# Jednoduchá lineární regrese II

**Opakování**

Studujeme regresní model

$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, kde

$\mathbf{y} = (y_1, \ldots, y_n)'$ - vektor pozorování závisle proměnné veličiny Y,

$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix}$ - regresní matice

(předpokládáme, že h(**X**) = p+1 < n)

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ - vektor náhodných odchylek, pro který platí $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X'X\beta} = \mathbf{X'y}$ - systém normálních rovnic

$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ – odhad vektoru $\boldsymbol{\beta}$ získaný metodou nejmenších čtverců

$\hat{\mathbf{y}} = \mathbf{Xb}$ – vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ - vektor reziduí

Vlastnosti odhadu **b**:

- odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$;
- odhad **b** je nestranný, tj. E(**b**) = **β**;
- odhad **b** má varianční matici var $\mathbf{b} = \sigma^2(\mathbf{X'X})^{-1}$;
- odhad $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X'X})^{-1})$;
- odhad **b** je nejlepší nestranný lineární odhad vektoru **β**.

Součty čtverců a index determinace:

$S_E = \mathbf{e'e}$ … reziduální součet čtverců (podíl $s^2 = \frac{S_E}{n-p-1}$ je odhad rozptylu $\sigma^2$)

$S_R = (\hat{\mathbf{y}} - \mathbf{m}_2)'(\hat{\mathbf{y}} - \mathbf{m}_2)$ … regresní součet čtverců, kde $\mathbf{m}_2$ je sloupcový vektor průměrů závisle proměnné veličiny Y

$S_T = (\mathbf{y} - \mathbf{m}_2)'(\mathbf{y} - \mathbf{m}_2)$ … celkový součet čtverců

Platí $S_T = S_R + S_E$

$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$ - index determinace ($0 \le ID^2 \le 1$), udává, jakou část variability závisle proměnné veličiny Y lze vysvětlit zvolenou regresní funkcí

## Intervaly spolehlivosti pro regresní parametry

$100(1-\alpha)\%$ interval spolehlivosti pro $\beta_j$ má meze:

$b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$,

kde $s_{b_j} = s\sqrt{v_{jj}}$ je směrodatná chyba odhadu $b_j$, $v_{jj}$ je j-tý diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$, j = 0, 1, …, p

## Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme $H_0$: $(\beta_1,\ldots,\beta_p)' = (0,\ldots,0)'$ proti $H_1$: $(\beta_1,\ldots,\beta_p)' \neq (0,\ldots,0)'$.

(Nulová hypotéza říká, že dostačující je model konstanty.)

Testová statistika: $F = \dfrac{S_R/p}{S_E/(n-p-1)}$ má rozložení F(p, n-p-1), pokud $H_0$ platí.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

$F \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

| zdroj variability | součet čtverců | stupně volnosti | podíl | statistika F |
|---|---|---|---|---|
| model | $S_R$ | p | $S_R/p$ | $\dfrac{S_R/p}{S_E/(n-p-1)}$ |
| reziduální | $S_E$ | n-p-1 | $S_E/(n-p-1)$ | - |
| celkový | $S_T$ | n-1 | - | - |

## Testování významnosti regresních parametrů (dílčí t-testy)

Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu $H_0$: $\beta_j = 0$ proti $H_1$: $\beta_j \neq 0$.

Testová statistika: $T_j = \dfrac{b_j}{s_{b_j}}$ má rozložení t(n-p-1), pokud $H_0$ platí.

Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty)$.

$T_j \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

## Nové poznatky

## Interval spolehlivosti pro teoretickou regresní funkci

Nechť $x_0$ je pevně zvolená hodnota nezávisle proměnné veličiny X. Vytvořme vektor $\mathbf{x}_0 = (1, f_1(x_0), \ldots, f_p(x_0))'$ a zabývejme se lineární kombinací $\mathbf{x}_0'\boldsymbol{\beta}$ složek vektoru regresních parametrů, tj. hodnotou $m(x_0;\beta_0,\beta_1,\ldots,\beta_p) = \sum_{j=0}^{p} \boldsymbol{\beta}_j f_j(x_0)$ teoretické regresní funkce v bodě $x_0$. Nestranným odhadem této lineární kombinace je $\mathbf{x}_0'\mathbf{b}$ s varianční maticí $\sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$. Protože $\mathbf{x}_0'\mathbf{b} \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \boldsymbol{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$, dostáváme, že $U = \dfrac{\mathbf{x}_0'\mathbf{b} - \mathbf{x}_0'\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim N(0,1)$. Jelikož $K = \dfrac{(n-p-1)s^2}{\sigma^2} \sim \chi^2(n-p-1)$,

plyne odtud, že náhodná veličina $T = \frac{\mathbf{x}_0'\mathbf{b} - \mathbf{x}_0'\boldsymbol{\beta}}{s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t(n-p-1)$. Vidíme tedy, že 100(1- α)% interval spolehlivosti pro $\mathbf{x}_0'\boldsymbol{\beta}$, tj. pro hodnotu regresní funkce $m(x_0;\beta_0,\beta_1,\ldots,\beta_p)$ má meze $\mathbf{x}_0'\mathbf{b} \pm t_{1-\alpha/2}(n-p-1)s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$.
Při spojitě se měnícím $\mathbf{x}_0$ vytvoří meze tohoto intervalu spolehlivosti tzv. pás spolehlivosti kolem regresní funkce.

