10. seminar

Problem 1: There are 15 workers in a manufactory. For any of them the number of work-shifts (variable X) and number of final products (variable Y) were recorded.

X: 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15
Y: 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81

a) Under the assumption that the regression line represents the dependence Y on X design the matrix of regressors, calculate the least square estimators of regression parameters and provide the sample regression function.
b) Find the estimator of variance σ^2 and the coefficient of determination and interpret it.
To make it easier for **b** = $\begin{pmatrix} 5,010 \\ 4,302 \end{pmatrix}$ the following statistics are calculated: S_E = **e'e** = 238,5169 and s^2_y = 121,4 (s^2_y is the realization of the sample variance for Y).
c) Find 95% confidence interval for regression parameters

d) At the significance level 0,05 carry out the overall F-test.

e) At the significance level 0,05 carry out the separate t-tests.

f) For 18 work-shifts estimate the number of final products.

g) Give a scatter plot with sample regression function .

Solution:
ad a) Matrix **X** (15x2) is formed of column of units and second column of values of X.
LSE of regession parameters are obtained using formula $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ where

X' X = $\begin{pmatrix} 15 & 274 \\ 274 & 5084 \end{pmatrix}$; (**X' X**)$^{-1}$ = $\begin{pmatrix} 4,2939 & -0,2314 \\ -0,2314 & 0,0127 \end{pmatrix}$; **X'y** = $\begin{pmatrix} 1254 \\ 23246 \end{pmatrix}$

thus **b** = $\begin{pmatrix} 4,2939 & -0,2314 \\ -0,2314 & 0,0127 \end{pmatrix} \cdot \begin{pmatrix} 1254 \\ 23246 \end{pmatrix} = \begin{pmatrix} 5,010 \\ 4,302 \end{pmatrix}$.

Hence the sample regression line is expressed as $\hat{y} = 5,010 + 4,302 \cdot x$

ad b)
The estimator of variance σ^2 follows: $s^2 = \frac{S_E}{n-p-1} = \frac{238,5169}{15-1-1} = 18,347$.

m_y = 83,6; s^2_y =121,4
S_T = (n-1). s^2_y =14.121,4 = 1699,6
S_R = S_T – S_E = 1699,6 – 238,5169 = 1461,0831.
The coefficient of determination follows: $ID = \frac{S_R}{S_T} = \frac{1461,0831}{1699,6} = 0,8597$.

Thus 85,97% of the variation of Y can be explained by the regression line.

ad c) To form the confidence interval we have to find the standard errors estimates s_{b_j}.
The needed diagonal elements of the matrix (**X' X**)$^{-1}$ follows:
v_{00} = 4,2939 and v_{11} = 0,0127. Thus
$s_{b_0} = s\sqrt{v_{00}} = \sqrt{18,347 \cdot 4,2939} = 8,875$ and
$s_{b_1} = s\sqrt{v_{11}} = \sqrt{18,347 \cdot 0,0127} = 0,482$.

Then the limits for the 95% confidence intervals for regression parameters β_0 and β_1 are derived from the formula: $b_j \pm t_{1-\alpha/2}(n-p-1) s_{b_j}$, j = 0, 1.

- For β_0 the limits are calculated as follows:

$$d = b_0 - t_{0,975}(13) s_{b_0} = 5,0101 - 2,1604 \cdot 8,8759 = -14,165$$
$$h = b_0 + t_{0,975}(13) s_{b_0} = 5,0101 + 2,1604 \cdot 8,8759 = 24,185$$

thus -14,1654 < β_0 < 24,1456 with the probability 95%.

- For β_1 the limits are calculated as follows:

$$d = b_1 - t_{0,975}(13) s_{b_1} = 4,3024 - 2,1604 \cdot 0,4827 = 3,259$$
$$h = b_1 + t_{0,975}(13) s_{b_1} = 4,3024 + 2,1604 \cdot 0,4827 = 5,345$$

thus 3,2596 < β_1 < 5,3452 with the probability 95%.

ad d) Carrying out the overall F-test we are testing
H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ at the significance level $\alpha = 0{,}05$.

