Probability and Statistics

7 – Regression

Stefan Heiss

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

January 26, 2024

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)

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

B is an unbiased estimator for β

(ii)

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

A is an unbiased estimator for α

(iii) For any x_0:

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

(iii) $\Rightarrow$ (ii) : (iii) with $x_0 = 0$ $\quad \sum x_i^2 = s_x + n\bar{x}^2$ $\quad$ see (1.2)

Proof of Theorem (7.5) (i)

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

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:

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

$$= \frac{1}{s_x} \left(\alpha \cdot \sum_{i=1}^{n} (x_i - \overline{x}) + \beta \cdot \sum_{i=1}^{n} x_i (x_i - \overline{x}) \right) = \frac{1}{s_x} (\beta \cdot s_x) = \beta$$

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

Proof of Theorem (7.5) (iii)

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

$\cdot (x_0 - \overline{x})$

From the proof of (i) given above, we have:

$A = \overline{Y} - B \cdot \overline{x}$ ✓

$$A + B x_0 = \overline{Y} + B (x_0 - \overline{x}) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_0 - \overline{x})(x_i - \overline{x})}{s_x} \right) Y_i$$

Proof of Theorem (7.5) (iii)

From the proof of (i) given above, we have:

$$A + B\,x_0 = \overline{Y} + B\,(x_0 - \overline{x}) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_0 - \overline{x})(x_i - \overline{x})}{s_x} \right) Y_i$$

Therefore, $A + Bx_0$ is a sum of the independent normally distributed random variables Y_i.

Hence, $A + Bx_0$ has a normal distribution with:

$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i\,, \quad Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma)$

$$E(A + B\,x_0) = E\Big(\overline{Y} + B\,(x_0 - \overline{x})\Big) = E(\overline{Y}) + E(B(x_0 - \overline{x}))$$

(i)

$$= \frac{1}{n} \sum_{i=1}^{n} (\alpha + \beta x_i) + \beta(x_0 - \overline{x}) = \alpha + \beta\overline{x} + \beta(x_0 - \overline{x})$$

$$= \alpha + \beta x_0$$

Proof of Theorem (7.5) (iii)

$$
\begin{aligned}
Var(A + Bx_0) &= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_0 - \overline{x})(x_i - \overline{x})}{s_x} \right)^2 \\
&= \frac{\sigma^2}{n^2 {s_x}^2} \sum_{i=1}^{n} (s_x + n(x_0 - \overline{x})(x_i - \overline{x}))^2 \\
&= \frac{\sigma^2}{n^2 {s_x}^2} \left(n{s_x}^2 + 2 s_x n (x_0 - \overline{x}) \overbrace{\sum_{i=1}^{n} (x_i - \overline{x})}^{0} + n^2 (\overline{x} - x_0)^2 \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{s_x} \right) \\
&= \frac{\sigma^2}{n^2 {s_x}^2} \left(n{s_x}^2 + n^2 (x_0 - \overline{x})^2 s_x \right) \\
&= \frac{\sigma^2}{n\, s_x} \left(s_x + n(x_0 - \overline{x})^2 \right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{s_x}(x_0 - \overline{x})^2
\end{aligned}
$$

Theorem (7.5)

σ not known

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

(iv) S_R is independent of A and B and:

without proof

$$S_R/\sigma^2 \sim \chi^2_{n-2}$$

(omitting the proof)

(v)

$$\sqrt{\frac{(n-2)\,s_x}{S_R}} \cdot (B - \beta) \sim t_{n-2}$$

(vi)

$$\sqrt{\frac{n(n-2)\,s_x}{S_R \sum_{i=1}^{n} x_i^2}} \cdot (A - \alpha) \sim t_{n-2}$$

Proof of Theorem (7.5) (v)

From (i) and (iv) it follows that

$$\frac{B - \beta}{\sigma / \sqrt{s_x}} \sim \mathcal{N}(0, 1)$$

and:

$$\frac{\frac{B-\beta}{\cancel{\sigma}/\sqrt{s_x}}}{\sqrt{\frac{S_R}{\cancel{\sigma^2}(n-2)}}} = \sqrt{\frac{(n-2)\, s_x}{S_R}} \cdot (B - \beta) \sim t_{n-2}$$

$\sim \chi^2_{n-2}$

Proof of Theorem (7.5) (vi)

