# MV011 Statistika I

# 10. Lineární regresní model

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

# Motivace

Často chceme prozkoumat vztah mezi dvěma veličinami, kde jedna z nich, tzv. „nezávisle proměnná“ $X$, má řídit druhou, tzv. „závisle proměnnou“ $Y$. Předpokládá se, že obě veličiny jsou spojité. Prvním krokem ve zkoumání by mělo být zakreslení dat do grafu. V řadě případů tento krok napoví mnohé o tom, co nás zajímá: Existuje vztah mezi oběma proměnnými (veličinami)? Pokud ano, pak rostou či klesají obě v jednom směru, nebo jedna klesá, když druhá roste? Je přímka vhodným modelem pro vyjádření vztahu mezi těmito dvěma veličinami? Chceme-li se dostat dále za tuto intuitivní úroveň analýzy, je lineární regrese často užitečným nástrojem. Tato metoda zahrnuje proložení přímky daty a analýzu statistických vlastností takovéto přímky.

# Regresní model

Pozorujeme dvojice $(x_i, Y_i),\ i = 1, \ldots, n$.
Předpokládáme

$$\boxed{Y_i = m(x_i) + \varepsilon_i}$$

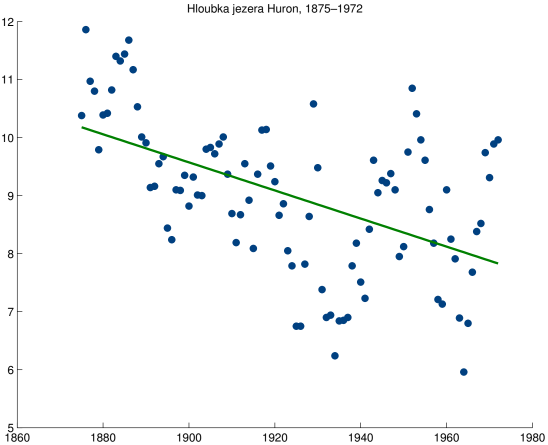
$x_i$ ... body pevného plánu,
$Y_i$ ... naměřené hodnoty
$\varepsilon_i$ ... chyby měření, $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$, $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$.

Hledáme **odhad** regresní funkce $m$.

# Motivační příklad

Regresní model

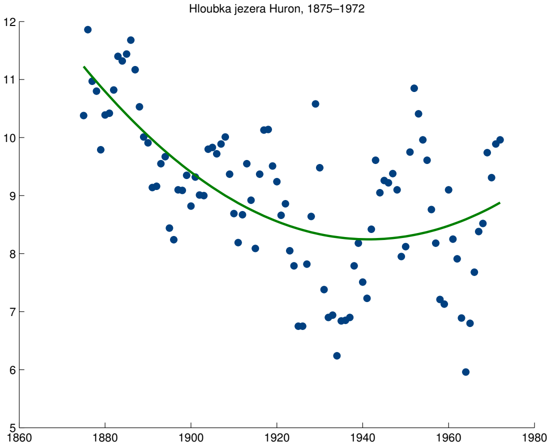
$$m(x) = \beta_0 + \beta_1 x$$



# Motivační příklad

Regresní model

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$



# Regresní model

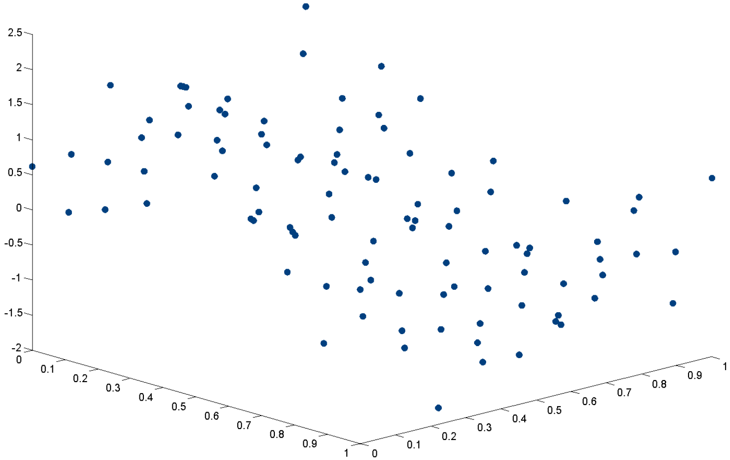
Pozorujeme trojice $(x_{i1}, x_{i2}, Y_i),\ i = 1, \ldots, n.$
Předpokládáme

$$\boxed{Y_i = m(x_{i1}, x_{i2}) + \varepsilon_i}$$

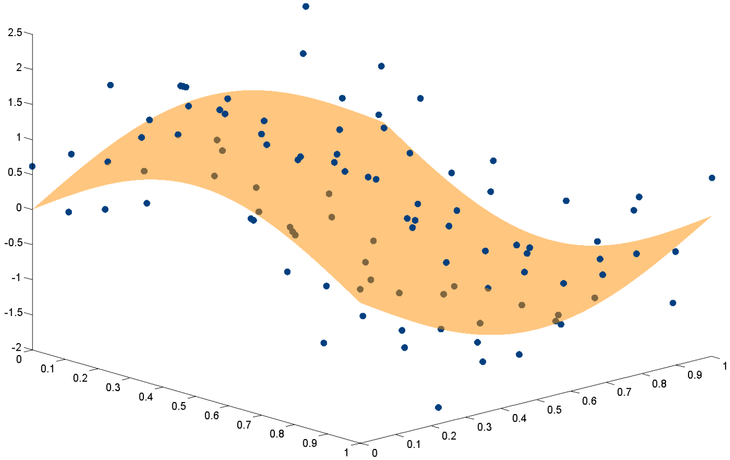
$(x_{i1}, x_{i2})$ ... body pevného plánu,
$Y_i$ ... naměřené hodnoty
$\varepsilon_i$ ... chyby měření, $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$, $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$.

Hledáme **odhad** regresní funkce $m$.

# Motivační příklad



# Motivační příklad



# Lineární regresní model

Předpokládejme, že mezi nějakými nenáhodnými veličinami $y, x_1, \ldots, x_k$ platí lineární vztah

$$y = \beta_1 x_1 + \cdots + \beta_k x_k,$$

ve kterém $\beta_1, \ldots, \beta_k$ jsou neznámé parametry. Informace o neznámých parametrech budeme získávat pomocí experimentu, a to tak, že opakovaně budeme měřit hodnoty veličiny $y$ při vybraných hodnotách proměnných $x_1, \ldots, x_k$. Při měřeních však vznikají chyby, což lze modelovat takto

$$Y = \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

kde $\varepsilon$ je náhodná chyba měření.

