Probability and Statistics

7 – Regression

Stefan Heiss

Technische Hochschule Ostwestfalen-Lippe
Dep. of Electrical Engineering and Computer Science

January 23, 2024

Linear Regression Equation

Definition (7.1)

A *linear regression equation* expresses a single *response variable* Y in terms of *input variables* $\vec{x} = (x_1, \ldots, x_r)$ and a *random error Z*:

x_i's without errors

$$Y = Y_{\vec{x}^{(i)}} = \underbrace{\alpha + \beta_1 x_1^{(i)} + \ldots + \beta_r x_r^{(i)}} + Z$$

The coefficients $\alpha, \beta_1, \ldots, \beta_r$ are called *regression coefficients* and Z is a random variable with:

$$E(Z) = 0$$

Linear Regression Equation

Definition (7.1)

If $r = 1$, the equation is called a *simple linear regression equation*:

$$Y = \alpha + \beta x + Z, \qquad E(Z) = 0$$

Simple Linear Regression Equation

Notation (7.2)

In the following, only the simple linear regression equation

$$Y \;=\; \alpha + \beta x + Z, \qquad E(Z) \;=\; 0$$

will be considered. Furthermore, let

$$x_1, \ldots, x_n$$

fixed input

be a finite sequence of input values and

$$Y_1, \ldots, Y_n$$

denote the corresponding response variables with $Y_i := Y_{x_i}$ for $i = 1, \ldots, n$.

Simple Linear Regression Equation

Notation (7.2)

Furthermore, the following notations will be fixed:

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s_x := \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})x_i = \left(\sum_{i=1}^{n} x_i^2\right) - n\,\overline{x}^2$$

$$\overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S_Y := \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})Y_i = \left(\sum_{i=1}^{n} Y_i^2\right) - n\,\overline{Y}^2$$

$$S_{xY} := \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \left(\sum_{i=1}^{n} x_i Y_i\right) - n\,\overline{x}\,\overline{Y}$$

Least Squares Estimators

Theorem (7.3)

The estimators A and B for α and β minimizing

$$S_R \;:=\; \sum_{i=1}^{n}(Y_i - A - Bx_i)^2$$

*for any sequence $y_1, \ldots, y_n$ of data sampled from $Y_1, \ldots, Y_n$ (*least squares estimators*) are given by:*

Least Squares Estimators

Theorem (7.3)

(i)

$$B \; := \; \frac{\left(\sum_{i=1}^{n} x_i Y_i\right) - n\,\overline{x}\,\overline{Y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\,\overline{x}^2} \; = \; \frac{S_{xY}}{s_x}$$

(ii)

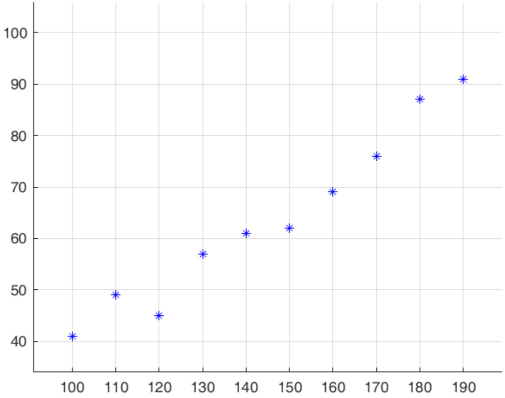
$$A \; := \; \overline{Y} \; - \; B\,\overline{x}$$

Pf.: Similar to (1.22)

Example

The following table contains 10 data pairs relating the yield of a laboratory experiment y_i to the temperature x_i at which the experiment was run.

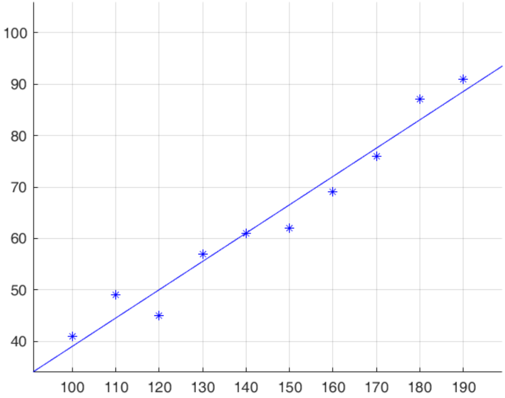
i	x_i	y_i
1	100	41
2	110	49
3	120	45
4	130	57
5	140	61
6	150	62
7	160	69
8	170	76
9	180	87
10	190	91



Example

The following table contains 10 data pairs relating the yield of a laboratory experiment y_i to the temperature x_i at which the experiment was run.