## Predikční interval spolehlivosti

V případě, kdy chceme zkonstruovat 100(1- α)% interval spolehlivosti nikoli pro hodnotu regresní funkce, ale pro i-tou predikovanou hodnotu $\hat{y}_i$ (tzv. predikční interval), dostaneme meze

$$\mathbf{x}_0'\mathbf{b} \pm t_{1-\alpha/2}(n-p-1)s\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

Vidíme, že tento predikční interval je širší než předešlý interval spolehlivosti. Je to interval, který nás informuje o tom, v jakém rozsahu můžeme očekávat jedno další pozorování s pravděpodobností aspoň 1- α.
Při spojitě se měnícím $\mathbf{x}_0$ vytvoří meze tohoto predikčního intervalu spolehlivosti tzv. predikční pás spolehlivosti kolem regresní funkce.

## Regresní přímka a její vlastnosti

Uvažujeme regresní model

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

(Parametr $\beta_0$ interpretujeme jako teoretickou hodnotu Y při x = 0 a $\beta_1$ udává změnu Y, když X se změní o jednotku. Systém normálních rovnic získáme derivováním výrazu

$S_E(\beta_0,\beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ parciálně podle $\beta_0$ a $\beta_1$:

$$\frac{\partial S_E(\beta_0,\beta_1)}{\partial \beta_0} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial S_E(\beta_0,\beta_1)}{\partial \beta_1} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

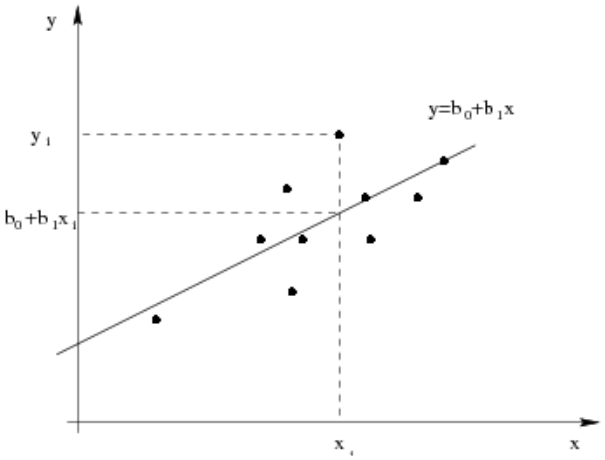
Řešením tohoto systému získáme odhady

$$b_0 = \frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}, \quad b_1 = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$$

Po jednoduchých úpravách dospějeme ke tvaru $b_1 = \frac{s_{XY}}{s_X^2}$, kde $s_{XY}$ je kovariance hodnot $(x_i, y_i)$, i = 1, ..., n a $s_X^2$ je rozptyl hodnot $x_1,\ldots,x_n$. Dále dostáváme

$b_0 = m_Y - b_1 m_X$, tedy regresní přímku můžeme vyjádřit ve tvaru

$$y = m_Y + \frac{s_{XY}}{s_X^{\,2}}(x - m_X) + \varepsilon .$$




Pro regresní přímku má reziduální součet čtverců tvar

$$S_E = \sum_{i=1}^{n} Y_i^{\,2} - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} x_i Y_i .$$

Odhad rozptylu: $s^2 = \frac{S_E}{n-2}$.

Index determinace:

$ID^2 = \frac{S_R}{S_T}$, kde

$$S_R = \sum_{i=1}^{n} (\hat{y}_i - m_Y)^2 = \sum_{i=1}^{n} \left[ m_Y + \frac{s_{XY}}{s_X^{\,2}}(x_i - m_X) - m_Y \right]^2 = \frac{s_{XY}^{\,2}}{s_X^{\,4}} \sum_{i=1}^{n} (x_i - m_Y)^2 = n \frac{s_{XY}^{\,2}}{s_X^{\,2}}$$

$S_T = \sum_{i=1}^{n} (y_i - m_Y)^2 = n s_Y^{\,2}$, tedy

$$ID^2 = \frac{n \frac{s_{XY}^{\,2}}{s_X^{\,2}}}{n s_Y^{\,2}} = \frac{s_{XY}^{\,2}}{s_X^{\,2} s_Y^{\,2}} = r_{XY}^{\,2} .$$

Vidíme tedy, že v případě regresní přímky je index determinace roven kvadrátu koeficientu korelace.

Test významnosti směrnice regresní přímky (tj. test $H_0$: $\beta_1 = 0$ proti $H_1$: $\beta_1 \neq 0$) je ekvivalentní hypotéze o nulovosti koeficientu korelace (tj. testu $H_0$: $\rho = 0$ proti $H_1$: $\rho \neq 0$). Jestliže koeficient korelace veličin X, Y je blízký 0, nemá smysl počítat parametry regresní přímky.

