Statistics

Lecture 12

Simple Linear Regression

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistics
INM/BASTA

Outline of the lecture

- Introduction
- Least Squares Method
- An application: the trend of a time series
- Summary & Background
- The Classical Assumptions
- The Coefficient of Determination (R^2)
- Further Theorems, Tests of Hypotheses and Confidence Intervals
- Two-sample *t*-test for the difference of the population means // $\sigma_X = \sigma_Y$
- Simple linear regression without the intercept term

Simple Linear Regression: Motivation

Motivation:

Assume a dataset $(y_i, x_i)_{i=1}^{n}$ of n statistical units, i.e. we are given n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of quantitative variables $(x_i, y_i \in \mathbb{R})$, such as

- x_i = investments and y_i = the resulting revenues
- x_i = particular times and y_i = the price of a stock at the given time
- x_i = the quantity of some goods supplied to a market and y_i = the resulting unit price for the goods
- etc.

Simple Linear Regression: Motivation

Given the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the measurements, we assume that there is a simple linear relationship between the values of x and Y of the form

$$Y \approx \beta_0 + \beta x \qquad \text{for some} \quad \beta_0, \beta \in \mathbb{R}$$

or rather

$$Y = \beta_0 + \beta x + \varepsilon \qquad \text{for some} \quad \beta_0, \beta \in \mathbb{R}$$

where ε is a random deviation.

We do not know the parameters β_0 and β, however…

Simple Linear Regression: Motivation

Based on the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the measurements, it is our purpose to find

the estimates b_0 and b

of

the unknown β_0 and β

The estimates b_0 and b are also denoted by $\hat{\beta}_0$ and $\hat{\beta}$, respectively, sometimes, i.e. the estimates are

$$b_0 = \hat{\beta}_0 \qquad \text{and} \qquad b = \hat{\beta}$$

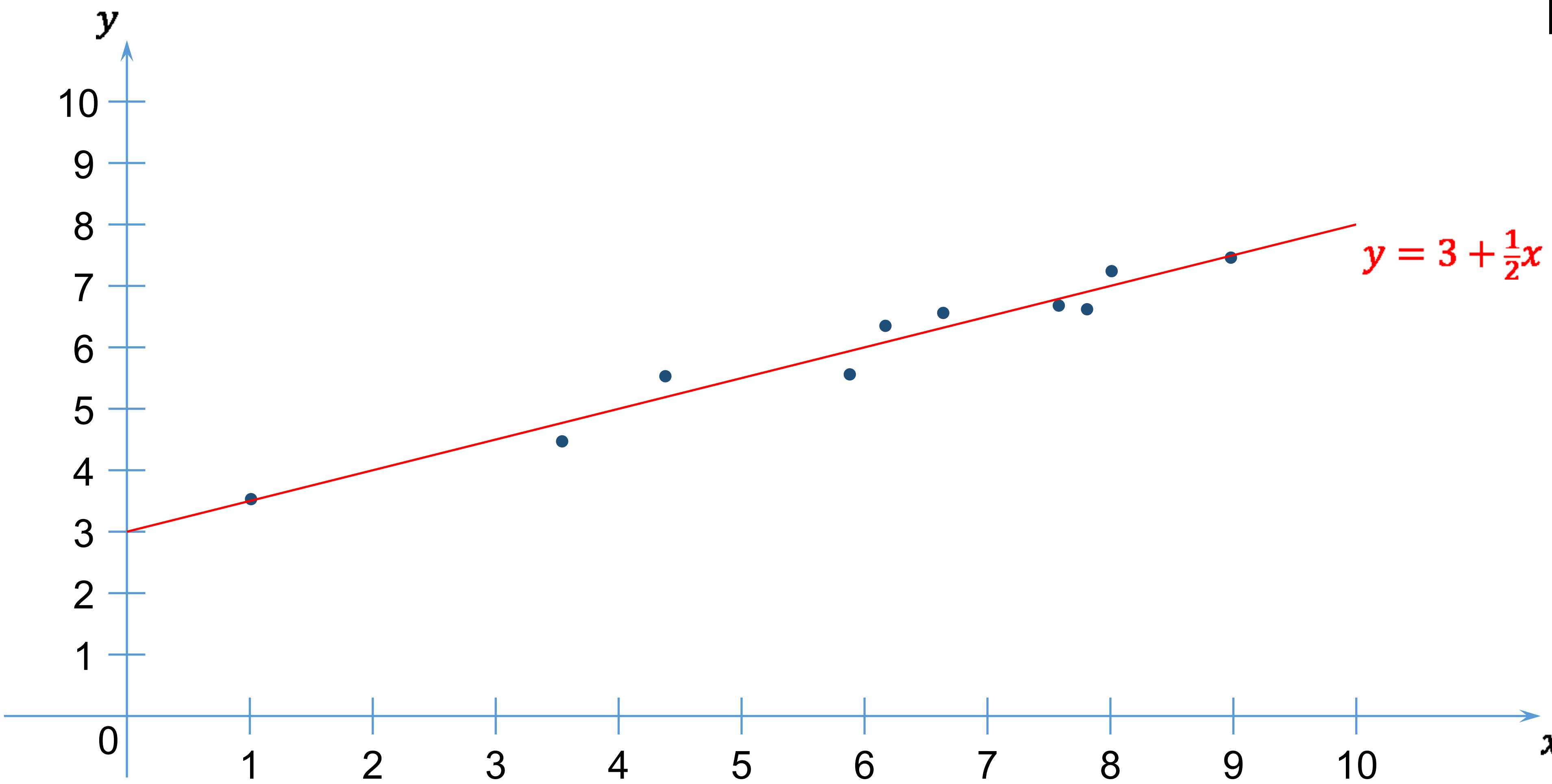
Simple Linear Regression: Example

We have got a sample of $n = 10$ observations:

i	x_i	y_i
□ 1	8.01	7.24
□ 2	7.81	6.62
□ 3	4.38	5.53
□ 4	3.54	4.47
□ 5	6.17	6.35
□ 6	6.64	6.56
□ 7	7.58	6.68
□ 8	8.98	7.46
□ 9	1.01	3.53
10	5.88	5.56

E.g.: x_i = temperature & y_i = the length of a metal rod

Simple Linear Regression: Example



Simple Linear Regression: Least Squares Method

We have got the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations.

For any $b_0, b \in \mathbb{R}$, the i-th **predicted (estimated) value** is

$$\hat{y}_i = b_0 + bx_i \qquad \text{for} \quad i = 1, 2, \ldots, n$$

The i-th **residual** is the difference

$$\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i \qquad \text{for} \quad i = 1, 2, \ldots, n$$

The **residual sum of squares** is

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (b_0 + bx_i - y_i)^2$$

Simple Linear Regression: Least Squares Method

Given the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations,

find $b_0, b \in \mathbb{R}$ so that the residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n} (b_0 + bx_i - y_i)^2 \quad \longrightarrow \quad \min$$

is minimized.

