\rho}_{xy}^2$

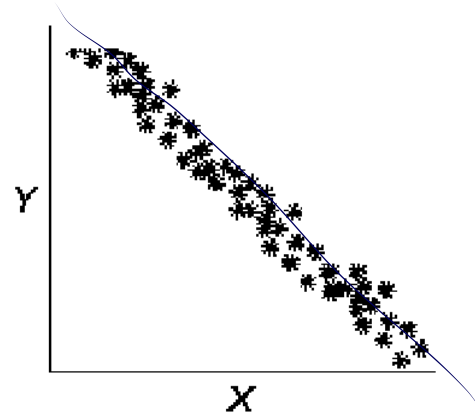
It is well known that

$|\rho_{xy}| = 1 \iff$ Random variables X and Y are linearly related to each other.



direct relationship

$$a) \hat{\rho}_{xy} \approx 1$$



inverse relationship

$$b) \hat{\rho}_{xy} \approx -1$$

Verification of the linear regression model

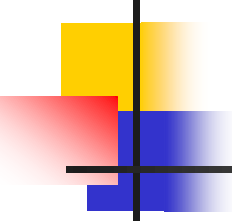
At this stage, it is usually found out how successfully the problems of specification and identification of the model are solved. What is the accuracy of forecasts for this model, and ultimately, how the constructed model corresponds to the simulated real economic phenomenon (object).

- Checking the adequacy of the constructed model:
 - F-test (Fisher test) helps to check statistical significance of the model.
- Checking the significance of the model's parameters (how do the parameters of the model differ from 0?):
 - Student's t-test helps to check the significance of the model's parameters (coefficients);
 - confidence intervals for the model parameters.
- Estimation of the accuracy for a model:
 - the coefficient of determination;
 - the standard error of a regression model;
 - error sum of squares (ESS);
 - MAD – the mean absolute deviation, and others.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{ESS}{TSS}$$
$$SEE = \sqrt{\frac{\sum e_i^2}{n-k-1}}$$

Handwritten notes: $\rightarrow 0$ (under ESS), $\rightarrow 0$ (under MAD), and $\sqrt{\frac{ESS}{n-k-1}}$ (under SEE).

$$\frac{1}{n} \sum |y_i - \bar{y}| = \frac{1}{n} \sum |e_i|$$

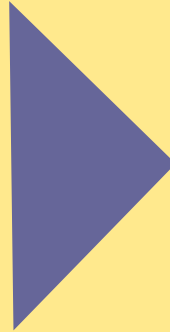


Analysis of residuals (errors) of a regression model

- One have to verify the assumptions of the regression analysis (**Gauss-Markov conditions**) using:
 1. a heteroscedasticity test;
 2. an autocorrelation test;
 3. a normality test;
 4. visual check of all the assumptions (using graphs).

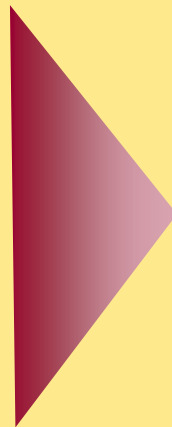
Goodness of fit

Quality indicators for regression coefficients



- Standard errors $s_{\hat{\alpha}}, s_{\hat{\beta}}$
- t-statistics
- P-value и др.
- Confidence intervals

Quality indicators for a regression equation



- Coefficient of determination R^2
- F-statistics
- Correlation coefficient
- Error sum of squares (ESS)
- Standard error s_{ε}
- Mean approximation error
- Mean absolute deviation (MAD)
- others

FORMULAE

Let k be a number factors in the regression model, for a simple linear model $k = 1$.