Interval spolehlivosti pro teoretickou regresní přímku při zadané hodnotě $x_0$ má meze:

$$d = b_0 + b_1 x_0 - s \cdot t_{1-\alpha/2}(n-2)\sqrt{\frac{1}{n} + \frac{(x_0 - m_X)^2}{\sum_{i=1}^{n} x_i^2 - n m_X^2}},$$

$$h = b_0 + b_1 x_0 + s \cdot t_{1-\alpha/2}(n-2)\sqrt{\frac{1}{n} + \frac{(x_0 - m_X)^2}{\sum_{i=1}^{n} x_i^2 - n m_X^2}}.$$

Predikční interval spolehlivosti pro budoucí pozorování y při zadané hodnotě $x_0$ má meze:

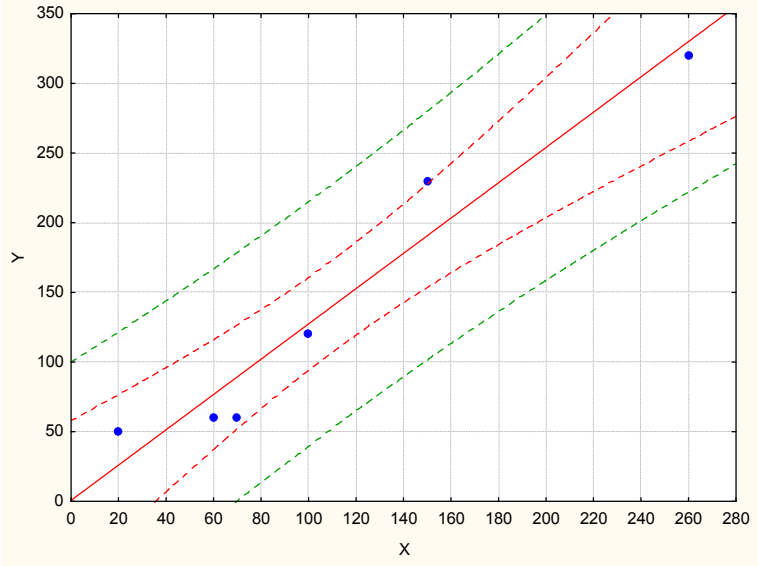
$$d = b_0 + b_1 x_0 - s \cdot t_{1-\alpha/2}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - m_X)^2}{\sum_{i=1}^{n} x_i^2 - n m_X^2}},$$

$$h = b_0 + b_1 x_0 + s \cdot t_{1-\alpha/2}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - m_X)^2}{\sum_{i=1}^{n} x_i^2 - n m_X^2}}.$$

Srovnání intervalu spolehlivosti a predikčního intervalu při zadané hodnotě $x_0$:

- oba intervaly jsou nejužší v místě $x_0 = m_X$,
- interval spolehlivosti pro dané $x_0$ je vždy užší než odpovídající predikční interval,
- predikční interval je určen pro individuální pozorování, zatímco interval spolehlivosti je určen pro hodnoty ležící na regresní přímce,
- s rostoucím rozsahem výběru se zmenšuje šířka obou intervalů.

Data s proloženou regresní přímkou, pásy spolehlivosti (červeně) a predikčními pásy (zeleně)




## Předpoklady použití regresní přímky:

- Závislost Y na X má lineární charakter.

- Pro celý rozsah uvažovaných hodnot nezávisle proměnné X je reziduální rozptyl $s^2$ konstantní (hovoříme o homoskedasticitě a znamená to, že variabilita hodnot závisle proměnné veličiny Y kolem regresní přímky je stejná pro všechny uvažované hodnoty nezávisle proměnné veličiny X).
- Hodnoty závisle proměnné veličiny Y mají normální rozložení pro dané hodnoty $x_i$ a jsou stochasticky nezávislé (to souvisí s uspořádáním experimentu).

Poznámka: Menší odchylky od normality a homoskedasticity je možno tolerovat.

## Sdružené regresní přímky

Uvažme nyní situaci, kdy obě veličiny Y a X jsou náhodné, přičemž samozřejmě předpokládáme, že X nezávisí na rušivé složce $\varepsilon$. Pak jde o případ oboustranné závislosti.

Závislost Y na X vystihuje 1. regresní přímka $Y = \beta_0 + \beta_1 X + \varepsilon$ a závislost X na Y vystihuje 2. regresní přímka $X = \alpha_0 + \alpha_1 Y + \delta$. Odhady $a_0, a_1$ regresních koeficientů $\alpha_0, \alpha_1$ v modelu $X_i = \alpha_0 + \alpha_1 Y_i + \delta_i$ získáme opět metodou nejmenších čtverců ve tvaru

$a_1 = \frac{s_{XY}}{s_Y^2}, a_0 = m_X - a_1 m_Y = m_X - \frac{s_{XY}}{s_Y^2} m_Y$. 2. regresní přímka má tedy rovnici:

$$X = m_X + \frac{s_{XY}}{s_Y^2}(Y - m_Y) + \delta.$$

1. a 2. regresní přímka se nazývají sdružené regresní přímky a odhady regresních koeficientů $b_1, a_1$ se nazývají odhady párově sdružených regresních koeficientů. Je zřejmé, že $b_1 a_1 = r_{XY}^2$. Rovnice sdružených regresních přímek můžeme tedy psát ve tvaru:

$$Y = m_Y + \frac{s_{XY}}{s_X^2}(x - m_X) + \varepsilon, \quad Y = m_X + \frac{1}{r_{XY}} \frac{s_X}{s_X}(x - m_Y) + \delta.$$