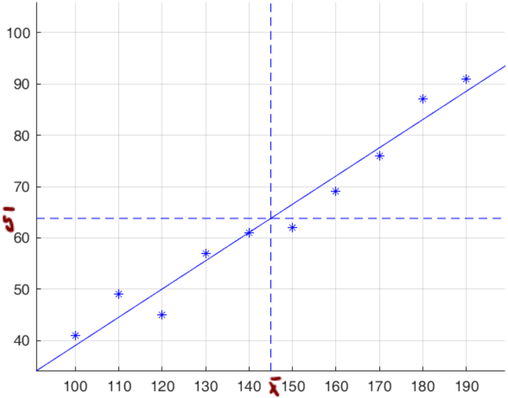
i	x_i	y_i
1	100	41
2	110	49
3	120	45
4	130	57
5	140	61
6	150	62
7	160	69
8	170	76
9	180	87
10	190	91



Example

The following table contains 10 data pairs relating the yield of a laboratory experiment y_i to the temperature x_i at which the experiment was run.

i	x_i	y_i
1	100	41
2	110	49
3	120	45
4	130	57
5	140	61
6	150	62
7	160	69
8	170	76
9	180	87
10	190	91



Least Squares Estimators

Lemma (7.4)

$$S_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2 = \frac{s_x \, S_Y - (S_{xY})^2}{s_x}$$

Proof of Lemma (7.4)

(7.3) (ii) $A = \overline{Y} - B\overline{x}$ (7.3) (i) : $B =$

$$\sum_{i=1}^{n}(Y_i - A - Bx_i)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y} + B\overline{x} - Bx_i)^2 = \sum_{i=1}^{n}\left((Y_i - \overline{Y}) + \frac{S_{xY}}{s_x}(\overline{x} - x_i)\right)^2$$

$$= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + 2\frac{S_{xY}}{s_x}\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})(\overline{x} - x_i)}_{=}\ \ (-(x_i - \overline{x})) + \left(\frac{S_{xY}}{s_x}\right)^2 \underbrace{\sum_{i=1}^{n}(\overline{x} - x_i)^2}_{=}$$

cf. with (7.2)

$$= S_Y + 2\frac{S_{xY}}{s_x}(-S_{xY}) + \left(\frac{S_{xY}}{s_x}\right)^2 s_x = S_Y - \frac{(S_{xY})^2}{s_x}$$

$$= \frac{s_x S_Y - (S_{xY})^2}{s_x}$$

Theorem (7.5)

If there exists some $\sigma > 0$, such that $Y_x \sim \mathcal{N}(\alpha + \beta x, \sigma)$ for all $x \in \mathbb{R}$, then:

(i)

$$B \;\sim\; \mathcal{N}\left(\beta,\, \sigma/\sqrt{s_x}\right)$$

(ii)

$$A \;\sim\; \mathcal{N}\left(\alpha,\, \sigma \cdot \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{n\, s_x}}\right)$$

(iii) For any x_0:

$$A \;+\; B\, x_0 \;\sim\; \mathcal{N}\left(\alpha + \beta x_0,\, \sigma \cdot \sqrt{\frac{s_x + n\,(x_0 - \overline{x})^2}{n\, s_x}}\right)$$

Proof of Theorem (7.5) (i)

$(x_i - \bar{x})(Y_i - \bar{Y}) = (x_i - \bar{x})Y_i - (x_i - \bar{x})\bar{Y}$

$$B \overset{(7.3)(i)}{=} \frac{S_{xY}}{s_x} = \frac{1}{s_x} \cdot \left(\sum_{i=1}^{n} (x_i - \bar{x})\, Y_i \; - \; \bar{Y} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{0} \right) = \frac{1}{s_x} \cdot \sum_{i=1}^{n} (x_i - \bar{x})\, Y_i$$

and B is therefore a linear combination of the independent normally distributed random variables $Y_1, \ldots, Y_n$. Hence B itself has a normal distribution with:

$$E(B) = \frac{1}{s_x} \cdot \sum_{i=1}^{n} (x_i - \bar{x})\, E(Y_i) = \frac{1}{s_x} \cdot \sum_{i=1}^{n} (x_i - \bar{x})\, (\alpha + \beta x_i)$$

$$= \frac{1}{s_x} \left(\alpha \cdot \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{0} + \beta \cdot \underbrace{\sum_{i=1}^{n} x_i (x_i - \bar{x})}_{\overset{(7.2)}{=} s_x} \right) = \frac{1}{s_x} (\beta \cdot s_x) = \beta$$

$$Var(B) = \frac{1}{{s_x}^2} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \underbrace{Var(Y_i)}_{\sigma^2} = \frac{\sigma^2 s_x}{{s_x}^2} = \frac{\sigma^2}{s_x}$$

overall assumption:

1.) X_i are independent

2.) $Y_i \sim \mathcal{N}(\ldots, \sigma)$