Opakované hodnoty sledovaných veličin budeme pro $i = 1, \ldots, n$ značit $Y_i, x_{i1}, \ldots, x_{ik}$, obdobně také náhodné chyby $\varepsilon_i$.

# Lineární regresní model

Celkově jsme dostali model

$$\begin{array}{rcl} Y_1 & = & \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ & \vdots & \\ Y_n & = & \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{array} \quad \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{\mathbf{X}\text{(matice plánu)}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

O náhodných chybách $\varepsilon_1, \ldots, \varepsilon_n$ budeme předpokládat, že jsou

- **nesystematické**, což lze matematicky vyjádřit požadavkem, že $\boxed{E\varepsilon_i = 0}$, $i = 1, \ldots, n$, tj. $\boxed{E\boldsymbol{\varepsilon} = \mathbf{0}}$ a tedy $\boxed{E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}}$
- **homogenní v rozptylu**, tj. že $\boxed{D\varepsilon_i = \sigma^2 > 0}$ pro $i = 1, \ldots, n$;
- jednotlivé náhodné chyby jsou **nekorelované**, tj. že $\boxed{C(\varepsilon_i, \varepsilon_j) = 0}$ pro $i \neq j, i, j = 1, \ldots, n$, tj. $\boxed{D\mathbf{Y} = D\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}_n}$, takže i měření jsou nekorelovaná.

# Terminologie

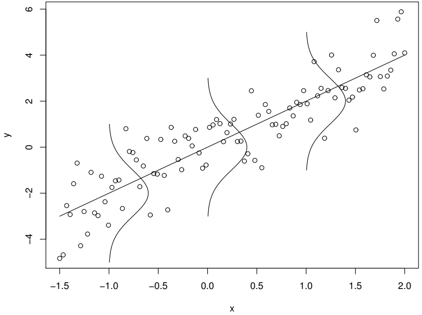
Používá se následující **terminologie** a značení

- parametry $\beta_1, \ldots, \beta_k$ se nazývají **regresní koeficienty** (**regression coefficients**);
- matice $\mathbf{X}$ obsahuje nenáhodné prvky $x_{ij}$ a nazývá se **regresní maticí** nebo **maticí plánu** (**design matrix**);
- popsaný model souhrnně zapíšeme jako $\boxed{\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)}$.

Takto zavedený model budeme nazývat **lineární regresní model**. Dále budeme předpokládat, že $\boxed{n > k}$ a o hodnosti matice $\mathbf{X}$ budeme předpokládat, že je rovna $k$, tj. $\boxed{h(\mathbf{X}) = k}$. Bude-li tento předpoklad splněn, budeme říkat, že jde **lineární regresní model plné hodnosti**. V tom případě jsou sloupce matice $\mathbf{X}$ nezávislé. V opačném případě, by bylo možné daný sloupec matice $\mathbf{X}$ napsat jako lineární kombinaci ostatních sloupců, což je možné interpretovat tak, že proměnná odpovídající danému sloupci je nadbytečná, protože ji lze vyjádřit jako lineární funkci ostatních proměnných.

# Příklad

## Regresní přímka v klasickém lineárním regresním modelu




**Jednoduchá lineární regrese:**
předpokládáme $Y_i$ $(i = 1, \dots, n)$ mají normální rozdělení

$$\boxed{Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)},$$

kde $x_i$ jsou dané konstanty, které nejsou všechny stejné.

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

V tomto případě

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

# Odhady neznámých parametrů

## Definice 1

Řekneme, že odhad $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je **lineárním odhadem** vektoru $\boldsymbol{\beta}$, jestliže existuje matice reálných čísel $\mathbf{B}_{k \times n}$ taková, že $\widehat{\boldsymbol{\beta}} = \mathbf{BY}$.
Dále řekneme, že odhad $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je **nestranným odhadem** vektoru $\boldsymbol{\beta}$, jestliže pro každé $\boldsymbol{\beta} \in \mathbb{R}^k$ platí $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.
Jestliže $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je takový lineární nestranný odhad vektoru parametrů $\boldsymbol{\beta}$, že pro každý jiný lineární nestranný odhad $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\mathbf{Y})$ je rozdíl variančních matic $D\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) - D\widehat{\boldsymbol{\beta}}(\mathbf{Y})$ **pozitivně semidefinitní matice**, potom budeme říkat, že $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je **nejlepší nestranný lineární odhad** (**Best Linear Unbiased Estimator**) parametrů $\boldsymbol{\beta}$, zkráceně $BLUE$ odhad.

# Metoda nejmenších čtverců

## Definice 2

Řekneme, že odhad $\widehat{\boldsymbol{\beta}}_{OLS}$ je odhadem parametru $\boldsymbol{\beta}$ **metodou nejmenších čtverců** (**Ordinary Least Square**), jestliže

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n} \left( Y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2$$

## Věta 3

*Odhad parametru $\boldsymbol{\beta}$ v modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ je tvaru*

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

# Důkaz

**Důkaz** Nejprve označme symbolem $\mathbf{x}_i'$ $i$-tý řádek matice plánu $\mathbf{X}$ a symbolem $\mathbf{X}_j$ $j$-tý sloupec této matice, tj.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \left( \mathbf{X}_1 \ldots \mathbf{X}_k \right)$$

Nutnou podmínkou pro extrém je, aby parciální derivace byly nulové, tj. pro $s = 1, \ldots, k$

$$0 = \frac{\partial}{\partial \beta_s} S(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_s} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_s} \sum_{i=1}^{n} \left( Y_i - \sum_{j=1}^{k} x_{ij} \beta_j \right)^2 .$$

Proto počítejme

$$
\begin{aligned}
\boxed{\frac{\partial}{\partial \beta_s} S(\boldsymbol{\beta})} &= \frac{\partial}{\partial \beta_s} \sum_{i=1}^{n} \left[ Y_i^2 - 2Y_i \sum_{j=1}^{k} x_{ij}\beta_j + \left( \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 \right] \\
&= -2 \sum_{i=1}^{n} Y_i x_{is} + 2 \sum_{i=1}^{n} \left( \sum_{j=1}^{k} x_{ij}\beta_j \right) x_{is} \\
&= -2 \sum_{i=1}^{n} Y_i x_{is} + 2 \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} x_{is} \beta_j = \boxed{0}
\end{aligned}
$$

tj.