Sdružené regresní přímky se protínají v bodě o souřadnicích $[m_X, m_Y]$. V případě, že náhodné veličiny X, Y jsou nekorelované, jsou odhady $b_1, a_1$ nulové a sdružené regresní přímky mají tvar $Y = m_Y + \varepsilon$, $Y = m_X + \delta$. Pokud mezi náhodnými veličinami X, Y existuje úplná lineární závislost, pak sdružené regresní přímky splynou. K tomu dojde tehdy, když $r_{XY}^2 = 1$, tj. $a_1 = \frac{1}{b_1}$.

Označíme-li $\varphi$ úhel, který svírají sdružené regresní přímky, pak z předešlých úvah plyne:

$\cos\varphi = 0 \Leftrightarrow$ mezi X a Y neexistuje žádná lineární závislost;

$\cos\varphi = 1 \Leftrightarrow$ mezi X a Y existuje úplná přímá lineární závislost;

$\cos\varphi = -1 \Leftrightarrow$ <=> mezi X a Y existuje úplná nepřímá lineární závislost.

**Příklad:**
Z fiktivního základního souboru všech vzorků oceli odpovídajících „všem myslitelným tavbám“ bylo do laboratoře dodáno 60 vzorků a zjištěny a hodnoty proměnné X – mez plasticity a Y – mez pevnosti. Datový soubor má tvar:

| | | | | | |
|---|---|---|---|---|---|
| 154 | 178 | 83 | 98 | 73 | 76 |
| 133 | 164 | 106 | 111 | 77 | 85 |
| 58 | 75 | 92 | 104 | 47 | 61 |
| 145 | 161 | 85 | 103 | 68 | 85 |
| 94 | 107 | 112 | 118 | 137 | 142 |
| 113 | 141 | 98 | 102 | 44 | 68 |
| 80 | 97 | 103 | 108 | 92 | 116 |
| 121 | 127 | 99 | 119 | 141 | 157 |
| 119 | 138 | 104 | 128 | 155 | 189 |
| 112 | 125 | 107 | 118 | 136 | 155 |
| 85 | 97 | 98 | 140 | 82 | 81 |
| 41 | 72 | 97 | 116 | 136 | 163 |
| 96 | 113 | 105 | 101 | 72 | 79 |
| 45 | 89 | 71 | 93 | 66 | 81 |
| 99 | 109 | 39 | 69 | 42 | 61 |
| 51 | 95 | 122 | 147 | 113 | 123 |
| 101 | 114 | 33 | 52 | 42 | 85 |
| 130 | 169 | 78 | 117 | 133 | 147 |
| 87 | 101 | 114 | 137 | 153 | 179 |
| 88 | 130 | 125 | 149 | 85 | 91 |

a) Určete regresní přímku meze pevnosti na mez plasticity.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu spolu s 95% pásem spolehlivosti a predikčním pásem spolehlivosti.
c) Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
d) Vypočtěte index determinace a interpretujte ho.
e) Určete regresní přímku meze plasticity na mez pevnosti.
f) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
g) Obě regresní přímky zakreslete do téhož dvourozměrného tečkového diagramu.

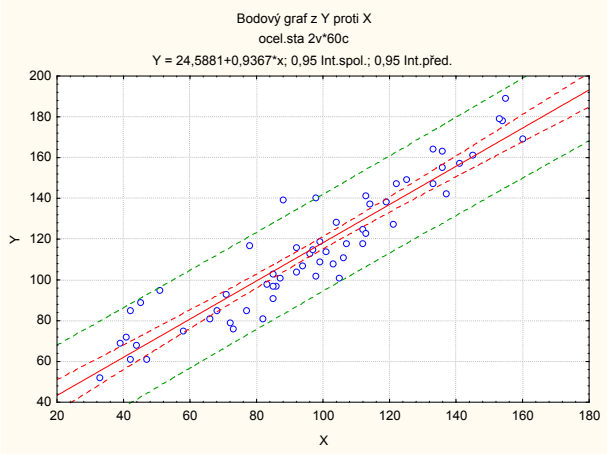
**Řešení:**
Nejprve vypočteme číselné charakteristiky obou proměnných:
$m_X = 95{,}9$, $m_Y = 114{,}4$,
$s_X^2 = 1052{,}40$, $s_Y^2 = 1057{,}21$, $s_X = 32{,}4$, $s_Y = 32{,}5$,
$s_{XY} = 985{,}76$, $r_{XY} = 0{,}936$.

ad a) Dosadíme do vzorců pro výpočet směrnice a úseku regresní přímky:

$$b_1 = \frac{s_{XY}}{s_X^2} = \frac{985{,}76}{1052{,}4} = 0{,}937\,,\ b_0 = m_Y - b_1 m_X = 114{,}4 - 0{,}937\ .\ 95{,}9 = 24{,}5$$

