Příklady na cvičení ke 13. přednášce

Příklad 1.: Porovnání koeficientu korelace s danou konstantou
Pro náhodný výběr rozsahu 50 z dvourozměrného normálního rozložení se skutečným koeficientem korelace ρ byl vypočten výběrový koeficient korelace $r_{12} = 0,5$. Na asymptotické hladině významnosti 0,05 testujte hypotézu H_0: $\rho = 0,6$ proti H_1: $\rho \neq 0,6$. Test proveďte pomocí kritického oboru i pomocí p-hodnoty.
Výsledek:
z = 0,5493, realizace u = -1,028.
Testování pomocí kritického oboru:
kritický obor $W = \left(-\infty, -u_{0,975}\right\rangle \cup \left\langle u_{0,975}, \infty\right) = \left(-\infty, -1,96\right\rangle \cup \left\langle 1,96, \infty\right)$, tedy H_0 nezamítáme na asymptotické hladině významnosti 0,05.
Testování pomocí p-hodnoty: p = 0,303, tedy H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Příklad 2.: Porovnání dvou koeficientů korelace
Jsou dány dva nezávislé náhodné výběry o rozsazích n = 35, $n^* = 40$, první pochází z dvourozměrného normálního rozložení s koeficientem korelace ρ, druhý pochází z dvourozměrného normálního rozložení s koeficientem korelace ρ^*. Výběrový koeficient korelace 1. výběru nabyl hodnoty $r_{12} = 0,4$, 2. výběru $r_{12}^* = 0,55$. Na asymptotické hladině významnosti 0,05 testujte H_0: $\rho = \rho^*$ proti H_1: $\rho \neq \rho^*$. Test proveďte pomocí kritického oboru i pomocí p-hodnoty.
Výsledek:
z = 0,423649, $z^* = 0,618381$, u = -0,8067.
Testování pomocí kritického oboru:
kritický obor $W = \left(-\infty, -u_{0,975}\right\rangle \cup \left\langle u_{0,975}, \infty\right) = \left(-\infty, -1,96\right\rangle \cup \left\langle 1,96, \infty\right)$, tedy H_0 nezamítáme na asymptotické hladině významnosti 0,05.
Testování pomocí p-hodnoty: p = 0,418, tedy H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Příklad 3.: Regresní přímka
V dílně pracuje 15 dělníků, u nichž byl zjištěn počet směn odpracovaných za měsíc (proměnná X) a počet zhotovených výrobků (proměnná Y).

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) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

b) Za předpokladu, že regresní přímka dobře vystihuje závislost Y na X, sestavte regresní matici, vypočtěte odhady regresních parametrů a napište rovnici regresní přímky.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

g) Vypočtěte regresní odhad počtu výrobků pro 18 odpracovaných směn.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Řešení:
ad a) Orientační ověření dvourozměrné normality, výpočet výběrového koeficientu korelace a testování hypotézy o nezávislosti veličin X a Y bylo provedeno v příkladu 5 ve cvičení k 12. přednášce, nyní se soustředíme na regresní analýzu.

ad b) Sestavíme regresní matici **X** typu 15x2, která má v 1. sloupci samé jedničky a ve 2. sloupci hodnoty proměnné X. Podle vzorce $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ získáme odhady regresních parametrů. Nejprve vypočítáme matici $\mathbf{X'X} = \begin{pmatrix} 15 & 274 \\ 274 & 5084 \end{pmatrix}$ a k ní inverzní matici $(\mathbf{X'X})^{-1} = \begin{pmatrix} 4,2939 & -0,2314 \\ -0,2314 & 0,0127 \end{pmatrix}$. Dále získáme součin $\mathbf{X'y} = \begin{pmatrix} 1254 \\ 23246 \end{pmatrix}$ a nakonec

$\mathbf{b} = \begin{pmatrix} 4,2939 & -0,2314 \\ -0,2314 & 0,0127 \end{pmatrix} . \begin{pmatrix} 1254 \\ 23246 \end{pmatrix} = \begin{pmatrix} 5,0101 \\ 4,3024 \end{pmatrix}$.