$$
\boxed{\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} x_{is} \beta_j = \sum_{i=1}^{n} Y_i x_{is}}.
$$

Nyní se budeme snažit vyjádřit předchozí rovnost maticově. Upravujme postupně levou a pravou stranu:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}x_{is}\beta_s = \sum_{i=1}^{n} x_{is} \underbrace{\sum_{j=1}^{k} x_{ij}\beta_j}_{\mathbf{x}_i'\boldsymbol{\beta}} = \sum_{i=1}^{n} x_{is}\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{X}_s' \underbrace{\begin{pmatrix} \mathbf{x}_1'\boldsymbol{\beta} \\ \vdots \\ \mathbf{x}_n'\boldsymbol{\beta} \end{pmatrix}}_{\mathbf{X}\boldsymbol{\beta}} = \mathbf{X}_s'\mathbf{X}\boldsymbol{\beta}$$

$$\sum_{i=1}^{n} Y_i x_{is} = \mathbf{X}_s'\mathbf{Y}$$

a celkově, zapíšeme-li $k$ rovnic pod sebe a uvažujeme-li obě strany rovnosti, dostaneme

$$\underbrace{\begin{pmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_k' \end{pmatrix} \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \boldsymbol{\beta}}_{\mathbf{X}'\mathbf{X}\boldsymbol{\beta}} = \underbrace{\begin{pmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_k' \end{pmatrix} \mathbf{Y}}_{\mathbf{X}'\mathbf{Y}} \qquad \ldots \qquad \text{tzv. normální rovnice}$$

Vzhledem k předpokladu $h(\mathbf{X}) = h(\mathbf{X}'\mathbf{X}) = k$,

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}.$$

Nyní zbývá dokázat, že tento extrém je také minimem, tj. že matice druhých parciálních derivací je pozitivně semidefinitní matice. Proto počítejme $(sh)$-tý prvek matice druhých parciálních derivací

$$\begin{aligned}\frac{\partial^2}{\partial\beta_s\partial\beta_h}S(\boldsymbol{\beta}) &= \frac{\partial}{\partial\beta_h}\left[-2\sum_{i=1}^{n}Y_i x_{is} + 2\sum_{i=1}^{n}\sum_{j=1}^{k}x_{ij}x_{is}\beta_j\right] \\ &= 2\sum_{i=1}^{n}x_{is}\underbrace{\frac{\partial}{\partial\beta_h}\left(\sum_{j=1}^{k}x_{ij}\beta_j\right)}_{x_{ih}} = 2\sum_{i=1}^{n}x_{is}x_{ih} = 2\mathbf{X}_s'\mathbf{X}_h\end{aligned}$$

Takže matice druhých parciálních derivací je

$$\left(\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_s\partial\beta_h}\right)_{s,h=1}^{k} = \left(\sum_{i=1}^{n}x_{is}x_{ih}\right)_{s,h=1}^{k} = \mathbf{X}'\mathbf{X} > 0,$$

tj. jde o pozitivně definitní matici a tím je věta dokázaná.

# Věty

**Věta 4 (Gaussova-Markovova věta)**

*Odhad* $\widehat{\boldsymbol{\beta}}_{OLS}$ *v modelu* $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *je* $BLUE$*-odhad (tj. je nejlepší nestranný lineární odhad) a jeho variační matice je rovna*

$$D\widehat{\boldsymbol{\beta}}_{OLS} = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

**Věta 5**

*Pro libovolný vektor* $\mathbf{c} \in \mathbb{R}^k$ *je* $\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS}$ $BLUE$*-odhad parametrické funkce* $\mathbf{c}'\boldsymbol{\beta}$ *a má rozptyl* $\sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$.

# Věty

## Věta 6

*Platí*

$$S_e = S(\widehat{\boldsymbol{\beta}}_{OLS}) = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'_{OLS}\mathbf{X}'\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y},$$

*kde* $\mathbf{H}$ *je tzv. „hat“ matice*

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'.$$

## Věta 7

*Odhad* $s^2 = \frac{S_e}{n-k}$ *je nestranným odhadem rozptylu* $\sigma^2$.

# Příklad

## Příklad 1

*V LRM* $(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta})$, $\mathbf{X} = \begin{pmatrix} 1 & -1 & -3 \\ 1 & -1 & -2 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \end{pmatrix}$, $\mathbf{Y} = \begin{pmatrix} 5 \\ 7 \\ 8 \\ 12 \\ 13 \\ 15 \end{pmatrix}$ *spočítejte MNČ-odhady vektoru parametrů* $\hat{\boldsymbol{\beta}}$, *aproximace* $\hat{\mathbf{Y}}$, *reziduální součty čtverců* $S_e$ *a* $s^2$.

# Řešení

**Řešení** Nejprve vypočteme matice

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & 12 \\ 0 & 12 & 28 \end{pmatrix}, \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{pmatrix} 0,5 & 0 & -0,0714 \\ 0 & 0,0357 & 0 \\ -0,0714 & 0 & 0,0153 \end{pmatrix}.$$

Odtud pak

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 10 \\ 1/3 \\ 3/2 \end{pmatrix} \text{ a } \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 5,17 \\ 6,67 \\ 8,17 \\ 11,83 \\ 13,33 \\ 14,83 \end{pmatrix}.$$

Nakonec ještě

$$S_e = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = 1/3,\ s^2 = \frac{S_e}{n-k} = \frac{1/3}{3} = 1/9.$$

# Testování hypotéz v lineárním regresním modelu

Díky předchozím větám dokážeme v lineárním regresním modelu plné hodnosti vypočítat nejen $OLS$-odhady neznámých parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, ale také máme k dispozici odhad neznámého rozptylu $\sigma^2$ a známe vlastnosti těchto odhadů. V dalším se zaměříme na stanovení jejich rozdělení v případě, že náhodný vektor $\mathbf{Y}$ má **vícerozměrné normální rozdělení**. Pak teprve budeme moci přejít k testování hypotéz o neznámých parametrech $\beta_1, \ldots, \beta_k$.
Jestliže náhodný vektor $\mathbf{Y}$ se řídí lineárním regresním modelem plné hodnosti, což zapisujeme $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, a navíc má **vícerozměrné normální rozdělení**, budeme psát

$$\boxed{\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)}.$$

# Věta

## Věta 8

*Mějme lineární regresní model plné hodnosti, přičemž* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$. *Pak platí*