Regresní přímka meze pevnosti na mez plasticity má tedy rovnici
$y = 24{,}5 + 0{,}937x$. Znamená to, že při nulové mezi plasticity by mez pevnosti byla 24,5. Pokud mez plasticity vzroste o jednotku, mez pevnosti vzroste o 0,937.

ad b) Vytvoříme dvourozměrný tečkový diagram s proloženou regresní přímkou:




ad c) Regresní odhad meze pevnosti pro mez plasticity = 60:
$\hat{y} = 24,5 + 0,937.60 = 80,72.$

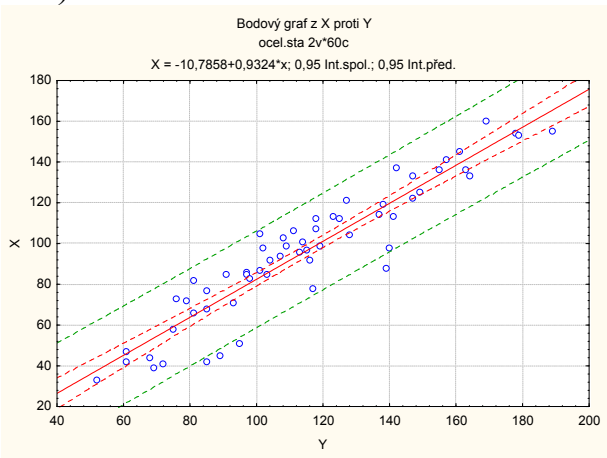
ad d) $ID^2 = r_{12}^2 = 0,936^2 = 0,876$. Znamená to, že 87,6% variability hodnot meze pevnosti je vysvětleno regresní přímkou.

ad e)
Dosadíme do vzorců pro výpočet směrnice a úseku 2. regresní přímky:

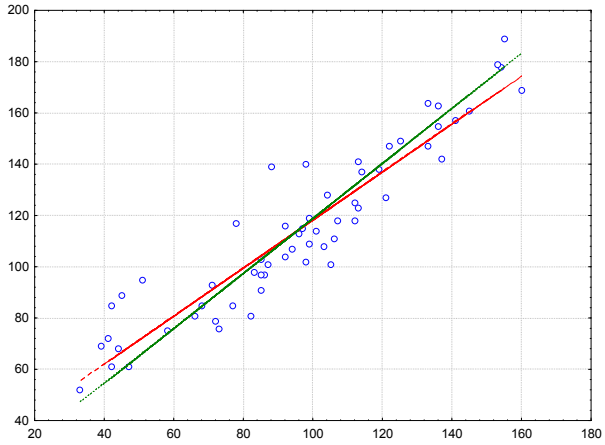
$$a_1 = \frac{s_{XY}}{s_Y^2} = \frac{985,76}{1057,21} = 0,932, a_0 = m_X - a_1 m_Y = 95,9 - 0,932 \cdot 114,4 = -10,7$$

Regresní přímka meze plasticity na mez pevnosti má tedy rovnici
x = -10,7 + 0,932y.

ad f)




ad g) 1. regresní přímka má rovnici y = 24,5 + 0,937x, 2. regresní přímka má rovnici x = -10,7 + 0,932y, tedy $y = \frac{10,7 + x}{0,932}$. Obě přímky zakreslíme do téhož dvourozměrného tečkového diagramu.




## Řešení v systému STATISTICA:

Odhad parametrů 1. regresní přímky:
Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

| N=60 | Výsledky regrese se závislou proměnnou : Y (ocel.sta)<br>R= ,93454811 R2= ,87338017 Upravené R2= ,87119707<br>F(1,58)=400,06 p<0,0000 Směrod. chyba odhadu : 11,768 | | | | | |
|---|---|---|---|---|---|---|
| | Beta | Sm.chyba beta | B | Sm.chyba B | t(58) | Úroveň p |
| Abs.člen | | | 24,58814 | 4,740272 | 5,18707 | 0,000003 |
| X | 0,934548 | 0,046724 | 0,93668 | 0,046830 | 20,00160 | 0,000000 |

Zakreslení regresních pásů do dvourozměrného tečkového diagramu s proloženou regresní přímkou:
Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zaškrtneme Regresní pásy Spolehl. – OK. Ve vytvořeném grafu pak 2x klikneme na pozadí grafu a v nabídce Regresní pásy vybereme Přidat nový pár pásů – zaškrtneme Predikční.
Analogicky získáme výsledky pro 2. regresní přímku:

| N=60 | Výsledky regrese se závislou proměnnou : X (ocel.sta)<br>R= ,93454811 R2= ,87338017 Upravené R2= ,87119707<br>F(1,58)=400,06 p<0,0000 Směrod. chyba odhadu : 11,741 | | | | | |
|---|---|---|---|---|---|---|
| | Beta | Sm.chyba beta | B | Sm.chyba B | t(58) | Úroveň p |
| Abs.člen | | | -10,7858 | 5,544250 | -1,94540 | 0,056579 |
| Y | 0,934548 | 0,046724 | 0,9324 | 0,046617 | 20,00160 | 0,000000 |

Nakreslení sdružených regresních přímek do jednoho diagramu:
K datovému souboru ocel.sta přidáme dvě nové proměnné y1 a y2. Do proměnné y1 uložíme predikované hodnoty meze pevnosti na mezi plasticity (do Dlouhého jména proměnné y1 napíšeme =24,58814 – 0,93668*x a do Dlouhého jména proměnné y2 napíšeme =(x+10,7858)/0,9324

Grafy – Bodové grafy – zaškrtneme Vícenásobný – Proměnné X: X, Y: Y, y1, y2 – OK. Ve vytvořeném grafu pak vypneme zobrazování značek pro y1, y2 a naopak zapneme Spojnici.