Regresní přímka má tedy rovnici y = 5,0101 + 4,3024 x.

ad c) Nyní vypočteme vektor regresních odhadů proměnné Y (vektor predikce):
$\hat{\mathbf{y}} = \mathbf{Xb}$ = (91,0574 95,3598 82,4527 78,1503 91,0574 82,4527 86,7551 95,3598 91,0574 65,2432 73,8480 86,7551 95,3598 69,5456 69,5456)‘.
Stanovíme vektor reziduí:
$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ = (0,9426 -2,3598 0,5473 1,8497 -0,0574 2,5473 -4,7551 2,6402 -1,0574 -5,2432 -0,8480 -0,7551 0,6402 -5,5456 11,4544)‘.
Pomocí vektoru reziduí vypočteme reziduální součet čtverců: $S_E = \mathbf{e'e} = 238,5169$.
Odhad rozptylu: $s^2 = \frac{S_E}{n-p-1} = \frac{238,5169}{15-1-1} = 18,3475$.
Dále potřebujeme celkový součet čtverců $S_T = (\mathbf{y} - \mathbf{m}_2)'(\mathbf{y} - \mathbf{m}_2)$, kde $\mathbf{m}_2$ je sloupcový vektor typu nx1 složený z průměru m_2 závisle proměnné veličiny Y. V našem případě je $m_2 = 83,6$. Po dosazení do vzorce pro celkový součet čtverců tedy dostaneme $S_T = 1699,6$. (Celkový součet čtverců lze získat také tak, že výběrový rozptyl veličiny Y vynásobíme n-1: $S_T = 14.121,4 = 1699,6$.) Regresní součet čtverců pak je: $S_R = S_T - S_E = 1699,6 - 238,5169 = 1461,0831$.
Index determinace: $ID^2 = \frac{S_R}{S_T} = \frac{1461,0831}{1699,6} = 0,8597$.
Znamená to, že variabilita hodnot závisle proměnné veličiny je z 85,97% vysvětlena regresní přímkou.
(V případě regresní přímky platí $ID^2 = r_{12}^2$. V našem případě bylo zjištěno, že $r_{12} = 0,9272$, tedy $ID^2 = 0,8597$.)

ad d) Vypočteme směrodatné chyby odhadů regresních parametrů b_0 a b_1. Přitom si uvědomíme, že $v_{00} = 4,2939$, $v_{11} = 0,0127$.

$s_{b_0} = s\sqrt{v_{00}} = \sqrt{18,3475} \cdot \sqrt{4,2939} = 8,8759$, $s_{b_1} = s\sqrt{v_{11}} = \sqrt{18,3475} \cdot \sqrt{0,0127} = 0,4827$.

Stanovíme meze 95% intervalů spolehlivosti pro regresní parametry β_0 a β_1. K tomu slouží vzorec $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$, j = 0, 1.

95% interval spolehlivosti pro β_0:

$d = b_0 - t_{0,975}(13)s_{b_0} = 5,0101 - 2,1604 \cdot 8,8759 = -14,1654$

$h = b_0 + t_{0,975}(13)s_{b_0} = 5,0101 + 2,1604 \cdot 8,8759 = 24,1856$

Znamená to, že -14,1654 < β_0 < 24,1456 s pravděpodobností aspoň 0,95.

95% interval spolehlivosti pro β_1:

$d = b_1 - t_{0,975}(13)s_{b_1} = 4,3024 - 2,1604 \cdot 0,4827 = 3,2596$

$h = b_1 + t_{0,975}(13)s_{b_1} = 4,3024 + 2,1604 \cdot 0,4827 = 5,3452$

Znamená to, že 3,2596 < β_1 < 5,3452 s pravděpodobností aspoň 0,95.

ad e) Provedení celkového F-testu: na hladině významnosti $\alpha = 0,05$ testujeme H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$.

Testová statistika $F = \dfrac{S_R / p}{S_E / (n-p-1)} = \dfrac{1461,0831/1}{238,5169/(15-1-1)} = 79,6341$,

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

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05. Výsledky testování významnosti modelu jako celku zapíšeme do tabulky ANOVA:

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	-	-

ad f) Provedení dílčích t-testů:

Na hladině významnosti $\alpha = 0,05$ testujeme H_0: $\beta_0 = 0$ proti H_1: $\beta_0 \neq 0$.