1. *OLS-odhad vektoru neznámých parametrů má normální rozdělení*

$$\widehat{\boldsymbol{\beta}}_{OLS} \quad \sim \quad N_k\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

2. *náhodná veličina*

$$K = \frac{n-k}{\sigma^2} s^2 \quad \sim \quad \chi^2(n-k)$$

3. *náhodná veličina* $\boxed{K = \frac{n-k}{\sigma^2} s^2}$ *a OLS-odhad* $\boxed{\widehat{\boldsymbol{\beta}}_{OLS}}$ *jsou nezávislé.*

# Test významnosti koeficientu $\beta_j$

## Věta 9

*V modelu* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *plné hodnosti pro každé* $\mathbf{c} \in \mathbb{R}^k$, $\mathbf{c} \neq \mathbf{0}$ *platí*

$$T = \frac{\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}'\boldsymbol{\beta}}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \quad \sim \quad t(n-k).$$

## Důsledek 10

*V modelu* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *plné hodnosti má* $100(1-\alpha)\%$ *interval spolehlivosti pro parametrickou funkci* $\mathbf{c}'\boldsymbol{\beta}$ *(kde* $\mathbf{c} \neq \mathbf{0}$*) tvar*

$$\left(\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} - s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}\, t_{1-\alpha/2}(n-k), \mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} + s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}\, t_{1-\alpha/2}(n-k)\right).$$

# Praktický test

Prakticky lze provést test hypotézy $H_0 : \mathbf{c}'\boldsymbol{\beta} = \gamma_0$ ($\gamma_0$ je dané reálné číslo) proti alternativě $H_1 : \mathbf{c}'\boldsymbol{\beta} \neq \gamma_0$ na hladině významnosti $\alpha$ tak, že hypotézu $H_0$ **zamítáme**, pokud platí

$$\frac{\left|\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} - \gamma_0\right|}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \geq t_{1-\alpha/2}(n-k)$$

# Poznámka

V praktických situacích se nejčastěji volí vektor $\mathbf{c}$ jako jednotkový s jedničkou na $j$-tém místě $\mathbf{c} = (0, \ldots, 1, 0, \ldots, 0)'$ a v tom případě $\mathbf{c}'\boldsymbol{\beta} = \beta_j$, takže

- (a) $100(1-\alpha)\%$ interval spolehlivosti má tvar (při značení $(\mathbf{X}'\mathbf{X})^{-1} = (v_{ij})_{i,j=1}^{k}$)

$$\left(\widehat{\beta}_{OLS,j} - s\sqrt{v_{jj}}\, t_{1-\alpha/2}(n-k)\ ,\ \widehat{\beta}_{OLS,j} + s\sqrt{v_{jj}}\, t_{1-\alpha/2}(n-k)\right).$$

- (b) Test hypotézy $H_0: \beta_j = \gamma_0$ ($\gamma_0$ je dané reálné číslo) proti alternativě $H_1: \beta_j \neq \gamma_0$ na hladině významnosti $\alpha$ se provede tak, že hypotézu $H_0$ **zamítáme**, pokud platí

$$\frac{\left|\widehat{\beta}_{OLS,j} - \gamma_0\right|}{s\sqrt{v_{jj}}} \geq t_{1-\alpha/2}(n-k).$$

# Test významnosti modelu

Zavedeme následující bloková značení:

$$\boldsymbol{\beta} = (\underbrace{\beta_1, \ldots, \beta_m}_{=\boldsymbol{\beta}_1'}, \underbrace{\beta_{m+1}, \ldots, \beta_k}_{=\boldsymbol{\beta}_2'})',$$

obdobně

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\widehat{\boldsymbol{\beta}}'_{OLS,1}, \widehat{\boldsymbol{\beta}}'_{OLS,2})'$$

a nakonec také pro matici

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix},$$

kde matice $\mathbf{V}_{11}$ je typu $m \times m$.

**Věta 11**

*V modelu $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ plné hodnosti platí, že statistika*

$$F = \frac{1}{s^2(k-m)} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_2\right)' \mathbf{V}_{22}^{-1} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_2\right) \quad \sim \quad F(k-m, n-k).$$

# Poznámka

Díky předcházející větě můžeme testovat nulovou hypotézu

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2,0},$$

(kde $\boldsymbol{\beta}_{2,0}$ je daný vektor reálných čísel, nejčastěji nulový vektor)

proti alternativě

$$H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_{2,0}$$

na hladině významnosti $\alpha$ tak, že hypotézu $H_0$ **zamítáme**, pokud platí

$$F_0 = \frac{1}{s^2(k-m)} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_{2,0}\right)' \mathbf{V}_{22}^{-1} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_{2,0}\right) \geq F_{1-\alpha}(k-m, n-k).$$

# Prakticky

Testujeme nulovou hypotézu

$$H_0 : (\beta_1, \dots, \beta_k)=(0,\dots,0)$$

proti alternativě

$$H_1 : \exists i > 0;\ \beta_i \neq 0$$

na hladině významnosti $\alpha$ tak, že hypotézu $H_0$ **zamítáme**, pokud platí

$$F_0 = \frac{s^2_{\widehat{Y}}}{s^2(k-1)} = \frac{ID}{1-ID}\frac{n-k}{k-1} \geq F_{1-\alpha}(k-1, n-k),$$

kde $s^2_{\widehat{Y}} = \frac{1}{n}\sum\limits_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$.

# Index determinace

Kvalitu modelu popisuje tzv. **Index Determinace**.