From (ii) and (iv) it follows that

$$\frac{A - \alpha}{\frac{\sigma \sqrt{\sum_{i=1}^{n} {x_i}^2}}{\sqrt{n\, s_x}}} \sim \mathcal{N}(0,1)$$

and:

$$\frac{\sqrt{n\, s_x} \cdot \frac{A-\alpha}{\sigma \sqrt{\sum_{i=1}^{n} {x_i}^2}}}{\sqrt{\frac{S_R}{\sigma^2\,(n-2)}}} = \sqrt{\frac{n(n-2)\, s_x}{S_R \sum_{i=1}^{n} {x_i}^2}} \cdot (A - \alpha) \sim t_{n-2}$$

Theorem (7.5)

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

(vii) For any x_0:

$$\sqrt{\frac{n(n-2)\, s_x}{S_R\,(s_x + n(x_0 - \overline{x})^2)}} \cdot \big((A + B\,x_0) - (\alpha + \beta\,x_0)\big) \quad \sim \quad t_{n-2}$$

(viii) If Y_0 denotes the response for the input value x_0, then $Y_0 \sim \mathcal{N}(\alpha + \beta x_0,\, \sigma)$ and:

$$\sqrt{\frac{n(n-2)\, s_x}{S_R\,((n+1)s_x + n(x_0 - \overline{x})^2)}} \cdot \big(Y_0 \;-\; (A + B\,x_0)\big) \quad \sim \quad t_{n-2}$$

Proof of Theorem (7.5) (vii)

From (iii) and (iv) it follows that

$$\frac{A + B\,x_0 \;-\; (\alpha + \beta\,x_0)}{\sigma\,\sqrt{\frac{s_x + n\,(x_0 - \overline{x})^2}{n\,s_x}}} \;\sim\; \mathcal{N}(0,1)$$

and:

$$\frac{\frac{A+Bx_0 \;-\; (\alpha+\beta\,x_0)}{\sigma\,\sqrt{\frac{s_x + n\,(x_0 - \overline{x})^2}{n\,s_x}}}}{\sqrt{\frac{S_R}{\sigma^2\,(n-2)}}} \;=\; \sqrt{\frac{n(n-2)\,s_x}{S_R(s_x \;+\; n(x_0 - \overline{x})^2)}}\cdot\left(A + B\,x_0 \;-\; (\alpha + \beta\,x_0)\right) \;\sim\; t_{n-2}$$

Proof of Theorem (7.5) (viii)

$$Var(A + Bx_0) = \frac{\sigma^2}{n\, s_x}\left(s_x + n(x_0 - \overline{x})^2\right)$$

$+\ \sigma^2$

$$Y_0 \sim \mathcal{N}(\alpha + \beta x_0, \sigma)$$

is independent of $Y_1, \ldots, Y_n$ (and consequently independent of $A + Bx_0$). It follows from (iii):

$$Y_0 - (A + Bx_0) \sim \mathcal{N}\left(0, \sqrt{\frac{\sigma^2}{n\, s_x}\left(ns_x + s_x + n(x_0 - \overline{x})^2\right)}\right)$$

$$= \mathcal{N}\left(0, \sigma \cdot \sqrt{\frac{(n+1)s_x + n(x_0 - \overline{x})^2}{n\, s_x}}\right)$$

and:

$$\frac{\dfrac{Y_0 - (A+Bx_0)}{\sigma\sqrt{\frac{(n+1)s_x + n\,(x_0 - \overline{x})^2}{n\, s_x}}}}{\sqrt{\dfrac{S_R}{\sigma^2\,(n-2)}}} = \sqrt{\frac{n(n-2)\, s_x}{S_R\left((n+1)s_x + n(x_0 - \overline{x})^2\right)}} \cdot \left(Y_0 - (A + B\, x_0)\right) \sim t_{n-2}$$

Notation (7.6)

Since the Greek character α is already used to denote one of the constants in our simple linear regression equation, we will use $1-\gamma$ and γ in the following to denote confidence and significance levels, respectively.

(7.7) Confidence interval for β with confidence level $1-\gamma$

Let

$$\delta_t \quad := \quad F^{-1}_{t_{n-2}}\left(1-\frac{\gamma}{2}\right)$$

Then:

$$\Pr\left(\left|\sqrt{\frac{(n-2)\,s_x}{S_R}}\cdot(B-\beta)\right| \;\leq\; \delta_t\right) \;=\; F_{t_{n-2}}(\delta_t) \;-\; F_{t_{n-2}}(-\delta_t)$$

$$= \; 1-\frac{\gamma}{2} \;-\; \left(1-\left(1-\frac{\gamma}{2}\right)\right) \;=\; 1-\gamma$$