The first-order optimality conditions are

$$\frac{\partial \text{RSS}}{\partial b_0} = 0 \qquad \text{and} \qquad \frac{\partial \text{RSS}}{\partial b} = 0$$

Simple Linear Regression: Least Squares Method

Given $\mathrm{RSS} = \sum_{i=1}^{n}(b_0 + bx_i - y_i)^2$, we obtain the system of two equations of two unknowns:

$$\frac{\partial \mathrm{RSS}}{\partial b_0} = \sum_{i=1}^{n} 2(b_0 + bx_i - y_i) = 0 \quad \text{and} \quad \frac{\partial \mathrm{RSS}}{\partial b} = \sum_{i=1}^{n} 2(b_0 + bx_i - y_i)x_i = 0$$

or

$$\left.\begin{aligned} n\,b_0 + \sum_{i=1}^{n} x_i\, b &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, b_0 + \sum_{i=1}^{n} x_i^2\, b &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\}$$

the normal equation

Simple Linear Regression: Least Squares Method

Hence, given the observations $(y_1, x_1),\ (y_2, x_2),\ \ldots,\ (y_n, x_n)$, the estimates are:

$$\hat{\beta}_0 = b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\, b\right) =$$

$$= \frac{\sum_{i=1}^{n} x_i x_i \sum_{j=1}^{n} y_j - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j y_j}{n\sum_{i=1}^{n} x_i x_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

and

$$\hat{\beta} = b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n\sum_{i=1}^{n} x_i x_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

An application: the trend of a time series

The trend of a time series

Assume a series of observations (such as the GDP, a stock price, etc.) $y_1, y_2, \dots, y_n$ at times $t = 1, 2, \dots, n$. Assuming that the observed quantity y follows the linear trend

$$y \approx \beta_0 + \beta t \qquad \text{for} \quad t = 1, 2, \dots, n$$

put

$$x_i := i \qquad \text{for} \quad i = 1, 2, \dots, n$$

and apply the least squares method (the linear regression) to find the (linear) trend of the time series.

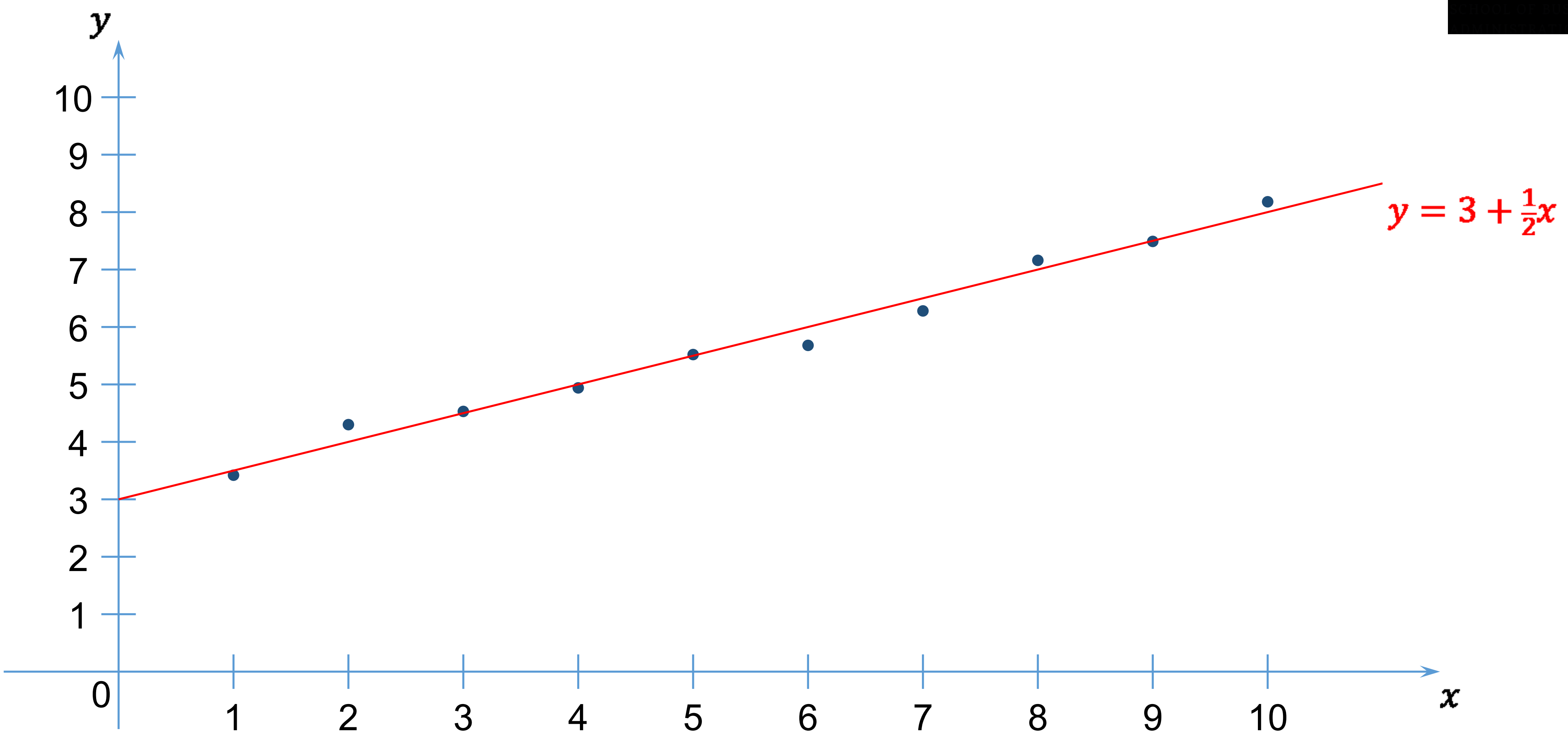
The trend of a time series: Example

We have got a sample of $n = 10$ observations:

i	x_i	y_i
□ 1	□ 1	3.42
□ 2	□ 2	4.30
□ 3	□ 3	4.53
□ 4	□ 4	4.94
□ 5	□ 5	5.52
□ 6	□ 6	5.68
□ 7	□ 7	6.28
□ 8	□ 8	7.16
□ 9	□ 9	7.49
10	10	8.18

Here, e.g.: x_i = year & y_i = the wealth

The trend of a time series: Example



Summary & Background

Simple Linear Regression: Summary & Background

We have got the sample of the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations.

The sample could have been obtained in either of the following two ways:

(1) A sample of n statistical units was selected from a larger population. Next, each of the statistical units was measured and we have obtained the pairs (y_i, x_i) for $i = 1, 2, \ldots, n$ thus. The values $x_i \in \mathbb{R}$ were measured exactly. (!) Assuming $y_i \approx \beta_0 + \beta x_i$, we have $y_i = \beta_0 + \beta x_i + \varepsilon_i$, where ε_i is a random deviation (error); the random deviation is caused by the intrinsic properties of the statistical unit (further unknown / "random" /

Simple Linear Regression: Summary & Background

We have got the sample of the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations.

The sample could have been obtained in either of the following two ways:

(2) We have prepared the values $x_1, x_2, \dots, x_n$ first, and these values $(x_1, x_2, \dots, x_n)$ are assumed to be known exactly. When doing the i-th measurement, we first set up the system (adjust the system's setting, e.g. the temperature, to x_i exactly) and we measure the value y_i of the dependent variable then. The random deviation ε_i here is caused either by the intrinsic properties of the system (further unknown / "random" /

Simple Linear Regression: Summary & Background

Remarks:

- In practice, the data may be obtained in either way, (1) or (2).
- In either case, (1) or (2), the independent values $x_1, x_2, \dots, x_n$ are assumed to be known exactly, i.e. without any measurement errors.
- Assuming $y_i \approx \beta_0 + \beta x_i$, even the dependent values y_i may be measured exactly, i.e. without any measurement error, the random deviation $\varepsilon_i = y_i - \beta_0 - \beta x_i$ being caused by the intrinsic properties (other unknown / "random" / unconsidered factors).
- For the purpose of the mathematical analysis, we assume the case (2) only.