## Test linearity regrese

Nechť hodnoty závisle proměnné veličiny Y jsou roztříděny do $r \geq 3$ skupin podle variant $x_{[1]}, ..., x_{[r]}$ nezávisle proměnné veličiny X, přičemž aspoň jedna skupina má více než jedno pozorování. Budeme předpokládat, že každá skupina hodnot má normální rozložení a že všechny skupiny mají týž rozptyl. Všech pozorování je n.
Charakter závislosti Y na X popíšeme regresní přímkou a budeme se zabývat testováním hypotézy, zda je regresní přímka vhodným modelem pro tato data.

Testová statistika: $F = \frac{(S_A - S_R)/(r-2)}{(S_T - S_A)/(n-r)}$

kde $S_R$ je regresní součet čtverců, $S_A$ je skupinový součet čtverců a $S_T$ je celkový součet čtverců (viz kapitola Jednofaktorová analýza rozptylu). Platí-li $H_0$, pak $F \sim F(r-2, n-r)$.
Kritický obor: $W = \langle F_{1-\alpha}(r-2,n-r), \infty)$
$F \in W$ => na hladině významnosti $\alpha$ zamítáme hypotézu, že přímka je vhodným regresním modelem závislosti Y na X.
Těsnost závislosti Y na X vyjádřenou skupinovými průměry měří poměr determinace $P^2 = SA/ST$. Nabývá hodnot z intervalu <0,1>. Čím je poměr determinace bližší jedné, tím je závislost silnější, čím je bližší nule, tím je závislost slabší.

**Příklad:** Máme k dispozici údaje o cenách 23 náhodně vybraných domů (veličina Y - v tisících $) a počtu jejich pokojů (veličina X) v jednom americkém městě.

| počet pokojů | cena |
|---|---|
| 5 | 155,168,180 |
| 6 | 166,172,179,190,200 |
| 7 | 210,215,218,225,230,245 |
| 8 | 213,225,240,247,249 |
| 9 | 267,275,290,298 |

Závislost ceny domu na počtu pokojů popište regresní přímkou. Na hladině významnosti 0,05 testujte hypotézu, že přímka je vhodným regresním modelem pro tato data. Těsnost závislosti vyjádřete poměrem determinace. Znázorněte data s proloženou regresní přímkou.

**Řešení:** Empirická regresní přímka má tvar y = 17,2885 + 28,5851 x,
$S_R = 30907{,}9041$, $S_T = 35870{,}6087$, $S_A = 32474{,}1087$,
$F = \frac{(32474{,}1087 - 30907{,}9041)/(5-2)}{(35870{,}6087 - 32474{,}1087)/(23-5)} = 2{,}768$, $F_{0,95}(3,18) = 3{,}161$,
kritický obor W = <3,161, ∞). Jelikož F ∉ W, nezamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem.
Poměr determinace: $P^2 = 32474{,}1087/35870{,}6087 = 0{,}9053$, tedy závislost ceny domu na počtu pokojů je v daném datovém souboru značně silná.

## Řešení v systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 23 případy:

| | 1 X | 2 Y |
|---|---|---|
| 1 | 5 | 155 |
| 2 | 5 | 168 |
| 3 | 5 | 180 |
| 4 | 6 | 166 |
| 5 | 6 | 172 |
| 6 | 6 | 179 |
| 7 | 6 | 190 |
| 8 | 6 | 200 |
| 9 | 7 | 210 |
| 10 | 7 | 215 |
| 11 | 7 | 218 |
| 12 | 7 | 225 |
| 13 | 7 | 230 |
| 14 | 7 | 245 |
| 15 | 8 | 213 |
| 16 | 8 | 225 |
| 17 | 8 | 240 |
| 18 | 8 | 247 |
| 19 | 8 | 249 |
| 20 | 9 | 267 |
| 21 | 9 | 275 |
| 22 | 9 | 290 |
| 23 | 9 | 298 |

Odhadneme parametry regresní přímky:

| N=23 | Výsledky regrese se závislou proměnnou : Y (ceny_bytu.sta)<br>R= ,92825096 R2= ,86164984 Upravené R2= ,85506173<br>F(1,21)=130,79 p<,00000 Směrod. chyba odhadu : 15,373 | | | | | |
|---|---|---|---|---|---|---|
| | Beta | Sm.chyba beta | B | Sm.chyba B | t(21) | Úroveň p |
| Abs.člen | | | 17,28851 | 18,00156 | 0,96039 | 0,347788 |
| X | 0,928251 | 0,081167 | 28,58506 | 2,49950 | 11,43629 | 0,000000 |