Testová statistika: $t_0 = \dfrac{b_0}{s_{b_0}} = \dfrac{5,0101}{8,8759} = 0,5645$,

kritický obor:

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

$= (-\infty, -2,1604\rangle \cup \langle 2,1604, \infty)$.

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_0 (tj. posunutí regresní přímky) nezamítáme na hladině významnosti 0,05.

Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_0. Vypočítali jsme, že -14,1654 < β_0 < 24,1456 s pravděpodobností aspoň 0,95. Protože tento interval obsahuje 0, hypotézu H_0: $\beta_0 = 0$ nezamítáme na hladině významnosti 0,05.

Na hladině významnosti $\alpha = 0,05$ testujeme H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$.

Testová statistika: $t_1 = \dfrac{b_1}{s_{b_1}} = \dfrac{4,3024}{0,4827} = 8,9132$,

kritický obor:

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

$$= (-\infty, -2{,}1604\rangle \cup \langle 2{,}1604, \infty)$$

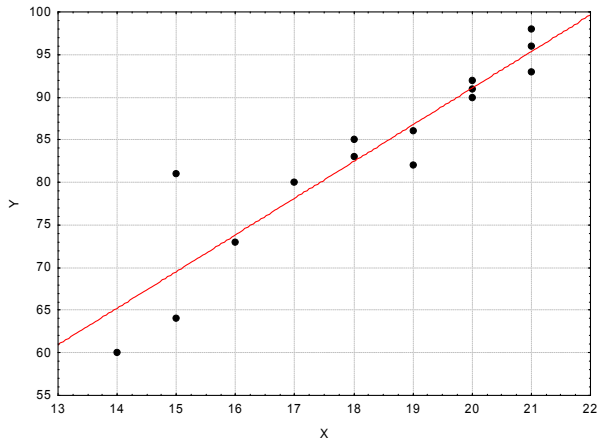
Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.

Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_1. Vypočítali jsme, že $3{,}2596 < \beta_1 < 5{,}3452$ s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, hypotézu H_0: $\beta_1 = 0$ zamítáme na hladině významnosti 0,05.

V případě modelu regresní přímky je dílčí t-test pro parametr β_1 ekvivalentní s celkovým F-testem.

ad g) Regresní odhad pro x = 18 dostaneme pouhým dosazením do rovnice regresní přímky: $\hat{y} = 5{,}0101 + 4{,}3024 \cdot 18 = 82{,}45$.

ad h)



Příklad 4.: Regresní parabola

U automobilu Škoda 120 byla změřena spotřeba benzínu (v l/100 km) v závislosti na rychlosti (v km/h).

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) Data znázorněte graficky dvourozměrným tečkovým diagramem a najděte vhodnou regresní funkci.

b) Sestavte regresní matici, vypočtěte odhady regresních parametrů, odhad rozptylu a index determinace.

c) Určete 95 % intervaly spolehlivosti pro regresní parametry.

d) Na hladině významnosti 0,05 proveďte celkový F-test.

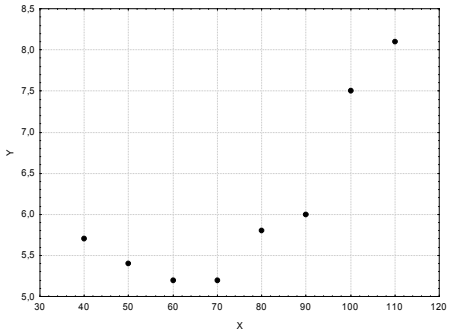
e) Na hladině významnosti 0,05 proveďte dílčí t-testy.

f) Určete regresní odhad spotřeby benzínu při rychlosti 80 km/h.

g) Znázorněte data s proloženou regresní funkcí.