Simple Linear Regression: Summary & Background

Terminology:

The intercept term → β_0

Parameters

Regression coefficients → β_0, β

$$Y = \beta_0 + \beta x + \varepsilon$$

Regressand (Y)

Predicand

Explained variable

Dependent variable

Endogenous variable

Controlled variable

Response

Outcome

Predicted variable

Measured variable

Regressor (x)

Predictor

Explanatory variable

Independent variable

Exogenous variable

Control variable

Stimulus

Covariate

Deviation (ε)

Error term

Disturbance

Noise

Simple Linear Regression: Summary & Background

Assumptions:

- The n values $x_1, x_2, \dots, x_n \in \mathbb{R}$ are known exactly, fixed, given before the measurements.
- We have n random variables $Y_1, Y_2, \dots, Y_n$
  and n random variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$.
- We assume that the random variables $Y_1, Y_2, \dots, Y_n$ are independent
  and the random variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent.

Simple Linear Regression: Summary & Background

<u>Simple (¡ not exact !) assumptions:</u>

- Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space.
- For $i = 1, 2, \dots, n$, let $\omega_i \in \Omega$ be the outcome of the random experiment.
- We assume that it holds

$$y_i = Y_i(\omega_i) = \beta_0 + \beta x_i + \varepsilon_i(\omega_i)$$

and the expected value

$$\mathrm{E}[Y_i] = \beta_0 + \beta x_i$$

or, equivalently,

$$\mathrm{E}[\varepsilon_i] = 0 \qquad \text{for} \quad i = 1, 2, \dots, n$$

Simple Linear Regression: Summary & Background

Recall:

$$y_i = Y_i(\omega_i) = \beta_0 + \beta x_i + \varepsilon_i(\omega_i)$$

In other words:

- The measured value y_i is the numerical outcome $Y_i(\omega_i)$ of the random experiment.
- The numerical outcome $Y_i(\omega_i)$ is obtained so that the numerical outcome $\varepsilon_i(\omega_i)$ of the random experiment is added to the given value $\beta_0 + \beta x_i$

Notice:

¡¡¡ The regressor values $x_1, x_2, \dots, x_n$ are known exactly !!!

¡¡¡ But the values of the parameters β_0 and β are unknown !!!

Simple Linear Regression: Summary & Background

Notation — notice that:

— the unknown quantities (unknown parameters β_0 and β with deviations ε_i) are denoted by Greek letters

— the estimates of the parameters are denoted by the respective Latin letters (b_0 and b) or by the hat "^" ($\hat{\beta}_0$ and $\hat{\beta}$), so that $b_0 = \hat{\beta}_0$ and $b = \hat{\beta}$ are the estimates of the parameters β_0 and β, respectively

We shall mainly use the Latin letters b_0 and b to denote the estimates of the parameters β_0 and β, respectively, here.

Simple Linear Regression: Summary & Background

Notation — notice that:

— the unknown quantities (unknown parameters β_0 and β with deviations ε_i) are denoted by Greek letters

— the estimates of the parameters are denoted by the respective Latin letters (b_0 and b) or by the hat "^" ($\hat{\beta}_0$ and $\hat{\beta}$), so that $b_0 = \hat{\beta}_0$ and $b = \hat{\beta}$ are the estimates of the parameters β_0 and β, respectively

— the predicted values of the dependent variable are denoted by the hat "^":

$$\hat{y}_i = b_0 + bx_i$$

Simple Linear Regression: Summary & Background

<u>Notation — notice that:</u>

— the <u>unknown</u> quantities (unknown parameters β_0 and β with **<u>deviations</u>** ε_i) are denoted by <u>Greek letters</u>

— the i-th **<u>residual</u>** is denoted by the respective <u>Latin letter</u> (e_i) <u>or</u> by the <u>hat</u> "^" ($\hat{\varepsilon}_i$), so that

$$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$$

is the i-th residual.

We shall mainly use the Latin letter e_i to denote the residual here.

The Classical Assumptions

Simple Linear Regression: Assumptions

The classical assumptions of the simple linear regression model:

- Assume n fixed values $x_1, x_2, \dots, x_n \in \mathbb{R}$, which are known exactly.
- Assume n independent & normally distributed random variables $Y_1, Y_2, \dots, Y_n$ such that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta x_i, \sigma^2) \qquad \text{for} \quad i = 1, 2, \dots, n$$

for some parameters $\beta_0, \beta \in \mathbb{R}$ and for some $\sigma^2 \in \mathbb{R}^+$.

That is

$$\mathrm{E}[Y_i] = \beta_0 + \beta x_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

linearity

and

$$\mathrm{Var}(Y_i) = \sigma^2 \qquad \text{for} \quad i = 1, 2, \dots, n$$

homoskedasticity

Simple Linear Regression: Assumptions

By introducing new random variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, it is equivalent to assume that

$$Y_i = \beta_0 + \beta x_i + \varepsilon_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

for some parameters $\beta_0, \beta \in \mathbb{R}$,

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad \text{for} \quad i = 1, 2, \dots, n$$

for some $\sigma^2 \in \mathbb{R}^+$, and

the deviations $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent.

Further Theorems, Tests of Hypotheses and Confidence Intervals

Simple Linear Regression: The Normal Equation

Given the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations,

recall that the **Residual Sum of Squares** for the estimates $b_0, b \in \mathbb{R}$ is

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - bx_i)^2$$

By letting $\frac{\partial \text{RSS}}{\partial b_0} = 0$ and $\frac{\partial \text{RSS}}{\partial b} = 0$, we obtain the system

$$\left.\begin{aligned} n\, b_0 + \sum_{i=1}^{n} x_i\, b &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, b_0 + \sum_{i=1}^{n} x_i^2\, b &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\}$$

the normal equation

Simple Linear Regression: The Normal Equation

Notice that by letting

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

the normal equation is written as:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\begin{pmatrix} b_0 \\ b \end{pmatrix} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

the normal equation

Under the assumption that $x_i \neq x_j$ for at least one $i \neq j$,

notice that the $\mathrm{rank}(\boldsymbol{X}) = 2$ and the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ of the system is non-singular.

Simple Linear Regression: The Normal Equation

Assume therefore, for simplicity, that $x_i \neq x_j$ for some $i \neq j$ in the following.

Then the **normal equation** is

$$X^T X \begin{pmatrix} b_0 \\ b \end{pmatrix} = X^T Y$$

We then know the matrix $X^T X$ is non-singular.

Let

$$C = (X^T X)^{-1}$$

Then the solution is:

$$\begin{pmatrix} b_0 \\ b \end{pmatrix} = C X^T Y$$

Simple Linear Regression: The Normal Equation

Since the normal equation is

$$n\, b_0 + \sum_{i=1}^{n} x_i\, b = \sum_{i=1}^{n} y_i$$
$$\sum_{i=1}^{n} x_i\, b_0 + \sum_{i=1}^{n} x_i^2\, b = \sum_{i=1}^{n} x_i y_i$$

the matrix of the system is

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

and its inverse is

$$\boldsymbol{C} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix} = \begin{pmatrix} \dfrac{\sum_{i=1}^{n} x_i^2}{\Delta} & -\dfrac{\sum_{i=1}^{n} x_i}{\Delta} \\ -\dfrac{\sum_{i=1}^{n} x_i}{\Delta} & \dfrac{n}{\Delta} \end{pmatrix}$$

Simple Linear Regression: The Normal Equation

The solution to the normal equation is $\begin{pmatrix} b_0 \\ b \end{pmatrix} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = CX^{\mathrm{T}}Y$,

hence

$$b_0 = c_{00}\sum_{i=1}^{n} y_i + c_{01}\sum_{i=1}^{n} x_i y_i = \frac{\sum_{i=1}^{n} x_i^2 \sum_{j=1}^{n} y_j - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j y_j}{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

and

$$b = c_{10}\sum_{i=1}^{n} y_i + c_{11}\sum_{i=1}^{n} x_i y_i = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

(as we already know, see above).

