7 Regression

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

$$Y \;=\; Y_{\mathbf{x}} \;=\; \alpha \,+\, \beta_1 x_1 \,+\, \ldots \,+\, \beta_r x_r \,+\, Z$$

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

$$E(Z) \;=\; 0$$

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

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

(7.2) Notation. In the following, only the simple linear regression equation above will be considered. Furthermore, let

$$x_1, \ldots, x_n$$

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 $i = 1, \ldots, n$.

Furthermore, the following notations will be fixed:

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

(7.3) Theorem. 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:

(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)

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

Proof: See (1.22). □

(7.4) Lemma.

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

Proof:

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

□

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

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

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

(iv) S_R is independent of A and B and:
$$S_R/\sigma^2 \sim \chi^2_{n-2}$$

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

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

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

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

Proof:

(i) By (7.3)(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:

$$\begin{aligned} E(B) &= \frac{1}{s_x} \cdot \sum_{i=1}^{n} (x_i - \overline{x})\, E(Y_i) \;=\; \frac{1}{s_x} \cdot \sum_{i=1}^{n} (x_i - \overline{x})\,(\alpha + \beta x_i) \\ &= \frac{1}{s_x} \left(\alpha \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \;+\; \beta \cdot \sum_{i=1}^{n} x_i (x_i - \overline{x}) \right) \;=\; \frac{1}{s_x}\,(\beta \cdot s_x) \;=\; \beta \end{aligned}$$

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

(ii) Follows from (iii) with $x_0 = 0$ and:

$$s_x \;+\; n\,\overline{x}^2 \;\overset{(7.2)}{=}\; \sum_{i=1}^{n} {x_i}^2$$

(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:

$$\begin{aligned} E(A \;+\; B\,x_0) &= E\Big(\overline{Y} \;+\; B\,(x_0 - \overline{x})\Big) \;=\; E(\overline{Y}) \;+\; E(B(x_0 - \overline{x})) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\alpha + \beta x_i) \;+\; \beta(x_0 - \overline{x}) \;=\; \alpha \;+\; \beta \overline{x} \;+\; \beta(x_0 - \overline{x}) \\ &= \alpha + \beta x_0 \end{aligned}$$

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

(iv) omitted

(v) From (i) and (iv) it follows that

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

and:

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

(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:

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

(vii) From (iii) and (iv) it follows that

$$\frac{A + Bx_0 - (\alpha + \beta\, x_0)}{\sigma\sqrt{\frac{s_x + n\,(x_0 - \overline{x})^2}{n\, s_x}}} \sim \mathcal{N}(0,1)$$

and:

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

(viii) As

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

is independent of $Y_1, \ldots, Y_n$ (and consequently independent of $A + Bx_0$), it follows from (iii) that $Y_0 - (A + Bx_0)$ has a normal distribution with

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

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

and:

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

□

(7.6) Notation. 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

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

Then:

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

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

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

$$\left|\sqrt{\frac{(n-2)\,s_x}{S_R(\mathbf{y})}}\cdot(B(\mathbf{y})-\beta)\right| \;\le\; \delta_t$$

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

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

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

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

- H_0 is rejected if $\quad \sqrt{\frac{(n-2)\,s_x}{S_R(\mathbf{y})}}\cdot|B(\mathbf{y})-\beta_0| \;>\; F^{-1}_{t_{n-2}}\left(1-\frac{\gamma}{2}\right)$
- H_0 is accepted if $\quad \sqrt{\frac{(n-2)\,s_x}{S_R(\mathbf{y})}}\cdot|B(\mathbf{y})-\beta_0| \;\le\; F^{-1}_{t_{n-2}}\left(1-\frac{\gamma}{2}\right)$

The p–value is:

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

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

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

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

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

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

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

(7.11) 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

The following MatLab script generates

(i) a scatter diagram for the data pairs,

(ii) the line of best fit,

(iii) the endpoints of $90\,\%$ confidence intervals for $\alpha + \beta x_0$ for all x_0 in the range of the displayed x-values,

(iv) the endpoints of $90\,\%$ prediction intervals for responses for all input values x_0 in the range of the displayed x-values.

