

# REGRESNÍ ANALÝZA

Mgr. Jiří Mazurek Ph.D.

# Regresní analýza

- Regresní analýza se zabývá závislostí kvantitativního znaku na kvantitativním znaku (nebo více kvantitativních znacích).
- V případě závislosti jednoho znaku na jednom znaku mluvíme o *jednoduché regresi*.
- U závislosti jednoho znaku na více kvantitativních znacích hovoříme o *vícenásobné (nebo mnohonásobné) regresi*.

# PODSTATA REGRESNÍ ANALÝZY

- Jednou ze základních úloh regresní analýzy je najít vztah závislé proměnné  $y$  na faktorech  $\vec{x} = (x_1, x_2, x_3, \dots, x_k)$
- Tvar závislosti  $y$  na  $x \rightarrow$  regresní analýza
- Míra závislosti  $y$  na  $x \rightarrow$  korelační analýza

# Data

- Průřezová data: jednotlivá pozorování více jednotek v jednom časovém intervalu (příjem domácnosti, spotřební chování)
- Časové řady: pozorování proměnné za jednu časovou jednotku
- „Panelová“ data: kombinace průřezových dat a časových řad

# Proměnné: $y = f(x_1, x_2, x_3, \dots, x_k)$

$y$	$(x_1, x_2, x_3, \dots, x_k)$
Predictand	Predictors
Regressand	Regressors
Vysvětlovaná proměnná	Vysvětlující proměnné
Závislá proměnná	Nezávislé proměnné
Endogenní proměnná	Exogenní proměnné
Cílová proměnná	Kontrolní proměnné

# Regresní funkce

## Funkce – formát:

- lineární
- log-lineární
- semi-logaritmická
- kvadratická

(*tvar*)

$$y = \alpha + \beta \cdot x$$

$$\log y = \alpha + \beta \cdot \log x$$

$$\log y = \alpha + \beta \cdot x$$

$$y = \alpha + \beta \cdot \log x$$

$$y = \alpha + \beta \cdot x + \gamma \cdot x^2$$

## Formalizace

$\bar{x}$   $y$   $a$   $b$   $x$   $e$   $r$   
(pozorované)

“skuteční” proměnné

$\mu$   $\hat{y}$   $\hat{\alpha}$   $\hat{\beta}$   $X$   $\varepsilon$   $\rho$