Sestavíme tabulku ANOVA:
Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

| Efekt | Analýza rozptylu (ceny_bytu.sta) | | | | |
|---|---|---|---|---|---|
| | Součet čtverců | sv | Průměr čtverců | F | Úroveň p |
| Regres. | 30907,90 | 1 | 30907,90 | 130,7888 | 0,000000 |
| Rezid. | 4962,70 | 21 | 236,32 | | |
| Celk. | 35870,61 | | | | |

Vidíme, že $S_R = 30907{,}9$, $S_T = 35870{,}61$

Provedeme jednofaktorovou analýzu rozptylu, abychom získali skupinový součet čtverců:
Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – Y, Grupovací - X – OK – OK – Analýza rozptylu.

| Proměnná | Analýza rozptylu (ceny_bytu.sta)<br>Označ. efekty jsou význ. na hlad. p < ,05000 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | SČ efekt | SV efekt | PČ efekt | SČ chyba | SV chyba | PČ chyba | F | p |
| Y | 32474,11 | 4 | 8118,527 | 3396,500 | 18 | 188,6944 | 43,02473 | 0,000000 |

Zde najdeme $S_A = 32474{,}11$.

Vypočteme testovou statistiku $F = \frac{(32474{,}1087 - 30907{,}9041)/(5-2)}{(35870{,}6087 - 32474{,}1087)/(23-5)} = 2{,}768$ a najdeme kritický obor $W = \langle 3{,}161, \infty)$. Jelikož $F \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem.

## Linearizující transformace

Odhad parametrů regresních funkcí, které nejsou lineární z hlediska parametrů, se neprovádí metodou nejmenších čtverců přímo, protože její použití vede k soustavě nelineárních rovnic. V některých speciálních případech však nelineární regresní funkci můžeme vhodnou transformací převést na lineární.
Např. máme exponenciální regresní funkci $y = \beta_0 \beta_1^{\,x}$. Provedeme logaritmickou transformaci $\ln y = \ln \beta_0 + x \ln \beta_1$, čímž získáme regresní funkci lineární v parametrech. Parametry $\ln \beta_0$ a $\ln \beta_1$ odhadneme metodou nejmenších čtverců a odlogaritmováním získáme odhady původních regresních koeficientů $\beta_0$, $\beta_1$.

### Přehled linearizujících transformací

| Funkce | Linearizující transformace |
|---|---|
| $y = \beta_0 \beta_1^{\,x}$ | $\ln y = \ln \beta_0 + x \ln \beta_1$ |
| $y = \beta_0 x^{\beta_1}$ | $\ln y = \ln \beta_0 + \beta_1 \ln x$ |

$y = \frac{\beta_0}{x^{\beta_1}}$ $\qquad$ $\ln y = \ln \beta_0 - \beta_1 \ln x$

$y = \frac{1}{\beta_0 + \beta_1 x}$ $\qquad$ $\frac{1}{y} = \beta_0 + \beta_1 x$

$y = \frac{x}{\beta_0 + \beta_1 x}$ $\qquad$ $\frac{x}{y} = \beta_0 + \beta_1 x$

**Příklad:** Hotelová společnost vlastnící 12 hotelů analyzuje vztah mezi celkovými měsíčními tržbami (veličina Y) a tržbami vyprodukovanými stravovacími úseky (veličina X).

| č. h. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x | 2,0 | 1,2 | 14,8 | 8,3 | 8,4 | 3,0 | 4,8 | 15,6 | 16,1 | 11,5 | 14,2 | 14,0 |
| y | 12,0 | 8,0 | 76,4 | 17,0 | 21,3 | 10,0 | 12,5 | 97,3 | 88,0 | 25,0 | 38,6 | 47,3 |

Popište tuto závislost exponenciální regresní funkcí $y = \beta_0 \beta_1^{\,x}$. Najděte odhady parametrů $\beta_0$, $\beta_1$ a vypočtěte predikovanou hodnotu celkových měsíčních tržeb pro x = 10.

**Řešení:** Provedeme logaritmickou transformaci $\ln y = \ln \beta_0 + x \ln \beta_1$. Metodou nejmenších čtverců získáme odhady $\ln b_0 = 1{,}8559$, $\ln b_1 = 0{,}1504$. Odlogaritmováním dostaneme $b_0 = 6{,}3973$, $b_1 = 1{,}1623$. Predikovaná hodnota y pro x = 10 je $6{,}3973.1{,}162310 = 28{,}7859$.