Simple Linear Regression

Recalling that the values $x_1, x_2, \dots, x_n \in \mathbb{R}$ are given, that we assume

$$Y_i = \beta_0 + \beta x_i + \varepsilon_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

with

$$\mathrm{E}[Y_i] = \beta_0 + \beta x_i \qquad \text{and} \qquad \mathrm{Var}(Y_i) = \sigma^2 \qquad \text{for} \quad i = 1, 2, \dots, n$$

or

$$\mathrm{E}[\varepsilon_i] = 0 \qquad \text{and} \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2 \qquad \text{for} \quad i = 1, 2, \dots, n$$

where the random variables $Y_1, Y_2, \dots, Y_n$, or $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, respectively, are independent, and that $y_1, y_2, \dots, y_n$ are some observations of the random variables $Y_1, Y_2, \dots, Y_n$, it follows that all the estimates

$$b_0 = \hat{\beta}_0 \qquad b = \hat{\beta} \qquad \hat{y}_1, \hat{y}_2, \dots, \hat{y}_n \qquad \text{RSS} \qquad \text{etc.}$$

Simple Linear Regression: Theorem 1

Theorem 1: It holds

$$\mathrm{E}[\hat{y}_i] = \mathrm{E}[b_0 + b x_i] = \beta_0 + \beta x_i = \mathrm{E}[Y_i]$$

and

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 \left(X C X^{\mathrm{T}} \right)_{ii} = \sigma^2 \frac{n x_i^2 - 2 x_i \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^2}{n \sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k \sum_{l=1}^{n} x_l}$$

Remark:
Recall that the **variance** of the random variable $\hat{y}_i$ is

$$\mathrm{Var}(\hat{y}_i) = \mathrm{E}[(\hat{y}_i - \mathrm{E}[\hat{y}_i])^2]$$

Simple Linear Regression

<u>**Remark:**</u> We actually have

$$\mathrm{cov}(\hat{y}_i, \hat{y}_j) = \sigma^2 \left(XCX^{\mathrm{T}}\right)_{ij} = \sigma^2 \frac{n x_i x_j - (x_i + x_j)\sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^2}{n \sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k \sum_{l=1}^{n} x_l}$$

where

$$\mathrm{cov}(\hat{y}_i, \hat{y}_j) = \mathrm{E}[(\hat{y}_i - \mathrm{E}[\hat{y}_i])(\hat{y}_j - \mathrm{E}[\hat{y}_j])]$$

is the **covariance** of the random variables $\hat{y}_i$ and $\hat{y}_j$

Observe that

$$\mathrm{Var}(\hat{y}_i) = \mathrm{cov}(\hat{y}_i, \hat{y}_i)$$

Simple Linear Regression: Theorem 2

Theorem 2: It holds

$$\mathrm{E}[e_i] = \mathrm{E}[Y_i - \hat{y}_i] = 0$$

and

$$\mathrm{Var}(e_i) = \mathrm{Var}(Y_i - \hat{y}_i) = \sigma^2 \left(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^{\mathrm{T}}\right)_{ii} = \sigma^2 \left(1 - \frac{n x_i^2 - 2x_i \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^2}{n \sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k \sum_{l=1}^{n} x_l}\right)$$

Remark:

Notice that the first equation follows by Theorem 1 ($\mathrm{E}[\hat{y}_i] = \mathrm{E}[Y_i]$) above too.

Simple Linear Regression

Remark: For $i \neq j$, we actually have

$$\mathrm{cov}(e_i, e_j) = \mathrm{cov}(Y_i - \hat{y}_i, Y_j - \hat{y}_j) = \sigma^2 (\boldsymbol{I} - \boldsymbol{XCX}^{\mathrm{T}})_{ij} =$$

$$= -\sigma^2 \frac{n x_i x_j - (x_i + x_j) \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^2}{n \sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k \sum_{l=1}^{n} x_l} =$$

$$= -\mathrm{cov}(\hat{y}_i, \hat{y}_j) \qquad \text{if} \quad i \neq j$$

Simple Linear Regression

Recall the **Residual Sum of Squares** is

$$\mathrm{RSS} = \sum_{i=1}^{n} (b_0 + b x_i - y_i)^2$$

The **residual variance** or the **Mean Square Error** is

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{\sum_{i=1}^{n} (y_i - b_0 - b x_i)^2}{n-2}$$

<u>Remark:</u>

The “2” in the denominator is the rank of the matrix $\boldsymbol{X}$.

(Recall that we assume $\operatorname{rank}(\boldsymbol{X}) = 2$, whence $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is non-singular.)

Simple Linear Regression: Theorem 3

Theorem 3:

$$\mathrm{E}[s^2] = \sigma^2$$

where the **residual variance** or the **Mean Square Error** is

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2}{n-2}$$

Remark: Theorem 3 provides an estimate of the unknown variance:

$$\sigma^2 \approx s^2$$

Notice that the residual variance is also denoted by $\hat{\sigma}^2$, i.e. $\hat{\sigma}^2 = s^2$,

and it is an estimate of the variance: we have $\sigma^2 \approx \hat{\sigma}^2 = s^2$.

Simple Linear Regression: Theorem 4

<u>Theorem 4:</u> It holds

$$\mathrm{E}[b_0] = \beta_0 \qquad \text{and} \qquad \mathrm{E}[b] = \beta$$

with

$$\mathrm{Var}(b_0) = \sigma^2 c_{00} = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

and

$$\mathrm{Var}(b) = \sigma^2 c_{11} = \sigma^2 \frac{n}{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

Simple Linear Regression

Remark: We also have

$$\operatorname{cov}(b_0, b) = \sigma^2 c_{01} = -\sigma^2 \frac{\sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}$$

Simple Linear Regression: Theorem 5

Theorem 5: For any $p_0, p \in \mathbb{R}$, such that $p_0 \neq 0$ or $p \neq 0$, it holds

$$\frac{(p_0 b_0 + p b) - (p_0 \beta_0 + p \beta)}{\sqrt{s^2}\sqrt{(p_0 \quad p)\, C \begin{pmatrix} p_0 \\ p \end{pmatrix}}} \sim t_{n-2}$$

where t_{n-2} denotes Student's *t*-distribution with $n - \text{rank}(\boldsymbol{X}) = n - 2$ d.f.

Remark:

Notice the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is positive definite.

Therefore, its inverse $\boldsymbol{C}$ is positive definite too,

Simple Linear Regression: Corollaries

Corollary 0: By considering $p_0 = 1$ and $p = 0$, we obtain:

$$\frac{b_0 - \beta_0}{\sqrt{s^2}\sqrt{c_{00}}} \sim t_{n-2}$$

Corollary 1: By considering $p_0 = 0$ and $p = 1$, we obtain:

$$\frac{b - \beta}{\sqrt{s^2}\sqrt{c_{11}}} \sim t_{n-2}$$

Remark: Use the corollaries of Theorem 5

— for *t*-tests about the parameters β_0 and β of the model,

Tests of hypotheses about the parameters β_0 and β

- Choose any $p_0, p \in \mathbb{R}$ such that $p_0 \neq 0$ or $p \neq 0$,
  and let $b_{00} \in \mathbb{R}$ be a prescribed number.
- We can then use Theorem 5 to test the null hyp. H_0 that $(p_0\beta_0 + p\beta) = b_{00}$.
- Choosing $p_0 = 1$ with $p = 0$ in particular, we can use Corollary 0
  ($\frac{b_0 - \beta_0}{\sqrt{s^2}\sqrt{c_{00}}} \sim t_{n-2}$) to test the null hypothesis H_0 that $\beta_0 = b_{00}$ or $\beta_0 = 0$
  (if we put $b_{00} = 0$ in particular).
- Choosing $p_0 = 0$ with $p = 1$ in particular, we can use Corollary 1
  ($\frac{b - \beta}{\sqrt{s^2}\sqrt{c_{11}}} \sim t_{n-2}$) to test the null hypothesis H_0 that $\beta = b_{00}$ or $\beta = 0$
  (if we put $b_{00} = 0$ in particular).

t-test for the parameter β_0 or β

<u>Notation:</u> Let

$$t_{n-2}(p)$$

denote the **quantile function of Student's *t*-distribution** with $n-2$ d.f.