odhadované proměnné

# Základní úloha

## Model jedoduché lineární regrese

$y = f(x)$ .....  $y = \alpha + \beta \cdot x + \varepsilon$  specifikace vztahu

Příklad:  $y =$  nákupy

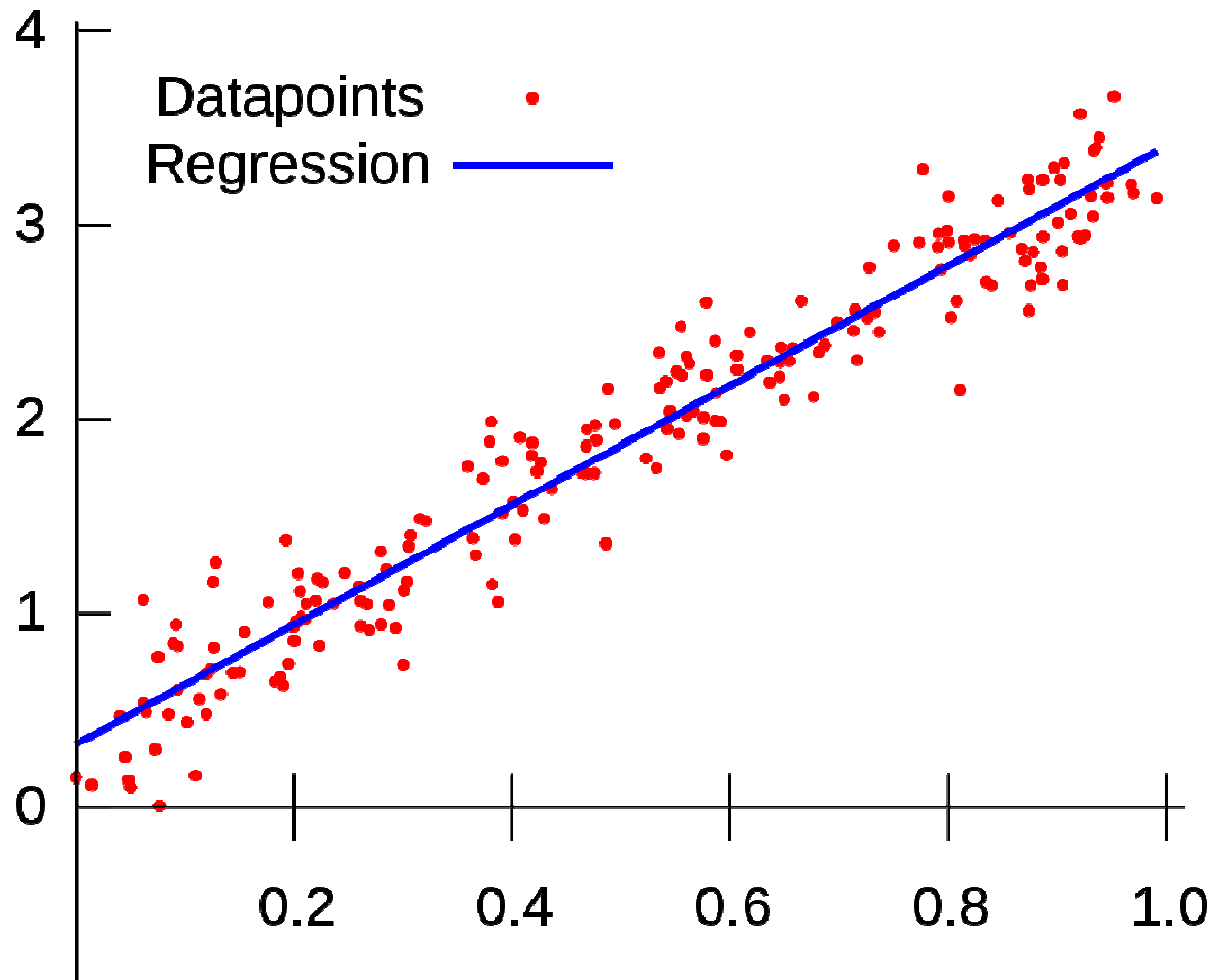
$x =$  příjem

$\varepsilon =$  reziduum (bílý šum)

Experiment: nastav  $x$ , pozoruj  $y$

Úkol: Jak přeložit přímku daty?