**Definice 12**

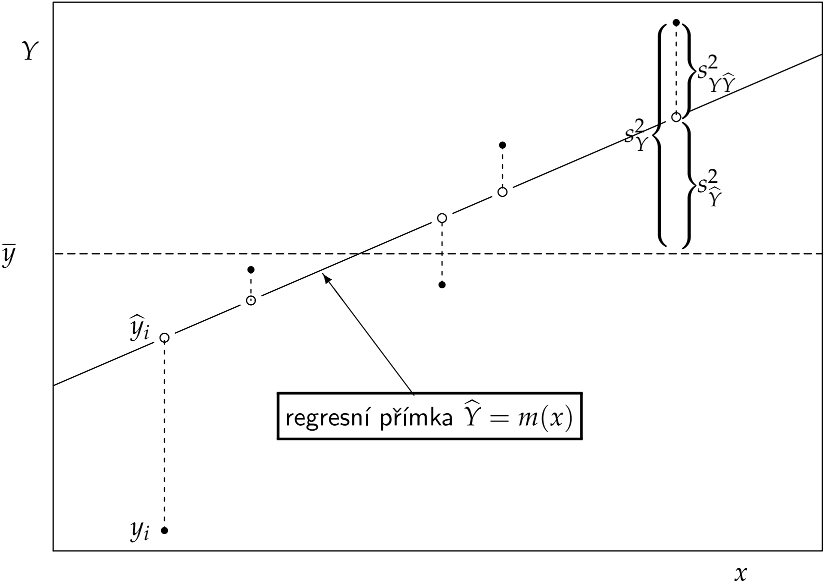
Nechť máme realizace $y_1, \ldots, y_n$ a jejich predikované hodnoty $\hat{y}_1, \ldots, \hat{y}_n$. **Index determinace** má tvar

$$ID = \frac{s^2_{\widehat{Y}}}{s^2_Y} = 1 - \frac{s^2_{Y\widehat{Y}}}{s^2_Y},$$

kde

$$s^2_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \quad s^2_{Y\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad s^2_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

# Graficky



$Y$
$s^2_{Y\widehat{Y}}$
$s^2_Y$
$s^2_{\widehat{Y}}$
$\overline{y}$
$\widehat{y}_i$
regresní přímka $\widehat{Y} = m(x)$
$y_i$
$x$

# Příklad

## Příklad 2

*Pro data*

| $x$ | $-2$ | $-1$ | $0$ | $1$ | $2$ |
|---|---|---|---|---|---|
| $Y$ | $-2$ | $1$ | $-2$ | $1$ | $-1$ |

*spočítejte MNČ-odhady vektoru parametrů $\hat{\boldsymbol{\beta}}$, aproximace $\hat{\mathbf{Y}}$, reziduální součty čtverců $s^2$ a index determinace $ID$ v následujících modelech. Odhadnuté regresní funkce znázorněte také graficky.*

1. $y = \beta_0 + \beta_1 x$
2. $y = \beta_1 x$
3. $y = \beta_0 + \beta_1 x + \beta_2 x^2$
4. $y = \beta_1 x + \beta_2 x^2$
5. $y = \beta_0 + \beta_1 x + \beta_2 e^x$

*Testujte významnost koeficientů $\beta_i$, testujte významnost modelu pomocí statistiky $F$. Porovnejte vhodnost regresních modelů pomocí $F$, $s^2$ a $ID$.*

# Řešení

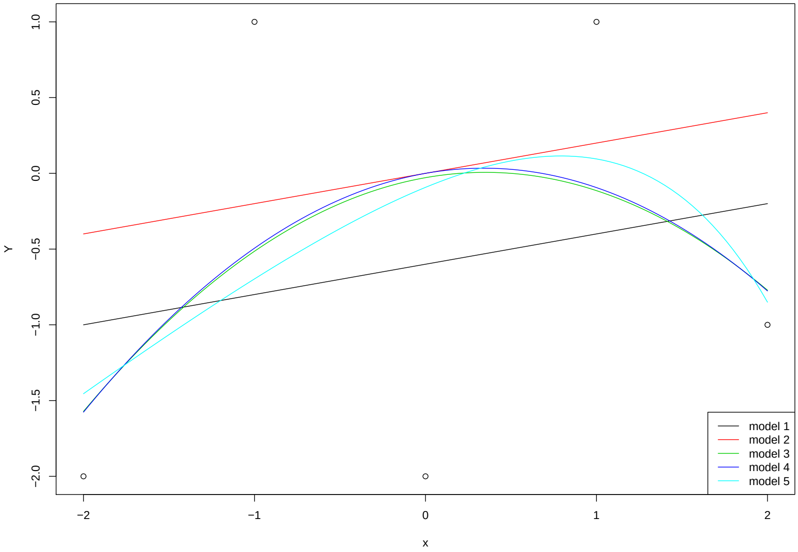
**Řešení** Pro jednotlivé modely počítejme postupně

1. $y = \beta_0 + \beta_1 x$
   $\hat{\boldsymbol{\beta}} = (-0,6; 0,2)'$, $\hat{\mathbf{Y}} = (-1; -0,8; -0,6; -0,4; -0,2)'$, $s^2 = 2,93$, $ID = 0,04348$, $F = 0,136$, $p$-hodnoty pro jednotlivé koeficienty: $(0,49; 0,73)$
2. $y = \beta_1 x$
   $\hat{\beta}_1 = 0,2$, $\hat{\mathbf{Y}} = (-0,4; -0,2; 0; 0,2; 0,4)'$, $s^2 = 2,65$, $ID = 0,0363$, $F = 0,15$, $p$-hodnoty pro jednotlivé koeficienty: $0,717$
3. $y = \beta_0 + \beta_1 x + \beta_2 x^2$
   $\hat{\boldsymbol{\beta}} = (-0,0286; 0,2; -0,2857)'$,
   $\hat{\mathbf{Y}} = (-1,5714; -0,5143; -0,0286; -0,1143; -0,7714)'$, $s^2 = 3,8286$,
   $ID = 0,1677$, $F = 0,2015$, $p$-hodnoty pro jednotlivé koeficienty:
   $(0,985; 0,777; 0,6396)$

# Řešení

4. $y = \beta_1 x + \beta_2 x^2$
   $\hat{\boldsymbol{\beta}} = (0,2; -0.2941)'$, $\hat{\mathbf{Y}} = (-1,576; -0,4941; 0; -0,0941; -0,776)'$, $s^2 = 2,55$, $ID = 0,3037$, $F = 0,654$, $p$-hodnoty pro jednotlivé koeficienty: $(0,718; 0,362)$
5. $y = \beta_0 + \beta_1 x + \beta_2 e^x$
   $\hat{\boldsymbol{\beta}} = (0,291; 0,847; -0,384)'$,
   $\hat{\mathbf{Y}} = (-1,4547; -0,6969; -0,0926; 0,0949; -0,851)'$, $s^2 = 3,8283$, $ID = 0,1677$, $F = 0,2015$, $p$-hodnoty pro jednotlivé koeficienty: $(0,8894; 0,59; 0,639)$.