The quantile function $t_{n-2}(p)$ is the function inverse to the cumulative distribution function $F(x)$ of **Student's *t*-distribution** with $n-2$ degrees of freedom, i.e.

$$t_{n-2}(p) = F^{-1}(p) \qquad \text{for} \quad p \in (0,1)$$

t-test for the parameter β_0 or β

<u>Notation:</u> Let

$$t_{n-2}(p)$$

denote the **quantile function of Student's *t*-distribution** with $n-2$ d.f.

In other words, if $0 < p < 1$, then $x = t_{n-2}(p)$ is the unique value such that

$$\int_{-\infty}^{t_{n-2}(p)} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = p$$

where $f(t)$ is the density of Student's *t*-distribution with $n-2$ d.f.

t-test for the parameter β_0

Choosing a value $b_{00} \in \mathbb{R}$, formulate the **null hypothesis**

$$H_0: \quad \beta_0 = b_{00}$$

formulate the **alternative hypothesis**

- two-sided: $H_1: \quad \beta_0 \neq b_{00}$
- one-sided: $H_1: \quad \beta_0 < b_{00}$
- one-sided: $H_1: \quad \beta_0 > b_{00}$

and use aforementioned Corollary 0 to conduct the test.

t-test for the parameter β_0

Having chosen the value $b_{00} \in \mathbb{R}$, such as $b_{00} = 0$, and assuming the null hypothesis $H_0\colon \beta_0 = b_{00}$ is true, calculate the statistic

$$T = \frac{b_0 - \beta_0}{\sqrt{s^2}\sqrt{c_{00}}} = \frac{b_0 - b_{00}}{\sqrt{s^2}\sqrt{c_{00}}} = \frac{b_0}{\sqrt{s^2}\sqrt{c_{00}}} =$$

$$= \frac{b_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2}{n-2}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}}}$$

t-test for the parameter β_0

The t-test for β_0 with two-sided alternative hypothesis ($\beta_0 \neq b_{00}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.
- the **critical value** is $c = t_{n-2}\left(1 - \frac{\alpha}{2}\right)$
- if $T \in (-\infty, -c] \cup [+c, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-c, +c)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β_0

The *t*-test for β_0 with one-sided alternative hypothesis ($\beta_0 < b_{00}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical value** is $c = t_{n-2}(1-\alpha)$
- if $T \in (-\infty, -c]$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-c, +\infty)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β_0

The t-test for β_0 with one-sided alternative hypothesis ($\beta_0 > b_{00}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.
- the **critical value** is $c = t_{n-2}(1-\alpha)$
- if $T \in [+c, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-\infty, +c)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β_0

¡¡¡WARNING!!! Do not use the aforementioned test unless you know what and why you are doing.

Given the model $(Y = \beta_0 + \beta X + \varepsilon)$, the test for $\beta_0 = 0$ actually means the test whether Y is directly proportional to x, i.e.

$$Y \approx \beta x \qquad \text{for some} \quad \beta \in \mathbb{R}$$

or rather

$$Y = \beta x + \varepsilon \qquad \text{for some} \quad \beta \in \mathbb{R}$$

Confidence interval for the parameter β_0

Let $x, y \in \mathbb{R}$ be any numbers such that $x < y$ and let $F(x)$ be the cumulative distribution function of Student's *t*-distribution with $n-2$ degrees of freedom. Then, by the definition of the cumulative distribution function and by Corollary 0, the probability

$$P\left(x < \frac{b_0 - \beta_0}{\sqrt{s^2}\sqrt{c_{00}}} \leq y\right) = F(y) - F(x)$$

Therefore

$$P\left(x\sqrt{s^2}\sqrt{c_{00}} < b_0 - \beta_0 \leq y\sqrt{s^2}\sqrt{c_{00}}\right) = F(y) - F(x)$$

$$P\left(b_0 - y\sqrt{s^2}\sqrt{c_{00}} \leq \quad \beta_0 \quad < b_0 - x\sqrt{s^2}\sqrt{c_{00}}\right) = F(y) - F(x)$$

Confidence interval for the parameter β_0

We have:

$$P\left(b_0 - y\sqrt{s^2}\sqrt{c_{00}} \le \beta_0 < b_0 - x\sqrt{s^2}\sqrt{c_{00}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\ \%$.

Let $y = t_{n-2}\left(1 - \frac{\alpha}{2}\right)$ and let $x = -y = -t_{n-2}\left(1 - \frac{\alpha}{2}\right) = t_{n-2}\left(\frac{\alpha}{2}\right)$.

Recall that $t_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta_0 \in \left[b_0 - t_{n-2}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{c_{00}},\ b_0 + t_{n-2}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{c_{00}}\right]$$

Confidence interval for the parameter β_0

We have:

$$P\left(b_0 - y\sqrt{s^2}\sqrt{c_{00}} \le \beta_0 < b_0 - x\sqrt{s^2}\sqrt{c_{00}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = t_{n-2}(1-\alpha)$ and let $x = -\infty$. Recall that $t_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta_0 \in \left[b_0 - t_{n-2}(1-\alpha)\sqrt{s^2}\sqrt{c_{00}},\ +\infty\right)$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the parameter β_0

We have:

$$P\left(b_0 - y\sqrt{s^2}\sqrt{c_{00}} \le \beta_0 < b_0 - x\sqrt{s^2}\sqrt{c_{00}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = +\infty$ and let $x = t_{n-2}(\alpha)$. Recall that $t_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta_0 \in \left(-\infty,\ b_0 + t_{n-2}(1-\alpha)\sqrt{s^2}\sqrt{c_{00}}\right]$$

is about $1-\alpha = 95\,\%$.

t-test for the parameter β

Choosing a value $b_{00} \in \mathbb{R}$, formulate the **null hypothesis**

$$H_0: \quad \beta = b_{00}$$

formulate the **alternative hypothesis**

- two-sided: $H_1: \quad \beta \neq b_{00}$
- one-sided: $H_1: \quad \beta < b_{00}$
- one-sided: $H_1: \quad \beta > b_{00}$

and use aforementioned Corollary 1 to conduct the test.

t-test for the parameter β

Having chosen the value $b_{00} \in \mathbb{R}$, such as $b_{00} = 0$, and assuming the null hypothesis $H_0\colon \beta = b_{00}$ is true, calculate the statistic

$$T = \frac{b-\beta}{\sqrt{s^2}\sqrt{c_{11}}} = \frac{b-b_{00}}{\sqrt{s^2}\sqrt{c_{11}}} = \frac{b}{\sqrt{s^2}\sqrt{c_{11}}} =$$

$$= \frac{b}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2}{n-2}}\sqrt{\frac{n}{n\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}x_i\sum_{j=1}^{n}x_j}}}$$

t-test for the parameter β

The *t*-test for β with two-sided alternative hypothesis ($\beta \neq b_{00}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical value** is $c = t_{n-2}\left(1 - \frac{\alpha}{2}\right)$
- if $T \in (-\infty, -c] \cup [+c, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-c, +c)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β

The t-test for β with one-sided alternative hypothesis ($\beta < b_{00}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical value** is $c = t_{n-2}(1-\alpha)$
- if $T \in (-\infty, -c]$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-c, +\infty)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β

The t-test for β with one-sided alternative hypothesis ($\beta > b_{00}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.