(7.7) Confidence interval for β with confidence level $1-\gamma$

Given sampled data values $\mathrm{y} = (y_1, \ldots, y_n)$ from $(Y_1, \ldots, Y_n)$, the value $B(\mathrm{y})$ and $S_R(\mathrm{y})$ can be calculate (substituting the Y_i's in the definition of B and S_R by the sampled values in y) and with a confidence of level $1-\gamma$:

$$\left|\sqrt{\frac{(n-2)\,s_x}{S_R(\mathrm{y})}} \cdot (B(\mathrm{y}) - \beta)\right| \leq \delta_t = F^{-1}_{t_{n-2}}\left(1 - \frac{\gamma}{2}\right)$$

$$\Longleftrightarrow \quad |B(\mathrm{y}) - \beta| \leq \sqrt{\frac{S_R(\mathrm{y})}{(n-2)\,s_x}} \cdot \delta_t = \sqrt{\frac{S_R(\mathrm{y})}{(n-2)\,s_x}} \cdot F^{-1}_{t_{n-2}}\left(1 - \frac{\gamma}{2}\right)$$

$$\Longleftrightarrow \quad \beta = B(\mathrm{y}) \pm \sqrt{\frac{S_R(\mathrm{y})}{(n-2)\,s_x}} \cdot F^{-1}_{t_{n-2}}\left(1 - \frac{\gamma}{2}\right)$$

(7.8) Hypothesis test of $H_0 : \beta = \beta_0$

Given a significance level γ and sampled data values $\mathrm{y} = (y_1, \ldots, y_n)$ from $(Y_1, \ldots, Y_n)$:

- H_0 is rejected if $\sqrt{\dfrac{(n-2)\, s_x}{S_R(\mathrm{y})}} \cdot |B(\mathrm{y}) - \beta_0| \; > \; F^{-1}_{t_{n-2}}\Big(1 - \dfrac{\gamma}{2}\Big)$
- H_0 is accepted if $\sqrt{\dfrac{(n-2)\, s_x}{S_R(\mathrm{y})}} \cdot |B(\mathrm{y}) - \beta_0| \; \leq \; F^{-1}_{t_{n-2}}\Big(1 - \dfrac{\gamma}{2}\Big)$

The p–value is:

$$\gamma_{\mathrm{y}} \; := \; 2\left(1 - F_{t_{n-2}}\left(\sqrt{\frac{(n-2)\, s_x}{S_R(\mathrm{y})}} \cdot |B(\mathrm{y}) - \beta_0|\right)\right)$$

(7.9) Confidence interval for $\alpha + \beta x_0$

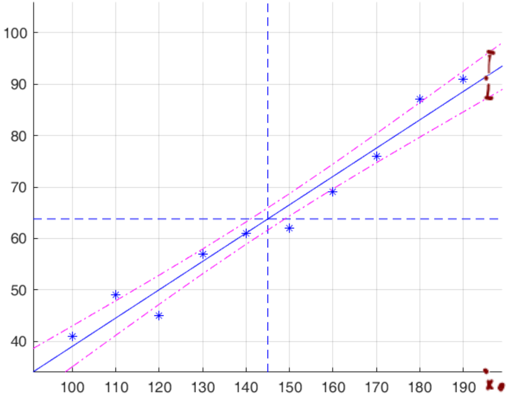
Given an input value x_0 and sampled data values $\mathrm{y} = (y_1, \dots, y_n)$ from $(Y_1, \dots, Y_n)$, we have with a confidence of $100(1-\gamma)\,\%$:

$$\alpha + \beta\, x_0 \;=\; A(\mathrm{y}) + B(\mathrm{y})\, x_0 \;\pm\; \sqrt{\frac{S_R(\mathrm{y}) \cdot (s_x + n(x_0 - \overline{x})^2)}{n(n-2)\, s_x}} \cdot F_{t_{n-2}}^{-1}\left(1 - \frac{\gamma}{2}\right)$$

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



conf. level of 50%

(7.10) Prediction interval for a response at input value x_0

For a given input value x_0 and sampled data values $\mathrm{y} = (y_1, \ldots, y_n)$ from $(Y_1, \ldots, Y_n)$ a $100(1-\gamma)$ percent confidence interval for the response value is given by:

$$A(\mathrm{y}) + B(\mathrm{y})\, x_0 \quad \pm \sqrt{\frac{S_R(\mathrm{y}) \cdot ((n+1)s_x + n(x_0 - \overline{x})^2)}{n(n-2)\, s_x}} \cdot F^{-1}_{t_{n-2}}\left(1 - \frac{\gamma}{2}\right)$$

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