Řešení:

ad a)



Z dvourozměrného tečkového diagramu je patrno, že vhodnou regresní funkcí bude parabola: $m(x;\beta_0,\beta_1,\beta_2) = \beta_0 + \beta_1 x + \beta_2 x^2$.

ad b) Regresní matice:

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

. Podle vzorce $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ získáme odhady regresních parametrů:

$$\mathbf{b} = \begin{pmatrix} 9{,}751786 \\ -0{,}150536 \\ 0{,}001244 \end{pmatrix}$$

. Regresní parabola má tedy rovnici:

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

Nyní vypočteme vektor regresních odhadů proměnné Y (vektor predikce):

$$\hat{\mathbf{y}} = \mathbf{Xb} = \begin{pmatrix} 5{,}720746 \\ 5{,}334986 \\ 5{,}198026 \\ 5{,}309866 \\ 5{,}670506 \\ 6{,}279946 \\ 7{,}138186 \\ 8{,}245226 \end{pmatrix}$$

. Vektor reziduí: $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -0{,}020746 \\ 0{,}065014 \\ 0{,}001974 \\ -0{,}109866 \\ 0{,}129494 \\ -0{,}279946 \\ 0{,}361814 \\ -0{,}145226 \end{pmatrix}$. Reziduální součet čtverců:

$S_E = \mathbf{e'e} = 0{,}263869$.

Odhad rozptylu: $s^2 = \frac{S_E}{n-p-1} = \frac{0{,}263869}{8-2-1} = 0{,}052774$.

Dále potřebujeme celkový součet čtverců $S_T = (\mathbf{y} - \mathbf{m}_2)'(\mathbf{y} - \mathbf{m}_2)$, kde $\mathbf{m}_2$ je sloupcový vektor typu nx1 složený z průměru m_2 závisle proměnné veličiny Y. V našem případě $m_2 = 6{,}1125$. Po dosazení do vzorce pro celkový součet čtverců tedy dostaneme $S_T = 8{,}32875$. (Celkový součet čtverců lze získat také tak, že výběrový rozptyl veličiny Y vynásobíme n-1: $S_T = 7.1{,}189821 = 8{,}32875$.) Regresní součet čtverců pak je: $S_R = S_T - S_E = 8{,}32875 - 0{,}263869 = 8{,}06488$.

Index determinace: $ID^2 = \frac{S_R}{S_T} = \frac{8{,}06488}{8{,}32875} = 0{,}9683$.

Znamená to, že variabilita hodnot závisle proměnné veličiny je z 96,83% vysvětlena regresní parabolou.

ad c) Podle vzorce $s_{b_j} = s\sqrt{v_{jj}}$, j = 0, 1, 2 vypočteme směrodatné chyby odhadů b_0, b_1 a b_2 regresních parametrů β_0, β_1 a β_2. Přitom si uvědomíme, že v_{00}, v_{11} a v_{22} jsou diagonální prvky matice $(\mathbf{X'\,X})^{-1}$. V našem případě $s_{b_0} = 0{,}945689$, $s_{b_1} = 0{,}026821$, $s_{b_2} = 0{,}000177$.

Stanovíme meze 95% intervalů spolehlivosti pro regresní parametry β_0 a β_1. K tomu slouží vzorec $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$, j = 0, 1, 2.

95% interval spolehlivosti pro β_0:

$d = b_0 - t_{0{,}975}(5)s_{b_0} = 9{,}751786 - 2{,}5706 \cdot 0{,}945689 = 7{,}3208$

$h = b_0 + t_{0{,}975}(5)s_{b_0} = 9{,}751786 + 2{,}5706 \cdot 0{,}945689 = 12{,}1828$

Znamená to, že $7{,}3208 < \beta_0 < 12{,}1828$ s pravděpodobností aspoň 0,95.

95% interval spolehlivosti pro β_1:

$d = b_1 - t_{0{,}975}(5)s_{b_1} = -0{,}150536 - 2{,}5706 \cdot 0{,}026821 = -0{,}2195$

$h = b_1 + t_{0{,}975}(5)s_{b_1} = -0{,}150536 + 2{,}5706 \cdot 0{,}026821 = -0{,}0816$

Znamená to, že $-0{,}2195 < \beta_1 < -0{,}0816$ s pravděpodobností aspoň 0,95.