**Řešení v systému STATISTICA:**
Vytvoříme datový soubor se dvěma proměnnými a 12 případy:

| | 1<br>Y | 2<br>X |
|---|---|---|
| 1 | 12 | 2 |
| 2 | 8 | 1,2 |
| 3 | 76,4 | 14,8 |
| 4 | 17 | 8,3 |
| 5 | 21,3 | 8,4 |
| 6 | 10 | 3 |
| 7 | 12,5 | 4,8 |
| 8 | 97,3 | 15,6 |
| 9 | 88 | 16,1 |
| 10 | 25 | 11,5 |
| 11 | 38,6 | 14,2 |
| 12 | 47,3 | 14 |

Přidáme novou proměnnou ln y. Do jejího Dlouhého jména napíšeme =log(y). Pak provedeme regresní analýzu se závisle proměnnou ln y a nezávisle proměnnou X:

| N=12 | Výsledky regrese se závislou proměnnou : ln y (hotely.sta)<br>R= ,95851605 R2= ,91875303 Upravené R2= ,91062833<br>F(1,10)=113,08 p<,00000 Směrod. chyba odhadu : ,26364 | | | | | |
|---|---|---|---|---|---|---|
| | Beta | Sm.chyba beta | B | Sm.chyba B | t(10) | Úroveň p |
| Abs.člen | | | 1,855881 | 0,154338 | 12,02480 | 0,000000 |
| X | 0,958516 | 0,090137 | 0,150428 | 0,014146 | 10,63398 | 0,000001 |

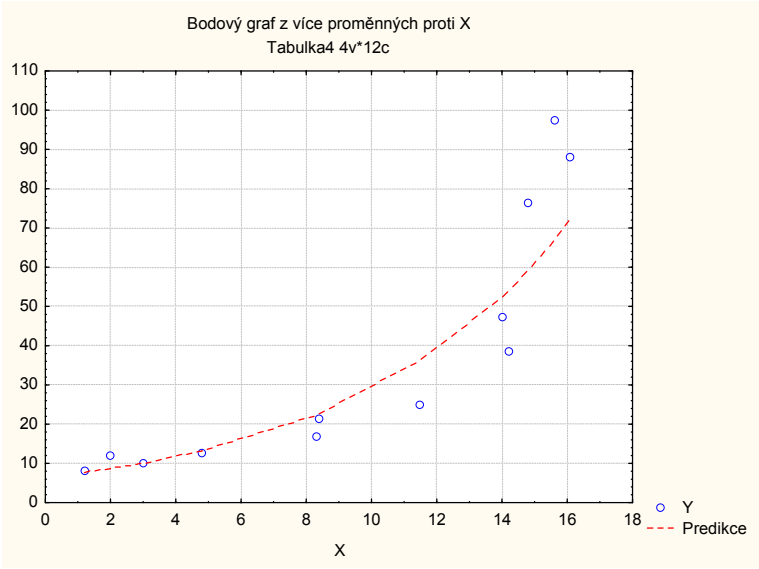
K výsledné tabulce přidáme novou proměnnou b, do jejíhož Dlouhého jména napíšeme =exp(B).

| N=12 | Výsledky regrese se závislou proměnnou : ln y (hotely.sta)<br>R= ,95851605 R2= ,91875303 Upravené R2= ,91062833<br>F(1,10)=113,08 p<,00000 Směrod. chyba odhadu : ,26364 | | | | | | |
|---|---|---|---|---|---|---|---|
| | Beta | Sm.chyba beta | B | Sm.chyba B | t(10) | Úroveň p | b =exp(B) |
| Abs.člen | | | 1,855881 | 0,154338 | 12,02480 | 0,000000 | 6,397333 |
| X | 0,958516 | 0,090137 | 0,150428 | 0,014146 | 10,63398 | 0,000001 | 1,162332 |

Vytvoříme ještě dvourozměrný tečkový diagram s proloženou exponenciálou. Na záložce Rezidua/předpoklady/předpovědi vybereme reziduální analýza – Uložit – Uložit rezidua & předpovědi – vybereme X, Y – OK.
Ve vzniklé tabulce odstraníme proměnné č. 5 až 10 a proměnnou rezidua přejmenujeme na Predikce. Do Dlouhého jména této proměnné napíšeme =exp(v3).
Tento datový soubor se čtyřmi proměnnými uspořádáme podle velikosti hodnot proměnné X: Data - Setřídit – Proměnná X – OK.

| | hotely.sta | | | |
|---|---|---|---|---|
| | 1 Y | 2 X | 3 Předpovědi | 4 Predikce |
| 1 | 8 | 1,2 | 2,04 | 7,66 |
| 1 | 12 | 2 | 2,16 | 8,64 |
| 3 | 10 | 3 | 2,31 | 10,05 |
| 4 | 12,5 | 4,8 | 2,58 | 13,17 |
| 5 | 17 | 8,3 | 3,10 | 22,30 |
| 6 | 21,3 | 8,4 | 3,12 | 22,63 |
| 7 | 25 | 11,5 | 3,59 | 36,08 |
| 8 | 47,3 | 14 | 3,96 | 52,56 |
| 9 | 38,6 | 14,2 | 3,99 | 54,16 |
| 10 | 76,4 | 14,8 | 4,08 | 59,28 |
| 11 | 97,3 | 15,6 | 4,20 | 66,86 |
| 12 | 88 | 16,1 | 4,28 | 72,08 |

Vytvoření grafu:
Grafy – Bodové grafy – zaškrtneme Vícenásobný – Proměnné X: X, Y: Y, Predikce – OK. Ve vytvořeném grafu pak vypneme zobrazování značek pro Predikce a naopak zapneme Spojnici.


Bodový graf z více proměnných proti X
Tabulka4 4v*12c
110
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
18
X
Y
Predikce