- the **critical value** is $c = t_{n-2}(1 - \alpha)$

- if $T \in [+c, +\infty)$, **the critical region**, then **reject** the null hypothesis

- if $T \in (-\infty, +c)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β

¡¡¡**REMARK**!!!

Given the model $(Y = \beta_0 + \beta X + \varepsilon)$, the test for $\beta = 0$ actually means the test whether Y is completely random and independent of x, i.e.

$$Y \approx \beta_0 \qquad \text{for some} \quad \beta_0 \in \mathbb{R}$$

or rather

$$Y = \beta_0 + \varepsilon \qquad \text{for some} \quad \beta_0 \in \mathbb{R}$$

i.e.

$$Y \sim \mathcal{N}(\beta_0, \sigma^2) \qquad \text{for some} \quad \beta_0 \in \mathbb{R} \text{ and } \sigma^2 \in \mathbb{R}_0^+$$

consequently, whether there is no correlation between the variables x and Y.

Confidence interval for the parameter β

Let $x, y \in \mathbb{R}$ be any numbers such that $x < y$ and let $F(x)$ be the cumulative distribution function of Student's *t*-distribution with $n-2$ degrees of freedom. Then, by the definition of the cumulative distribution function and by Corollary 1, the probability

$$P\left(x < \frac{b-\beta}{\sqrt{s^2}\sqrt{c_{11}}} \leq y\right) = F(y) - F(x)$$

Therefore

$$P\left(x\sqrt{s^2}\sqrt{c_{11}} < b - \beta \leq y\sqrt{s^2}\sqrt{c_{11}}\right) = F(y) - F(x)$$

$$P\left(b - y\sqrt{s^2}\sqrt{c_{11}} \leq \quad \beta \quad < b - x\sqrt{s^2}\sqrt{c_{11}}\right) = F(y) - F(x)$$

Confidence interval for the parameter β

We have:

$$P\left(b - y\sqrt{s^2}\sqrt{c_{11}} \leq \beta < b - x\sqrt{s^2}\sqrt{c_{11}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = t_{n-2}\left(1 - \frac{\alpha}{2}\right)$ and let $x = -y = -t_{n-2}\left(1 - \frac{\alpha}{2}\right) = t_{n-2}\left(\frac{\alpha}{2}\right)$.

Recall that $t_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta \in \left[b - t_{n-2}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{c_{11}},\ b + t_{n-2}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{c_{11}}\right]$$

Confidence interval for the parameter β

We have:

$$P\left(b - y\sqrt{s^2}\sqrt{c_{11}} \le \beta < b - x\sqrt{s^2}\sqrt{c_{11}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = t_{n-2}(1-\alpha)$ and let $x = -\infty$. Recall that $t_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta \in \left[b - t_{n-2}(1-\alpha)\sqrt{s^2}\sqrt{c_{11}},\ +\infty\right)$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the parameter β

We have:

$$P\left(b - y\sqrt{s^2}\sqrt{c_{11}} \le \beta < b - x\sqrt{s^2}\sqrt{c_{11}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\ \%$.

Let $y = +\infty$ and let $x = t_{n-2}(\alpha)$. Recall that $t_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta \in \left(-\infty,\ b + t_{n-2}(1-\alpha)\sqrt{s^2}\sqrt{c_{11}}\right]$$

is about $1-\alpha = 95\ \%$.

Simple Linear Regression: Theorem 6

Theorem 6: The random vectors

$$(\hat{y}_i)_{i=1}^{n} = (b_0 + bx_i)_{i=1}^{n}$$

and

$$(e_i)_{i=1}^{n} = (y_i - \hat{y}_i)_{i=1}^{n}$$

are independent.

Simple Linear Regression: Theorem 7

Theorem 7:

$$\frac{\mathrm{RSS}}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2_{n-2}$$

where χ^2_{n-2} denotes Pearson's χ^2-distribution with $n-2$ degrees of freedom

where we subtract the $2 = \mathrm{rank}(\boldsymbol{X})$.

Test of hypothesis about the variance σ^2

Remark:

Theorem 7 $(RSS/\sigma^2 \sim \chi^2_{n-2})$ can be used to conduct the χ^2-test for the variance.

- Let $\sigma_0^2 \in \mathbb{R}_0^+$ be a prescribed number.
- Formulate the null hypothesis:

$$H_0: \quad \sigma^2 = \sigma_0^2$$

- Formulate the alternative hypothesis
  - two-sided: $H_1: \quad \sigma^2 \neq \sigma_0^2$
  - one-sided: $H_1: \quad \sigma^2 < \sigma_0^2$
  - one-sided: $H_1: \quad \sigma^2 > \sigma_0^2$

χ^2-test for the variance σ^2

<u>Notation:</u> Let

$$\chi^2_{n-2}(p)$$

denote the **quantile function of Pearson's χ^2-distribution** with $n-2$ d.f.

The quantile function $\chi^2_{n-2}(p)$ is the function inverse to the cumulative distribution function $F(x)$ of **Pearson's χ^2-distribution** with $n-2$ degrees of freedom, i.e.

$$\chi^2_{n-2}(p) = F^{-1}(p) \qquad \text{for} \quad p \in (0,1)$$

χ^2-test for the variance σ^2

<u>Notation:</u> Let

$$\chi^2_{n-2}(p)$$

denote the **quantile function of Pearson's χ^2-distribution** with $n-2$ d.f.

In other words, if $0 < p < 1$, then $x = \chi^2_{n-2}(p)$ is the unique value such that

$$\int_{-\infty}^{\chi^2_{n-2}(p)} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = p$$

where $f(t)$ is the density of Pearson's χ^2-distribution with $n-2$ d.f.

χ^2-test for the variance σ^2

Having chosen the value $\sigma_0^2 \in \mathbb{R}_0^+$ and assuming the null hypothesis $H_0\colon \sigma^2 = \sigma_0^2$ is true, calculate the statistic

$$X^2 = \frac{\mathrm{RSS}}{\sigma^2} = \frac{\mathrm{RSS}}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma_0^2}$$

χ^2-test for the variance σ^2

The χ^2-test for σ^2 with two-sided alternative hypothesis ($\sigma^2 \neq \sigma_0^2$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical values** are $c = \chi^2_{n-2}\left(\frac{\alpha}{2}\right)$ and $d = \chi^2_{n-2}\left(1 - \frac{\alpha}{2}\right)$
- if $X^2 \in [0, c] \cup [d, +\infty)$, **the critical region**, then **<u>reject</u>** the null hypothesis
- if $X^2 \in (c, d)$, then **<u>do not reject</u>** (or <u>fail to reject</u>) the null hypothesis

χ^2-test for the variance σ^2

The χ^2-test for σ^2 with one-sided alternative hypothesis ($\sigma^2 < \sigma_0^2$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical value** is $c = \chi^2_{n-2}(\alpha)$
- if $X^2 \in [0, c]$, **the critical region**, then **reject** the null hypothesis
- if $X^2 \in (c, +\infty)$, then **do not reject** (or fail to reject) the null hypothesis

χ^2-test for the variance σ^2

The χ^2-test for σ^2 with one-sided alternative hypothesis ($\sigma^2 > \sigma_0^2$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.
- the **critical value** is $d = \chi^2_{n-2}(1-\alpha)$
- if $X^2 \in [d, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $X^2 \in [0, d)$, then **do not reject** (or fail to reject) the null hypothesis

Confidence interval for the variance σ^2

Let $x, y \in \mathbb{R}^+$ be any numbers such that $x < y$ and let $F(x)$ be the cumulative distribution function of Pearson's χ^2-distribution with $n-2$ degrees of freedom. Then, by the definition of the cumulative distribution function and by Theorem 7, the probability

$$P\left(x < \frac{\text{RSS}}{\sigma^2} \leq y\right) = F(y) - F(x)$$

Therefore

$$P\left(\frac{\text{RSS}}{y} \leq \sigma^2 < \frac{\text{RSS}}{x}\right) = F(y) - F(x)$$