# Řešení



1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Y
−2
−1
0
1
2
x
model 1
model 2
model 3
model 4
model 5

<u>MODEL I</u>: **Regresní přímka** $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$; $n > 2$.

Matice plánu $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix}$, $\quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum\limits_{i=1}^{n} Y_i \\ \sum\limits_{i=1}^{n} x_i Y_i \end{pmatrix}$

Model bude plné hodnosti, pokud všechny hodnoty $x_1, \dots, x_n$ nebudou stejné.
**Normální rovnice** jsou tvaru:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} Y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i \end{aligned}$$

# Speciální modely lineární regrese

<u>MODEL II</u>: **Regrese procházející počátkem** $Y_i = \beta x_i + \varepsilon_i$, $i = 1, \dots, n$; $n > 1$.

**Matice plánu** $\mathbf{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $\quad \mathbf{X}'\mathbf{X} = \left(\sum_{i=1}^{n} x_i^2\right)$, $\quad \mathbf{X}'\mathbf{Y} = \left(\sum_{i=1}^{n} x_i Y_i\right)$

a model bude plné hodnosti, pokud alespoň jedna z hodnot $x_1, \dots, x_n$ bude různá od nuly.

**Normální rovnice**:

$$\beta \sum_{i=1}^{n} x_i^2 \quad = \quad \sum_{i=1}^{n} x_i Y_i$$

# Speciální modely lineární regrese

<u>Model III</u>: **Kvadratická regrese** $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, $i = 1, \ldots, n$; $n > 3$.

**Matice plánu** $\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$, $\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum\limits_{i=1}^n x_i & \sum\limits_{i=1}^n x_i^2 \\ \sum\limits_{i=1}^n x_i & \sum\limits_{i=1}^n x_i^2 & \sum\limits_{i=1}^n x_i^3 \\ \sum\limits_{i=1}^n x_i^2 & \sum\limits_{i=1}^n x_i^3 & \sum\limits_{i=1}^n x_i^4 \end{pmatrix}$,

$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum\limits_{i=1}^n Y_i \\ \sum\limits_{i=1}^n x_i Y_i \\ \sum\limits_{i=1}^n x_i^2 Y_i \end{pmatrix}$, **Norm. rov.**:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^n x_i + \beta_2 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n Y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 + \beta_2 \sum_{i=1}^n x_i^3 &= \sum_{i=1}^n x_i Y_i \\ \beta_0 \sum_{i=1}^n x_i^2 + \beta_1 \sum_{i=1}^n x_i^3 + \beta_2 \sum_{i=1}^n x_i^4 &= \sum_{i=1}^n x_i^2 Y_i \end{aligned}$$

# Speciální modely lineární regrese

<u>Model IV</u>: **Polynomická regrese** $Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_m x_i^m + \varepsilon_i$, $i = 1, \ldots, n$; $n > m+1$.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum\limits_{i=1}^n x_i & \cdots & \sum\limits_{i=1}^n x_i^m \\ \sum\limits_{i=1}^n x_i & \sum\limits_{i=1}^n x_i^2 & \cdots & \sum\limits_{i=1}^n x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum\limits_{i=1}^n x_i^m & \sum\limits_{i=1}^n x_i^{m+1} & \cdots & \sum\limits_{i=1}^n x_i^{2m} \end{pmatrix},$$

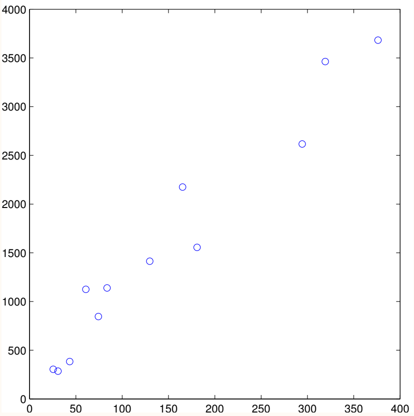
$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum\limits_{i=1}^n Y_i \\ \sum\limits_{i=1}^n x_i Y_i \\ \vdots \\ \sum\limits_{i=1}^n x_i^m Y_i \end{pmatrix}$$

# Příklad

## Příklad 3

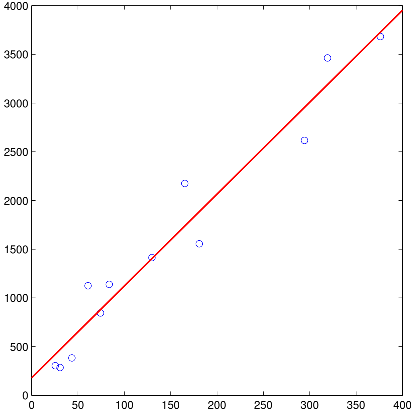
*Analyzujte data o počtu pracovních hodin za měsíc $Y$ spojených s provozováním anesteziologické služby v závislosti na velikosti spádové populace nemocnice $X$ (v tisících). Údaje byly získány ve 12 nemocnicích ve Spojených státech.*

| $i$ | $Y$ | $X$ |
|---|---|---|
| 1 | 304,37 | 25,5 |
| 2 | 2616,32 | 294,3 |
| 3 | 1139,12 | 83,7 |
| 4 | 285,43 | 30,7 |
| 5 | 1413,77 | 129,8 |
| 6 | 1555,68 | 180,8 |
| 7 | 383,78 | 43,4 |
| 8 | 2174,27 | 165,2 |
| 9 | 845,30 | 74,3 |
| 10 | 1125,28 | 60,8 |
| 11 | 3462,60 | 319,2 |
| 12 | 3682,33 | 376,2 |



# Řešení

Graf naznačuje lineární vztah mezi pracovní dobou a velikostí populace, a tak budeme pokračovat kvantifikací tohoto vztahu pomocí přímky $y = \beta_0 + \beta_1 x$.