# Jak proložit danými body přímkou?





# Metoda nejmenších čtverců

---

Pro každý bod  $y_i = a + b.x_i + e_i, \quad i = 1, 2, \dots, n$

Chyba (*nepozorujeme*)  $e_i = y_i - a - b.x_i$

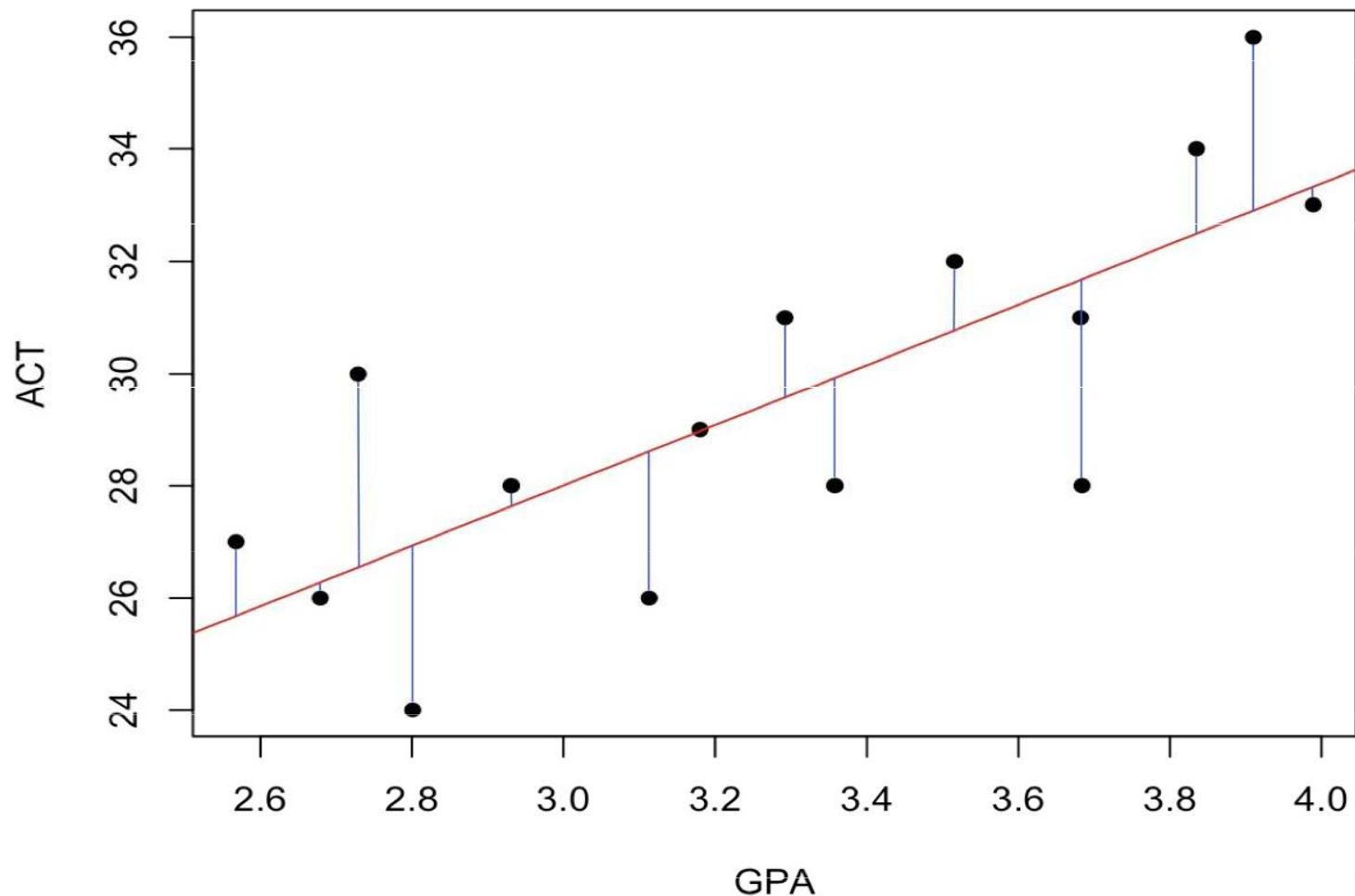
Potřebujeme minimalizovat chybu. Necht' střední hodnota je 0, potom musíme minimalizovat druhé mocniny chyby  $e^2$  (jméno - *LSE*)



reg\_front.svg

# Metoda nejmenších čtverců

GPA vs. ACT scores for 15 students



# Jednoduchá lineární regrese

$$\min\left(\sum_{i=1}^n (y - \alpha - \beta \cdot x_i)^2\right)$$

$$\Rightarrow \alpha, \beta = \arg \min\left(\sum_{i=1}^n (y_i - \alpha - \beta \cdot x_i)^2\right)$$

$$\left| \frac{\partial S}{\partial \alpha} = 0 = -2 \cdot \sum_{i=1}^n (y_i - \alpha - \beta \cdot x_i) \right.$$

$$\left. \frac{\partial S}{\partial \beta} = 0 = -2 \cdot \sum_{i=1}^n (y_i - \alpha - \beta \cdot x_i) \cdot x_i \right.$$

$$\sum y_i - n\alpha - \beta \cdot \sum x_i = 0$$

$$\sum y_i \cdot x_i - \alpha \cdot \sum x_i - \beta \cdot \sum x_i^2 = 0$$

$$\beta = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

# Jednoduchá lineární regrese

- Jednoduchá lineární regrese je speciálním případem vícenásobní regrese
- Jednoduchá lineární regrese má pouze jednu vysvětlující proměnnou, vícenásobná regrese má dvě nebo více vysvětlujících proměnných