Confidence interval for the variance σ^2

We have:

$$P\left(\frac{\mathrm{RSS}}{y} \le \sigma^2 < \frac{\mathrm{RSS}}{x}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = \chi^2_{n-2}\left(1 - \frac{\alpha}{2}\right)$ and let $x = \chi^2_{n-2}\left(\frac{\alpha}{2}\right)$. Recall that $\chi^2_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \sigma^2 \in \left[\frac{\mathrm{RSS}}{\chi^2_{n-2}\left(1-\frac{\alpha}{2}\right)}, \frac{\mathrm{RSS}}{\chi^2_{n-2}\left(\frac{\alpha}{2}\right)}\right]$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the variance σ^2

We have:

$$P\left(\frac{\text{RSS}}{y} \le \sigma^2 < \frac{\text{RSS}}{x}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = \chi^2_{n-2}(1-\alpha)$ and let $x \searrow 0$. Recall that $\chi^2_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \sigma^2 \in \left[\frac{\text{RSS}}{\chi^2_{n-2}(1-\alpha)},\ +\infty\right)$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the variance σ^2

We have:

$$P\left(\frac{\mathrm{RSS}}{y} \le \sigma^2 < \frac{\mathrm{RSS}}{x}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = +\infty$ and let $x = \chi^2_{n-2}(\alpha)$. Recall that $\chi^2_{n-2}(p) = F^{-1}(p)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \sigma^2 \in \left[\, 0,\ \frac{\mathrm{RSS}}{\chi^2_{n-2}(\alpha)} \,\right]$$

is about $1-\alpha = 95\,\%$.

Simple Linear Regression: Theorem 8

Theorem 8:

$$\frac{(b_0-\beta_0 \quad b-\beta)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\begin{pmatrix} b_0-\beta_0 \\ b-\beta \end{pmatrix}}{\mathrm{RSS}} \Bigg/ \frac{2}{n-2} \;\sim\; F_{2,n-2}$$

where $F_{2,n-2}$ denotes Fisher's *F*-distribution with 2 and $n-2$ d.f.

Hypotheses about all the parameters

Remark: The theorem can be used to test the null hypothesis that

$$H_0: \quad \beta_0 = \bar{\beta}_0 \quad \text{and} \quad \beta = \bar{\beta}$$

where $\bar{\beta}_0, \bar{\beta} \in \mathbb{R}$ are some prescribed values, such as $\bar{\beta}_0 = \bar{\beta} = 0$, i.e.

$$H_0: \quad \beta_0 = \beta = 0$$

- **Be cautious** because this test actually means the test whether Y is just a random error, i.e. $Y = \varepsilon$, i.e. $Y \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 \in \mathbb{R}_0^+$ (see above).
- To conduct the test, find the critical value $c > 0$ so that $\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$, where $f(x)$ is the density of Fisher's *F*-distribution with 2 and $n-2$ d.f. The critical region is $[c, +\infty)$, i.e. reject the hypothesis if the statistic $F \in [c, +\infty)$.

The Coefficient of Determination (R^2)

The Coefficient of Determination (R^2)

Theorem:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

Proof — to see that:

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (b_0 + bx_i) = nb_0 + \sum_{i=1}^{n} bx_i =$$

$$= n\frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i b\right) + \sum_{i=1}^{n} bx_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} bx_i + \sum_{i=1}^{n} bx_i = \sum_{i=1}^{n} y_i$$

The Coefficient of Determination (R^2)

Recall the **Residual Sum of Squares:**

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Introduce the **Regression Sum of Squares:**

$$\text{RegSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Introduce the **Total Sum of Squares:**

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The Coefficient of Determination (R^2)

Theorem:

$$\text{TSS} = \text{RSS} + \text{RegSS}$$

or

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

The Coefficient of Determination (R^2)

<u>Notice first:</u>

$$\hat{y}_i = b_0 + bx_i$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i = nb_0 + b\sum_{i=1}^{n} x_i$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = b_0 + b\frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = b_0 + b\bar{x}$$

Therefore:

$$\hat{y}_i - \bar{y} = bx_i - b\hat{x}$$

The Coefficient of Determination (R^2)

Notice second:

$$\hat{\beta} = b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n\sum_{i=1}^{n} x_i x_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j} =$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j + \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j + \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j} =$$

$$= \frac{n\sum_{i=1}^{n}\left(x_i y_i - x_i \frac{\sum_{j=1}^{n} y_j}{n} - \frac{\sum_{j=1}^{n} x_j}{n} y_i + \frac{\sum_{j=1}^{n} x_j}{n}\frac{\sum_{j=1}^{n} y_j}{n}\right)}{n\sum_{i=1}^{n}\left(x_i^2 - 2x_i \frac{\sum_{j=1}^{n} x_j}{n} + \frac{\sum_{j=1}^{n} x_j}{n}\frac{\sum_{j=1}^{n} x_j}{n}\right)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The Coefficient of Determination (R^2)

Notice third:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-\hat{y}_i)(\hat{y}_i-\bar{y})=$$

$$=\sum_{i=1}^{n}(y_i-\bar{y}+b\bar{x}-bx_i)(bx_i-b\bar{x})=\sum_{i=1}^{n}(y_i-\bar{y})b(x_i-\bar{x})-b^2(\bar{x}-x_i)^2=$$

$$=b\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})-b\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(\bar{x}-x_i)^2 \quad =0$$

The Coefficient of Determination (R^2)

Notice now:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2 =$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 =$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \text{RSS} + \text{RegSS}$$

The Coefficient of Determination (R^2)

<u>**Recall:**</u>

Residual Sum of Squares: $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Regression Sum of Squares: $\text{RegSS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Total Sum of Squares: $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

and

$$\text{RSS} + \text{RegSS} = \text{TSS}$$

Then the <u>**Coefficient of Determination**</u> is

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{RegSS}}{\text{TSS}}$$

The Coefficient of Determination (R^2)

The coefficient of determination

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{RegSS}}{\text{TSS}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