- standard errors of the regression coefficients

$$s_{\hat{\alpha}} = s_{\varepsilon} \sqrt{\frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}}$$
$$s_{\hat{\beta}} = \frac{s_{\varepsilon}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

- standard errors of the regression equation

$$s_{\varepsilon} = \sqrt{\frac{\sum_i (y_i - \hat{y}_{x_i})^2}{n - k - 2}}$$

$$F = \frac{ESS/k}{RSS/(n - k - 1)} \quad \text{or} \quad F = \frac{n - k - 1}{k} \cdot \frac{R^2}{1 - R^2} \propto F(k, n - k - 1)$$

- F-statistics

$$t = \frac{\hat{\alpha}}{s_{\hat{\alpha}}} \propto t(n - k - 1) \quad t = \frac{\hat{\beta}}{s_{\hat{\beta}}} \propto t(n - k - 1)$$

- t-statistics for coefficients $\hat{\alpha}$ and $\hat{\beta}$

Statistical significance of the regression equation

Consider the model (1) satisfied assumptions 1-4 and additionally assume that the fifth one holds, i.e. $\varepsilon_i \propto N(0, \sigma^2), i = 1, n$.

1. Null-hypothesis: $H_0 : R^2 = 0$


Alternative hypothesis: $H_1 : R^2 \neq 0$

2. Significance level: α

3. F-statistics: $F = (n - 2) \frac{R^2}{1 - R^2} \propto F(1, n - 2)$

4. Critical region: $\omega^{cr} = (F_{\alpha, 1, n-2}^{cr}, \infty)$

5. The decision rule is:

if $F_{obs} \in \omega^{cr}$, *i.e.* $F_{obs} > F_{\alpha, 1, n-2}^{cr}$ 

The null -hypothesis of the statistical insignificance of the regression equation is rejected and the hypothesis H_1 is accepted. This means that the regression equation is statistical significant for the given significance level α

Analysis of statistical significance of the regression coefficients

$$s_{\hat{\alpha}} = s_{\varepsilon} \sqrt{\frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}}$$
$$s_{\hat{\beta}} = \frac{s_{\varepsilon}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

1. Null hypothesis:

$$H_0 : \beta = 0$$

Alternative hypothesis:

$$H_1 : \beta \neq 0$$

2. Significance level:

$$\alpha$$

3. t-statistics:

$$t = \frac{\hat{\beta}}{s_{\hat{\beta}}} \propto t(n-2)$$

4. Critical region:

$$\omega^{cr} = (-\infty, -t_{\alpha/2, n-2}^{cr}) \cup (t_{\alpha/2, n-2}^{cr}, \infty)$$

5. The decision rule is:

$$\text{if } t_{obs} \in \omega^{cr}, \text{ i.e. } |t_{obs}| > t_{\alpha/2, n-2}^{cr} \quad \Rightarrow$$

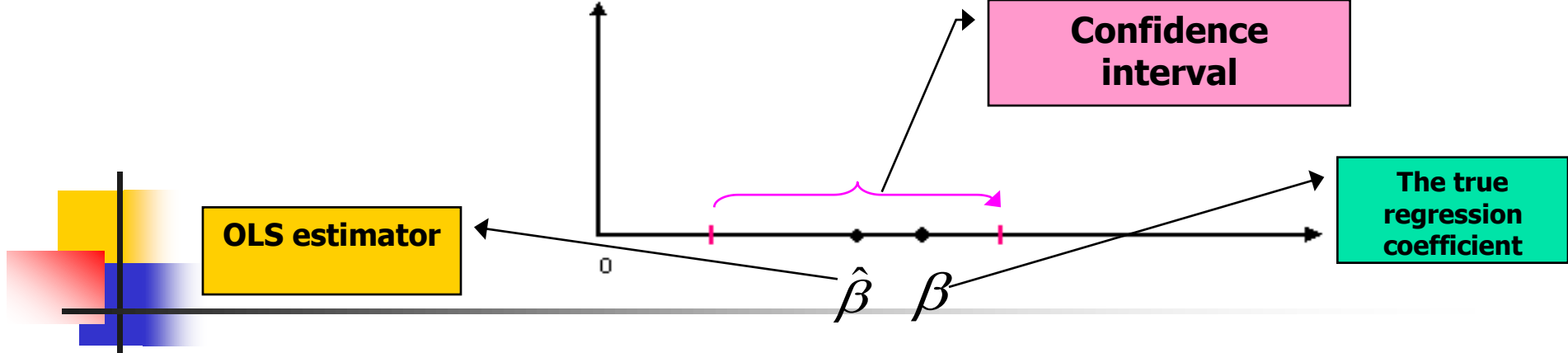
The null-hypothesis of the statistical insignificance of the regression coefficient $\hat{\beta}$ is rejected and the hypothesis H_1 is accepted.