# Vícenásobná lineární regrese

- Nejčastěji odhadovaná funkce

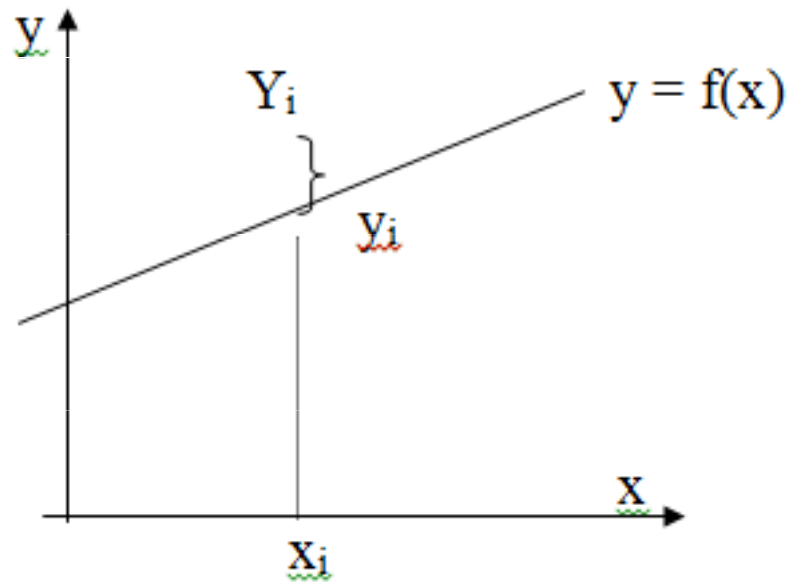
$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- Nejjednodušším případem je jednoduchá lineární regrese

$$y = \beta_1 + \beta_2 x_2.$$

# Grafická interpretace

- $Y_i$  = empirické (měřené) hodnoty závislé proměnné,
- $y_i$  = teoretické hodnoty závislé proměnné,
- $\varepsilon_i$  = residua.
- Vztah mezi  $Y_i$  a  $y_i$ :  $y_i = Y_i + \varepsilon_i$



# Předpokládané statistické vlastnosti náhodné složky

1. Střední hodnota  $\varepsilon_i$  je nula, tj.  $E(\varepsilon_i) = 0$  pro každé  $i$ .
2. Rozptyl  $\varepsilon_i$  je konstantní, nezávislý na  $i$ , tj.  $Var(\varepsilon_i) = \sigma^2$  pro každé  $i$ .
3. Veličiny  $\varepsilon_i, \varepsilon_j$  jsou nekorelované, tj.  $Cov(\varepsilon_i, \varepsilon_j) = 0$  pro  $i \neq j$ .
4. Veličiny  $\varepsilon_i$  mají normální rozdělení, tj.  $\varepsilon_i \sim N(0, \sigma^2)$  pro každé  $i$ .