is a measure how well the regression line $y = b_0 + bx$ fits the observed data

$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$

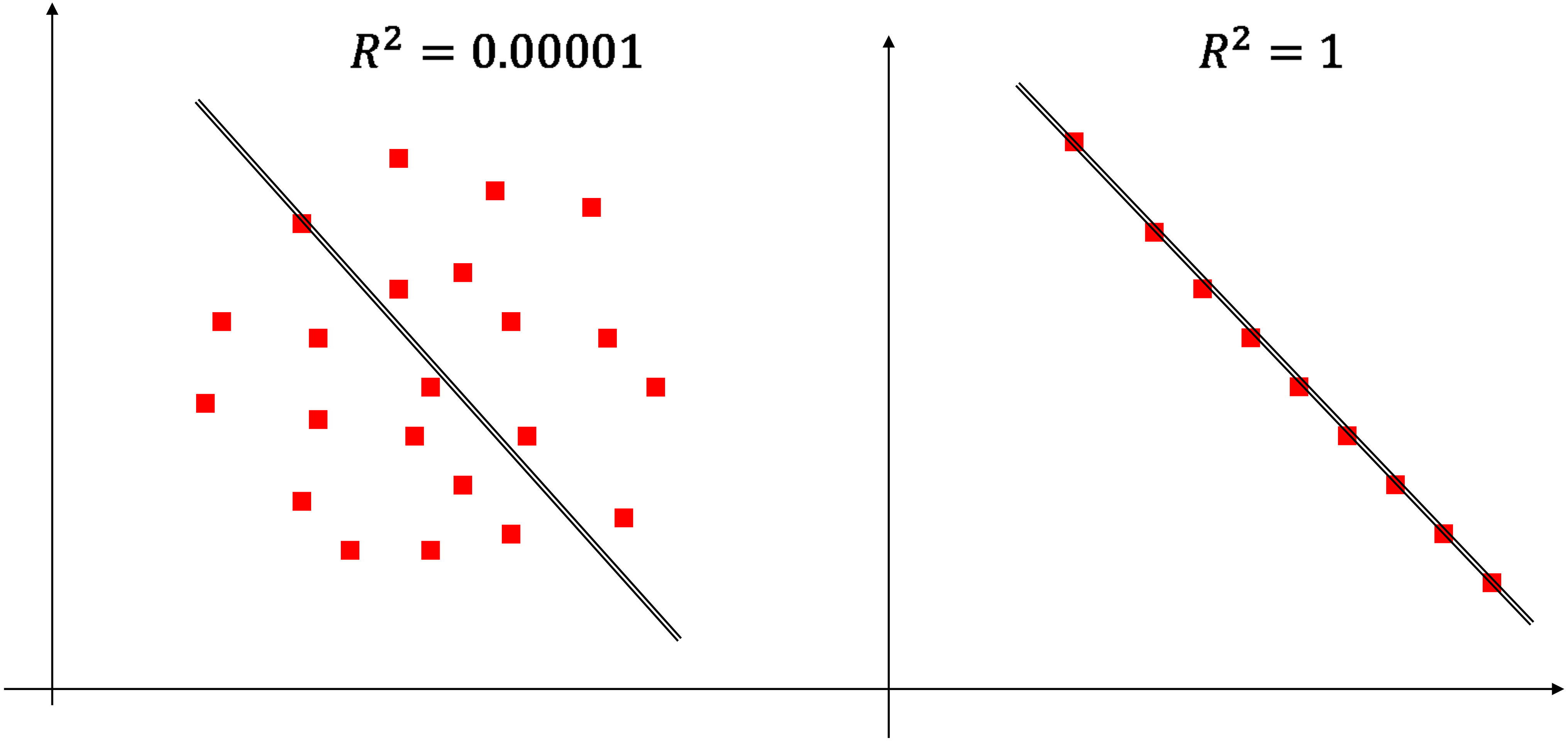
It holds

$$0 \le R^2 \le 1$$

If $R^2 \nearrow 1$, the fit is good.

If $R^2 \searrow 0$, the fit is poor.

The Coefficient of Determination (R^2)



The Coefficient of Determination (R^2): Theorem 9

Theorem 9: Under the hypothesis

$$H_0: \quad \beta = 0$$

it holds

$$\frac{\text{RegSS}}{\text{RSS}} \Big/ \frac{1}{n-2} = \frac{R^2}{1-R^2}(n-2) \sim F_{1,n-2}$$

where $F_{1,n-2}$ denotes Fisher's *F*-distribution with 1 and $n-2$ degrees of freedom.

Remark: Since, except the intercept term β_0, we have only one regression coefficient β, this *F*-test is equivalent with the *t*-test for the coefficient β (see Corollary 1 above).

F-test for the null hypothesis H_0: $\beta=0$

- Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.
- Find the **critical value** $c > 0$ so that $\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$, where f is the density of the *F*-distribution with 1 and $n-2$ degrees of freedom.
- Calculate the statistic

$$F = \frac{R^2}{1-R^2}(n-2) = \frac{\text{RegSS}}{\text{RSS}} \Big/ \frac{1}{n-2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

- If $F \in [c, +\infty)$, **the critical region**, then **<u>reject</u>** the null hypothesis.
- If $F \in [0, c)$, then **<u>do not reject</u>** (or <u>fail to reject</u>) the null hypothesis.

Two-sample *t*-test for the difference of the population means // $\sigma_X = \sigma_Y$

Two-sample *t*-test for the diff. of the pop. means // $\sigma_X = \sigma_Y$

<u>**Motivation:**</u>

We have two unknown random variables Y' and Y''. We ask (test the hypothesis) whether the population means of both random variables are the same.

We assume that both random variables are normal, i.e. $Y' \sim \mathcal{N}(\mu', \sigma'^2)$ and $Y'' \sim \mathcal{N}(\mu'', \sigma''^2)$, for some $\mu', \mu'' \in \mathbb{R}$ and for some $\sigma'^2, \sigma''^2 \in \mathbb{R}_0^+$.

Although we do not know the means μ', μ'' nor the variances σ'^2, σ''^2, we assume that

¡¡¡ ¡¡¡ ¡¡¡ $\sigma'^2 = \sigma''^2$!!! !!! !!!

Two-sample *t*-test for the diff. of the pop. means // $\sigma_X = \sigma_Y$

Having the m observations $y'_1, y'_2, \dots, y'_m$ of the random variable $Y' \sim \mathcal{N}(\mu', \sigma^2)$ and having the n observations $y''_1, y''_2, \dots, y''_n$ of the random variable $Y'' \sim \mathcal{N}(\mu'', \sigma^2)$, we formulate the **null hypothesis**:

both samples come from the same population:
the values of the population means are the same

$$H_0: \quad \mu' = \mu''$$

Recall that we do not know the true population means μ' and μ''. **We only test the hypothesis** by means of two samples of m and n measurements **with the same variance.**

Two-sample *t*-test for the diff. of the pop. means // $\sigma_X=\sigma_Y$

Now, transform the problem into the problem of linear regression: Put

$$N = m + n$$

and

$$x_i = 0 \quad \text{and} \quad y_i = y_i' \quad \text{for} \quad i = 1, 2, \dots, m$$

with

$$x_j = 1 \quad \text{and} \quad y_j = y_{j-m}'' \quad \text{for} \quad j = m+1,\, m+2,\, \dots,\, m+n$$

Consider now the model

$$Y_\ell = \mu' + (\mu'' - \mu')x_\ell + \varepsilon_\ell \quad \text{for} \quad \ell = 1, 2, \dots, N$$

i.e. the model with $\beta_0 = \mu'$ and $\beta = \mu'' - \mu'$.

Finally, conduct the *t*-test that $\beta = 0$ (see above).

Simple linear regression without the intercept term

Simple linear regression without the intercept term

Let n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of some observations be given.

We sometimes know that the dependent variable Y should be directly proportional to the independent variable x, i.e. the relationship between the values of x and Y is of the form

$$Y \approx \beta x \qquad \text{for some} \quad \beta \in \mathbb{R}$$

or rather

$$Y = \beta x + \varepsilon \qquad \text{for some} \quad \beta \in \mathbb{R}$$

where ε is a random deviation.

Notice there is no intercept term β_0 now.

Simple linear regression without the intercept term

We then proceed as before. Having got the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations, and given an estimate $b \in \mathbb{R}$, the i-th **predicted value** is

$$\hat{y}_i = bx_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

the i-th **residual** is

$$e_i = y_i - \hat{y}_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

and the **residual sum of squares** is

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (bx_i - y_i)^2$$

Simple linear regression without the intercept term

Given the n pairs (y_1, x_1), (y_2, x_2), ..., (y_n, x_n) of the observations,

find $b \in \mathbb{R}$ so that the residual sum of squares

$$\mathrm{RSS} = \sum_{i=1}^{n} (bx_i - y_i)^2 \quad \longrightarrow \quad \min$$

is minimized.

The first-order optimality condition is

$$\frac{\partial \mathrm{RSS}}{\partial b} = \sum_{i=1}^{n} 2(bx_i - y_i)x_i = 0$$

Simple linear regression without the intercept term

We have hence the equation

$$b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

the normal equation

and its solution is

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i}$$