The realization of the test statistic $F = \frac{S_R/p}{S_E/(n-p-1)}$ can be found in the last column of the following ANOVA table :

zdroj variab.	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R = 1461,0831	p = 1	S_R/p=1461,0831	79,6341
reziduální	S_E = 238,5169	n-p-1 = 13	S_E/(n-p-1)=18,3475	-
celkový	S_T = 1699,6	n-1 = 14	-	-

thus $F = 79,634$ and the critical region has the form

$$W = \langle F_{1-\alpha}(p, n-p-1), \infty) = \langle F_{0,95}(1,13), \infty) = \langle 4,6672, \infty).$$

Since $F \in W$ we reject the null at 0.05; thus the parameter β_1 (the slope) is relevant in our model.

ad e) Carrying out the separate t-tests we are testing
I. H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ at the significance level $\alpha = 0{,}05$.

The realization of the test statistic follows: $t_0 = \frac{b_0}{s_{b_0}} = \frac{5,0101}{8,8759} = 0,564$,

and the critical region has the form:

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0,975}(13)\rangle \cup \langle t_{0,975}(13), \infty) = $$
$$= (-\infty, -2,1604\rangle \cup \langle 2,1604, \infty)$$

Since $t_0 \notin W$ we do not reject the null at 0.05; thus the parameter β_0 is not relevant in our model.

II. H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ at the significance level $\alpha = 0{,}05$.

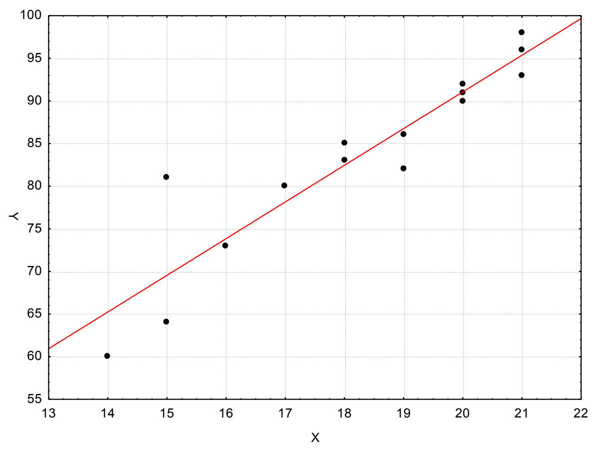
The realization of the test statistic follows: $t_1 = \frac{b_1}{s_{b_1}} = \frac{4,3024}{0,4827} = 8,913$,

and the critical region has the form:

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0,975}(13)\rangle \cup \langle t_{0,975}(13), \infty) = $$
$$= (-\infty, -2,1604\rangle \cup \langle 2,1604, \infty).$$

Since $t_1 \in W$ we reject the null at 0.05; thus the parameter β_1 is relevant in our model.
In case of the regression line the t-test for β_1 is equivalent with overall F-test.
ad f) for x = 18 the regression estimate follows: $y = 5,0101 + 4,3024 \cdot 18 = 82,45$.

ad g)



Problem 2.: Considering a car Škoda 120, the petrol consumption (1 liter/100 km) is dependent on the speed (km/hour).

rychlost	40	50	60	70	80	90	100	110
spotřeba	5,7	5,4	5,2	5,2	5,8	6,0	7,5	8,1

a) Give a scatter plot for the data and suggest the form of regression function.

b) Design the matrix of regressors, calculate the least square estimators of regression parameters, find the estimator of variance σ^2 find the coefficient of determination and interpret it.

c) Find 95% confidence interval for regression parameters

d) At the significance level 0,05 carry out the overall F-test.

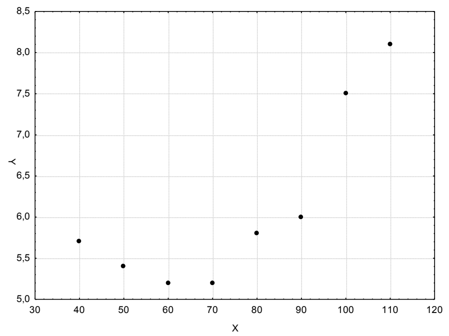
e) At the significance level 0,05 carry out the separate t-tests.

f) For the speed 80 km/hour estimate the petrol consumption.

g) Give a scatter plot with sample regression function .

Solution

ad a)



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

ad b)

$$\mathbf{X} = \begin{pmatrix} 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 70 & 4900 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \\ 1 & 100 & 10000 \\ 1 & 110 & 12100 \end{pmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 9,75178 \\ -0,15053 \\ 0,00124 \end{pmatrix}.$$

$y = 9,751786 - 0,150536x + 0,001244x^2$.