95% interval spolehlivosti pro β_2:

$d = b_2 - t_{0{,}975}(5)s_{b_2} = 0{,}001244 - 2{,}5706 \cdot 0{,}000177 = 0{,}0008$

$h = b_2 + t_{0{,}975}(5)s_{b_2} = 0{,}001244 + 2{,}5706 \cdot 0{,}000177 = 0{,}0017$

Znamená to, že $0{,}0008 < \beta_2 < 0{,}0017$ s pravděpodobností aspoň 0,95.

ad d) Provedení celkového F-testu: na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $(\beta_1, \beta_2) = (0, 0)$ proti H_1: $(\beta_1, \beta_2) \neq (0, 0)$.

Testová statistika $F = \frac{S_R / p}{S_E / (n-p-1)} = \frac{8{,}06488 / 2}{0{,}263869 / (8-2-1)} = 76{,}41$,

kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty) = \langle F_{0{,}95}(2,5), \infty) = \langle 19{,}2964, \infty)$.

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 zamítáme na hladině významnosti 0,05. Výsledky testování významnosti modelu jako celku zapíšeme do tabulky ANOVA:

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) Provedení dílčích t-testů:

Na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_0 = 0$ proti H_1: $\beta_0 \neq 0$.

Testová statistika: $t_0 = \frac{b_0}{s_{b_0}} = \frac{9{,}751786}{0{,}945689} = 10{,}3118$,

kritický obor:

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

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_0 zamítáme na hladině významnosti 0,05.
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_0.
Vypočítali jsme, že $7{,}3208 < \beta_0 < 12{,}1828$ s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, hypotézu H_0: $\beta_0 = 0$ zamítáme na hladině významnosti 0,05.
Na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$.

Testová statistika: $t_1 = \frac{b_1}{s_{b_1}} = \frac{-0{,}150536}{0{,}026821} = -5{,}6126$,

kritický obor:

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

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 zamítáme na hladině významnosti 0,05.
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_1.
Vypočítali jsme, že $-0{,}2195 < \beta_1 < -0{,}0816$ s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, hypotézu H_0: $\beta_1 = 0$ zamítáme na hladině významnosti 0,05.
Na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_2 = 0$ proti H_1: $\beta_2 \neq 0$.

Testová statistika: $t_2 = \frac{b_2}{s_{b_2}} = \frac{0{,}001244}{0{,}000177} = 7{,}0282$,

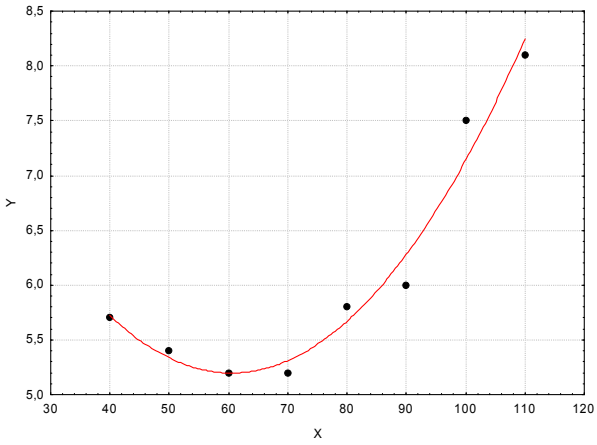
kritický obor:

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

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_2 zamítáme na hladině významnosti 0,05.
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_2.
Vypočítali jsme, že $0{,}0008 < \beta_2 < 0{,}0017$ s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, hypotézu H_0: $\beta_2 = 0$ zamítáme na hladině významnosti 0,05.

ad f) Regresní odhad pro x = 80 dostaneme pouhým dosazením do rovnice regresní paraboly:
$\hat{y} = 9{,}751786 - 0{,}150536 \cdot 80 + 0{,}001244 \cdot 80^2 = 5{,}67$.

ad g)


8,5
8,0
7,5
7,0
6,5
6,0
5,5
5,0
Y
30
40
50
60
70
80
90
100
110
120
X