# Řešení

| Parametr | Koeficient | $SE$ koef. | $t$-statistika | $p$-hodnota |
|---|---|---|---|---|
| $\beta_0$ | 180,658 | 128,381 | 1,407 | 0,1896823 |
| $\beta_1$ | 9,429 | 0,681 | 13,847 | 7,520972e-08 |

Z tabulky tedy dostáváme:

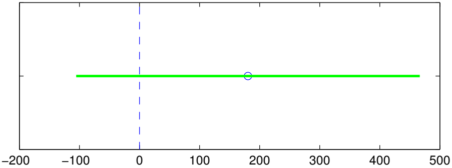
$$\text{pracovní doba } = 180{,}658 + 9{,}429 \cdot \text{velikost populace.}$$

**Co je na tom divného?**

# Řešení

Oboustranný interval spolehlivosti pro $\beta_0$

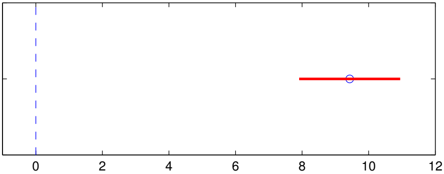
$$180{,}6575 \pm 2{,}228 \cdot 128{,}3812 = 180{,}6575 \pm 286{,}051$$




$$(-105{,}394;\ 466{,}709)$$

Oboustranný interval spolehlivosti pro $\beta_1$

$$9{,}429 \pm 2{,}228 \cdot 0{,}681 = 9{,}429 \pm 1{,}517$$




$$(7{,}912;\ 10{,}946)$$

# Řešení

Uvažujeme **regresi procházející počátkem** (plná čára) a výsledek srovnáme s obecnou regresní přímkou (čárkovaná čára).

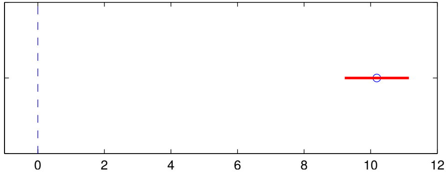
$$\widehat{\beta}_1^* = 10,185 \quad \widehat{s}_{\beta_1^*} = 0,4371,$$

$t^*$ =3,30157,

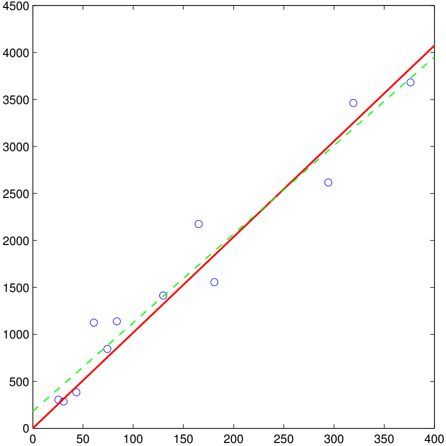
$p^*$-hodnota=1,0318e-10

Oboustranný interval spolehlivosti pro $\beta_1^*$

$$10{,}185 \pm 2{,}2 \cdot 0{,}4371 = 10{,}185 \pm 0{,}962$$




$$(9{,}223;\ 11{,}147)$$




pracovní doba $=$ $10{,}185 \cdot$ velikost populace.

# Příklad

## Příklad 4

*U 118 podniků řepařské oblasti v České Republice byl sledován hektarový výnos cukrovky ve vztahu ke spotřebě průmyslových hnojiv.*
*Data jsou uložena v souboru „`cukrovka.Rdata`“ ve 4 sloupcích:*

1. *dolní hranice spotřeby* $K_2O$ *(kg/ha)*
2. *horní hranice spotřeby* $K_2O$ *(kg/ha)*
3. *četnosti*
4. *průměrné výnosy cukrovky (q/ha)*

**a)** *odhadněte parametry regresní funkce tvaru*

$$y = \beta_0 + \beta_1 x$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$y = \beta_0 + \beta_1 x^{0{,}5}$$

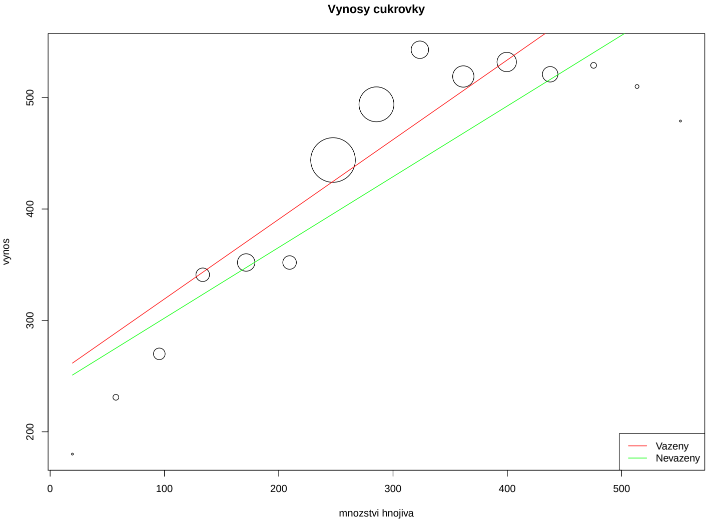
*Poznámka: Za hodnoty nezávisle proměnné volte střed intervalu.*

**b)** *Porovnejte vhodnost tří použitých regresních modelů.*

# Řešení

Regresní model I

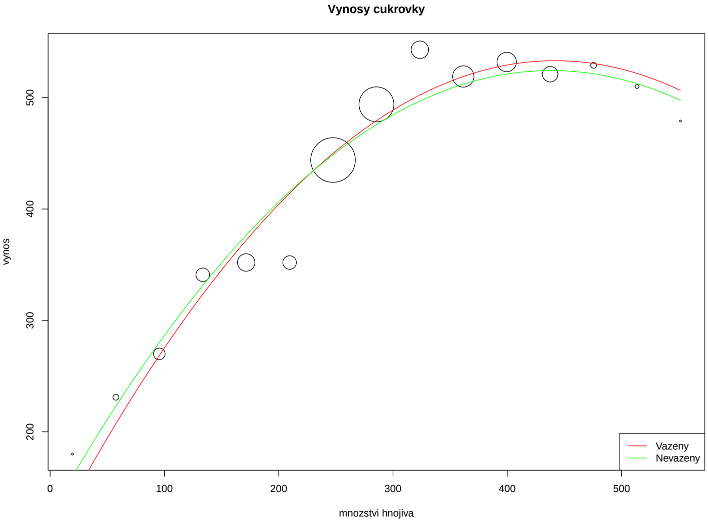
$$m(x) = \beta_0 + \beta_1 x$$



# Řešení

Regresní model II

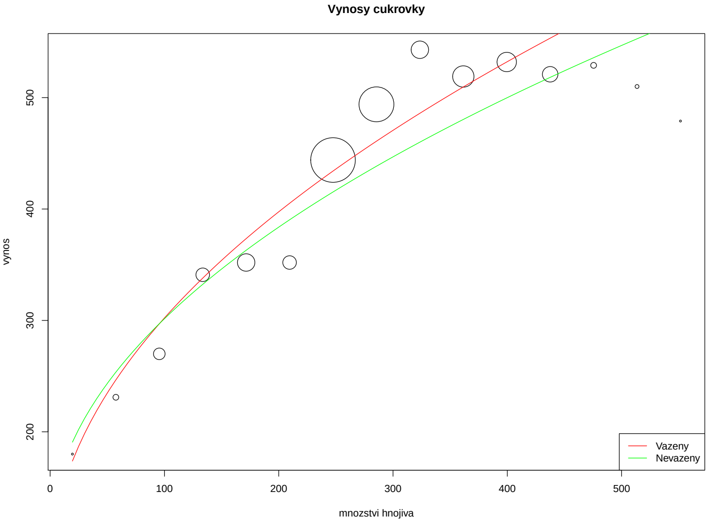
$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$