Confidence intervals for the coefficients of the linear regression equation

Let β be an arbitrary unknown parameter of the model (1), $\gamma = 1 - \alpha$ be the confidence probability, then we have

$$P(-t_{\alpha/2, n-2}^{cr} < \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} < t_{\alpha/2, n-2}^{cr}) = 1 - \alpha.$$

The confidence interval is a range of values for a parameter so defined that there is a specified probability γ that the value of the parameter lies within it.



Let us find the confidence interval that covers the true coefficient β with the given probability $\gamma = 1 - \alpha$.

$$-t_{\alpha/2, n-2}^{kp} < \frac{\hat{\beta} - \beta}{s_{\hat{\beta}_1}} < t_{\alpha/2, n-2}^{kp} \quad \Rightarrow$$

Confidence interval for β is

$$\hat{\beta} - s_{\hat{\beta}} \cdot t_{\alpha/2, n-2}^{cr} < \beta < \hat{\beta} + s_{\hat{\beta}} \cdot t_{\alpha/2, n-2}^{cr}.$$

Confidence interval for α is

$$\hat{\alpha} - s_{\hat{\alpha}} \cdot t_{\alpha/2, n-2}^{cr} < \alpha < \hat{\alpha} + s_{\hat{\alpha}} \cdot t_{\alpha/2, n-2}^{cr}.$$

The principle of simplicity: *"From two or more nearly equivalent to «good» models for forecasting and analysis the simplest one should be chosen"*

Evaluation of the accuracy and selection the «best» model

Mean absolute deviation:

$$MAD = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_{x_i}|,$$

where y_i is the i th observation of Y ,

\hat{y}_{x_i} is a fitted value derived from the regression equation.

Mean approximation error:

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_{x_i}}{y_i} \right| \cdot 100\%.$$

If \bar{A} does not exceed 8-10%, the accuracy of a linear model is considered as good.

Forecasting using the linear regression model

Confidence
probability: $\gamma = 1 - \alpha$

The forecast at point $x = x_0$ is

$$\hat{y}_{x_0} = \hat{\alpha} + \hat{\beta}x_0.$$

A confidence interval for a true predicted value is

$$\hat{y}_{x_0} - s_{\hat{y}_{x_0}} \cdot t_{\alpha/2, n-2}^{cr} < y_{x_0}^{forecast} < \hat{y}_{x_0} + s_{\hat{y}_{x_0}} \cdot t_{\alpha/2, n-2}^{cr}.$$

$$s_{\hat{y}_{x_0}} = s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- Standard forecast error

Example

Уравнение регрессии:

$$\hat{y}_{x_i} = 369 + 116,8x_i$$

$$s_{\hat{\alpha}} = 190, s_{\hat{\beta}} = 17,1$$

Доверительный интервал для β

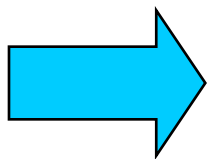
$$116,8 - 17,1 \cdot t_{0,05;18}^{кр} < \beta < 116,8 + 17,1 \cdot t_{0,05;18}^{кр} \quad \Longleftrightarrow \quad 80,9 < \beta < 152,7$$

Доверительный интервал для α

$$369 - 190 \cdot t_{0,05;18}^{кр} < \alpha < 369 + 190 \cdot t_{0,05;18}^{кр} \quad \Longleftrightarrow \quad -30 < \alpha < 768$$

Доверительный интервал для прогноза $x=20$:

$$\hat{y}_{20} = 369 + 116,8 \cdot 20 = 602,6$$
$$s_{\varepsilon} = 441,9 \quad s_{\hat{y}_{x_0}} = 441,9 \sqrt{1 + \frac{1}{20} + \frac{(20 - 9,5)^2}{17088771}} = 1,025$$



$$602,6 - s_{\hat{y}_0} \cdot t_{0,05;18}^{кр} < y_{20}^{прогн} < 602,6 + s_{\hat{y}_0} \cdot t_{0,05;18}^{кр}$$

$$600,45 < y_{20}^{прогн} < 604,75 \quad t_{0,05;18}^{кр} = 2,10$$