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{pmatrix} 5,7207 \\ 5,3349 \\ 5,1980 \\ 5,3098 \\ 5,6705 \\ 6,2799 \\ 7,1381 \\ 8,2452 \end{pmatrix}; \quad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -0,0207 \\ 0,0650 \\ 0,0019 \\ -0,1098 \\ 0,1294 \\ -0,2799 \\ 0,3618 \\ -0,1452 \end{pmatrix}; \quad S_E = \mathbf{e}'\mathbf{e} = 0,263869.$$

$$s^2 = \frac{S_E}{n-p-1} = \frac{0,263869}{8-2-1} = 0,0527.$$

$S_T = (\mathbf{y} - \mathbf{m}_2)'(\mathbf{y} - \mathbf{m}_2)$, where $\mathbf{m}_2$ is a column vector (nx1) of m_2 (sample mean of Y);
$m_2 = 6,1125$. $S_T = 8,32875$. (Or it can be calculated: $S_T = (n-1).\ s^2_y$)
$S_R = S_T - S_E = 8,32875 - 0,263869 = 8,06488$.

$$ID = \frac{S_R}{S_T} = \frac{8,06488}{8,32875} = 0,968.$$

ad c)

I for β_0:

$$d = b_0 - t_{0,975}(5)s_{b_0} = 9,751786 - 2,5706 \cdot 0,945689 = 7,320$$
$$h = b_0 + t_{0,975}(5)s_{b_0} = 9,751786 + 2,5706 \cdot 0,945689 = 12,182$$

II for β_1:

$$d = b_1 - t_{0,975}(5)s_{b_1} = -0,150536 - 2,5706 \cdot 0,026821 = -0,219$$
$$h = b_1 + t_{0,975}(5)s_{b_1} = -0,150536 + 2,5706 \cdot 0,026821 = -0,081$$

III for β_2:

$$d = b_2 - t_{0,975}(5)s_{b_2} = 0,001244 - 2,5706 \cdot 0,000170 = 0,000\ldots$$
$$h = b_2 + t_{0,975}(5)s_{b_2} = 0,001244 + 2,5706 \cdot 0,000170 = 0,001\ldots$$

ad d) F-test; $\alpha = 0,05$ H_0: $(\beta_1, \beta_2) = (0, 0)$ versus H_1: $(\beta_1, \beta_2) \neq (0, 0)$.

$$F = \frac{S_R/p}{S_E/(n-p-1)} = \frac{8,06488/2}{0,263869/(8-2-1)} = 76,4,$$
$$W = \langle F_{1-\alpha}(p, n-p-1), \infty) = \langle F_{0,95}(2,5), \infty) = \langle 19,296; \infty).$$

zdroj variab.	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R = 8,06488	p = 2	S_R/p=4,03244	76,41
reziduální	S_E = 0,263869	n-p-1 = 5	$S_E/(n-p-1)$=0,05277	-
celkový	S_T = 8,32875	n-1 = 7	-	-

ad e) t-tests; $\alpha = 0{,}05$

I for β_0: H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$.

$$t_0 = \frac{b_0}{s_{b_0}} = \frac{9{,}751786}{0{,}945689} = 10{,}311,$$

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)) \cup (t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0{,}975}(5)) \cup (t_{0{,}975}(5), \infty) = = (-\infty, -2{,}5706) \cup (2{,}5706, \infty).$$

II for β_1: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.

$$t_1 = \frac{b_1}{s_{b_1}} = \frac{-0{,}150536}{0{,}026821} = -5{,}612,$$

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)) \cup (t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0{,}975}(5)) \cup (t_{0{,}975}(5), \infty) = = (-\infty, -2{,}5706) \cup (2{,}5706, \infty).$$

III for β_2: H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$.

$$t_2 = \frac{b_2}{s_{b_2}} = \frac{0{,}001244}{0{,}000177} = 7{,}028,$$

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)) \cup (t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0{,}975}(5)) \cup (t_{0{,}975}(5), \infty) = = (-\infty, -2{,}5706) \cup (2{,}5706, \infty).$$

ad f) for x = 80 the regression estimate follows:

$$\hat{y} = 9{,}751786 - 0{,}150536 \cdot 80 + 0{,}001244 \cdot 80^2 = 5{,}67.$$

ad g)