The plot generated with the MatLab-script is shown in Figure 26.

```
%% -------------------------------------------------------------------
%  Scatter plot
%  -------------------------------------------------------------------
clear;
clc;
format compact;
```

```
x = 100:10:190;
y = [41 49 45 57 61 62 69 76 87 91];

scatter(x,y,'b*')
grid on;

% add some margin to the plot
d = 0.1;
a = axis();
w = a(2)-a(1);
a(1) = a(1)-d*w;
a(2) = a(2)+d*w;
w = a(4)-a(3);
a(3) = a(3)-d*w;
a(4) = a(4)+d*w;
axis(a);

pause
%% ---------------------------------------------------------------
%  Line of best fit
%  ---------------------------------------------------------------
hold on

n = length(x)
x_bar = mean(x)
s_x = x*x' - n*x_bar^2
y_bar = mean(y)
s_xy = x*y' - n*x_bar*y_bar

B = s_xy/s_x
A = y_bar - B*x_bar

x = [a(1) x a(2)];
plot(x,A + B*x,'b')

pause
%% ---------------------------------------------------------------
%  Lines through mean values: x = x_bar, y = y_bar
%  ---------------------------------------------------------------
plot([x_bar, x_bar], [a(3), a(4)], 'b--');
plot([a(1), a(2)], [y_bar, y_bar], 'b--');

pause
%% ---------------------------------------------------------------
%  Confidence intervals for expected response values
%  (mean values of response values)
%  for confidence level 1-gamma
%  ---------------------------------------------------------------
gamma = 0.1;

s_y = y*y' - n*y_bar^2
s_R = (s_x*s_y - s_xy^2)/s_x

delta = sqrt(s_R*(s_x + n*(x-x_bar).^2)/(n*(n-2)*s_x))...
    * tinv(1-gamma/2,n-2);
plot(x,A + B*x + delta, '-. m')
```

```
plot(x,A + B*x - delta, '-. m')

pause
%% ------------------------------------------------------------------
%  Prediction intervals for response values
%  for confidence level 1-gamma
%  ------------------------------------------------------------------
delta = sqrt(s_R*((n+1)*s_x + n*(x-x_bar).^2)/(n*(n-2)*s_x))...
    * tinv(1-gamma/2,n-2);
plot(x,A + B*x + delta, '-. r')
plot(x,A + B*x - delta, '-. r')
```

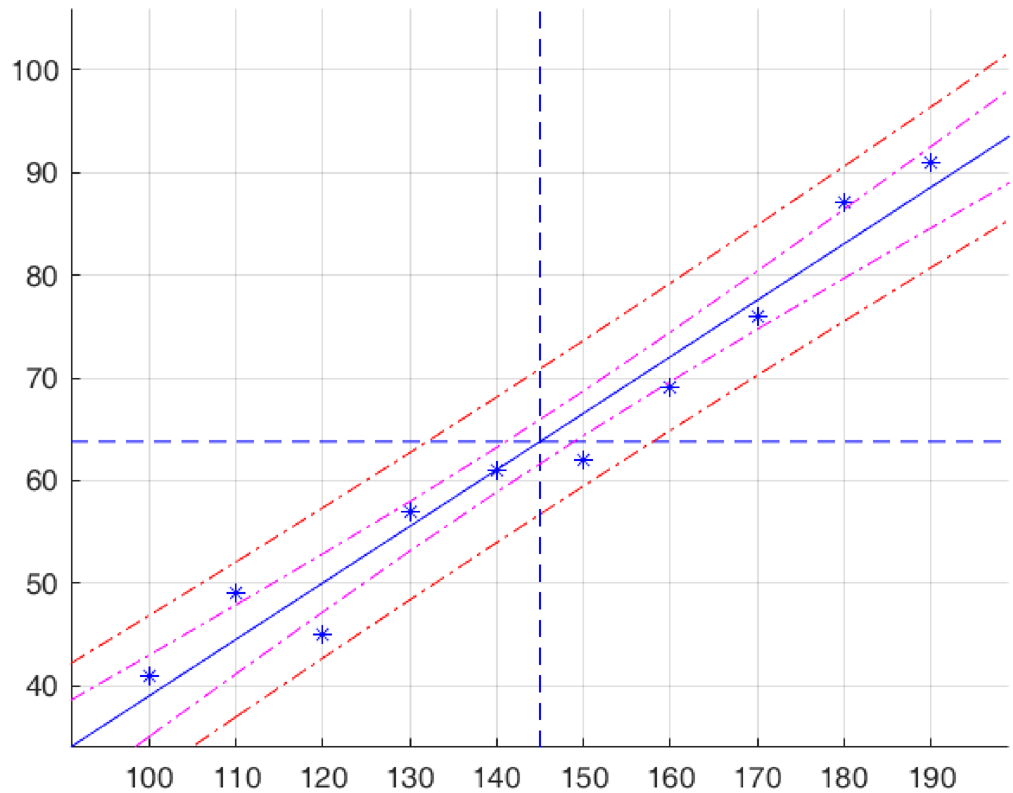

Figure 26: Scatter diagram, line of best fit, bounds for 90 % confidence intervals for $\alpha + \beta x$ and bounds for 90 % prediction intervals for responses