# Řešení

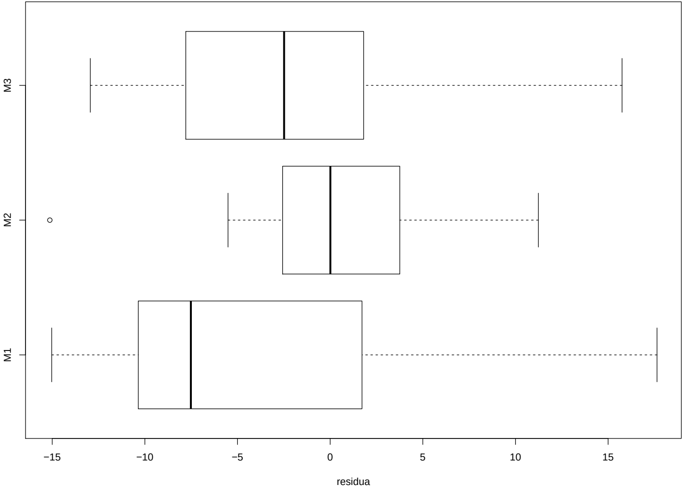
Regresní model III

$$m(x) = \beta_0 + \beta_1 x^{0,5}$$



# Řešení

Srovnání modelů pomocí residuí



# Úlohy k procvičení

## Příklad 5.1

*V LRM* $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$, $\boldsymbol{X} = \begin{pmatrix} 1 & -3 & 9 \\ 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix}$, $\boldsymbol{Y} = \begin{pmatrix} 7 \\ 4 \\ 2 \\ 2 \\ 5 \\ 8 \end{pmatrix}$ *spočítejte MNČ-odhady vektoru parametrů* $\hat{\boldsymbol{\beta}}$, *aproximace* $\hat{\boldsymbol{Y}}$, *reziduální součty čtverců* $S_e$ *a* $s^2$.

[$\hat{\boldsymbol{\beta}} = (1,5; 0,1786; 0,6786)'$, $\hat{\boldsymbol{Y}} = (7,0714; 3,8571; 2; 2,3571; 4,5714; 8,1429)'$, $S_e = 0,3571$, $s^2 = 0,119$.]

# Úlohy k procvičení

## Příklad 5.2

*Pro data*

| $x$ | $-2$ | $-1$ | $0$ | $1$ | $2$ |
|---|---|---|---|---|---|
| $Y$ | $0$ | $2$ | $3$ | $3$ | $1$ |

*spočítejte MNČ-odhady vektoru parametrů* $\hat{\boldsymbol{\beta}}$, *aproximace* $\widehat{\boldsymbol{Y}}$, *reziduální součty čtverců* $S_e$ *a* $s^2$ *ve dvou modelech. Který model je vhodnější? (Proč?) Oba modely vykreslete.*

**a** *model s regresní funkcí* $Y = \beta_0 + \beta_1 x + \beta_2 x^2$

**(b)** *model s maticí plánu* $\boldsymbol{X} = \begin{pmatrix} 1 & 4 \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 4 \end{pmatrix}$

*[(a)* $\hat{\boldsymbol{\beta}} = (3,09; 0,3; -0,64)'$, $\widehat{\boldsymbol{Y}} = (-0,086; 2,143; 3,086; 2,743; 1,114)'$, $S_e = 0,114$, $s^2 = 0,057$.
*(b)* $\hat{\boldsymbol{\beta}} = (3,17; -0,67)'$, $\widehat{\boldsymbol{Y}} = (0,5; 2,5; 0; 2,5; 0,5)'$, $S_e = 10$, $s^2 = 3,33$.*]*

# Úlohy k procvičení

**Příklad 5.3**

*Pomocí regresní přímky procházející počátkem spočítejte MNČ-odhady vektoru parametrů $\hat{\boldsymbol{\beta}}$, aproximace $\hat{\boldsymbol{Y}}$, reziduální součty čtverců $S_e$ a $s^2$ v LRM $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ pro data*

| $x$ | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| $Y$ | 0,18 | 0,35 | 0,48 | 0,65 | 0,84 | 0,97 |

*Jedná se o měření teplotní délkové roztažnosti měděné trubky. Rozdíl teploty od referenční 20 °C je $x$, prodloužení tyče je měřená veličina $Y$.*

$[\hat{\boldsymbol{\beta}} = 0,0164$, $\widehat{\boldsymbol{Y}} = (0,164; 0,328; 0,493; 0,657; 0,821; 0,985)'$, $S_e = 0,0015$, $s^2 = 0,0003.]$

# Úlohy k procvičení

## Příklad 5.4

*U 19 vzorků potravinářské pšenice byl zjišťován obsah zinku v zrnu (proměnná $Y$), v kořenech (proměnná $X_1$), v otrubách (proměnná $X_2$) a ve stonku a listech (proměnná $X_3$). Data jsou uložena v souboru „`psenice.Rdata`“.*

**a)** *Předpokládejte, že je vhodný regresní model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

*Odhadněte regresní koeficienty a rozptyl, vypočtěte vektor predikce a index determinace. Proveďte celkový $F$-test a dílčí $t$-testy. Hladinu významnosti volte $0,05$. Normalitu reziduí posuďte graficky pomocí funkce `qqnorm`.*

**b)** *Z regresního modelu odstraňte ty proměnné, jejichž regresní koeficienty se ukázaly nevýznamné pro $\alpha = 0,05$. Sestavte nový regresní model a proveďte v něm všechny úkoly z bodu a).*