# Regresní koeficienty

- Vektor regresních koeficientů získáme z vektorové rovnice :

$$\vec{b}^T = (X^T \cdot X)^{-1} X^T \cdot \vec{Y},$$

- kde  $X$  je tzv. matice regresorů

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

$$\vec{Y} = (Y_1, Y_2, \dots, Y_n)^T$$



# Příklad

- Odhadněte závislost spotřeby elektrické energie ( $Y$ ) na délce elektrického vedení ( $X_1$ ) a odběru energie ( $X_2$ ). Jsou k dispozici následující výběrová data:

$X_1$	$X_2$	$Y$
1,2	3,6	3,2
1,3	3,7	3,3
1,3	3,8	3,4
1,4	3,8	3,5
1,4	3,9	3,6
1,5	3,9	3,6
1,5	4	3,7
1,6	4	3,8
1,6	4,1	3,9
1,7	4,2	4

# Příklad - řešení

- Tabulka představuje body, z nichž získáme potřebné matice  $X$  a  $Y$ .

$$X = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix}$$

$$\bar{Y} = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix}$$

# Příklad - řešení

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} = \begin{bmatrix} 10 & 14,5 & 39 \\ 14,5 & 21,25 & 56,8 \\ 39 & 56,8 & 152,4 \end{bmatrix}$$

# Příklad - řešení

$$(X^T \cdot X)^{-1} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix}$$

$$X^T \cdot Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix} = \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix}$$

# Příklad - řešení

$$\vec{b}^T = (X^T \cdot X)^{-1} \cdot X^T \cdot \vec{Y} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} \cdot \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix} = \begin{bmatrix} -0,78 \\ 0,60 \\ 0,90 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

Hledaná regresní funkce má tedy rovnici  $\vec{Y} = -0,78 + 0,60x_1 + 0,90x_2$ .

# Teoretické hodnoty

- Teoretické hodnoty obdržíme dosazením do regresní rovnice za  $x_1$  a  $x_2$  postupně z tabulky vstupních dat:

$$\widehat{Y}_1 = -0,78 + 0,60 \cdot 1,2 + 0,90 \cdot 3,6 = 3,18,$$

$$\widehat{Y}_2 = -0,78 + 0,60 \cdot 1,3 + 0,90 \cdot 3,7 = 3,33,$$

...

$$\widehat{Y}_{10} = -0,78 + 0,60 \cdot 1,7 + 0,90 \cdot 4,2 = 4,02.$$



# Vektor reziduálních odchylek:

- Rozdíl teoretické a skutečné hodnoty, představuje *vektor reziduálních odchylek*:

$$\bar{e} = \bar{Y} - \hat{Y} = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,9 \\ 3,9 \\ 4 \end{bmatrix} - \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} 0,02 \\ -0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ -0,03 \\ 0,02 \\ 0,02 \\ 0,03 \\ -0,02 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ e_{10} \end{bmatrix}$$



# Rozptyl odhadu regresních koeficientů

- Protože při výpočtu regresních koeficientů se jedná o odhady, je účelné také nalézt rozptyly těchto odhadů, které vyjadřují přesnost odhadů. Získáme je jako prvky hlavní diagonály matice:

$$\text{Var}(\bar{b}) = s^2 \cdot (X^T \cdot X)^{-1},$$

kde  $s^2 = \frac{\sum_{i=1}^n e_i^2}{n-k}$  je odhad rozptylu veličiny  $\varepsilon$ . Přitom

$e_i$  =  $i$ -tá reziduální odchylka,

$n$  = počet bodů,

$k$  = počet parametrů regresního modelu.

# Příklad - řešení

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-k} = \frac{0,006}{10-3} = 0,0008571.$$

$$\text{Var}(\vec{b}) = s^2 \cdot (X^T \cdot X)^{-1} = 0,0008571 \cdot \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} = \begin{bmatrix} 0,2102 & 0,0926 & -0,0883 \\ 0,0926 & 0,0514 & -0,0429 \\ 0,0883 & -0,0429 & 0,0386 \end{bmatrix}.$$

# Rozptyly regresních koeficientů

- Diagonálu poslední matice tvoří rozptyly jednotlivých regresních koeficientů:

$s^2(b_0) = 0,2102$ , odtud směrodatná odchylka je  $s(b_0) = 0,4584$ .

$s^2(b_1) = 0,0514$ , odtud směrodatná odchylka je  $s(b_1) = 0,2267$ .

$s^2(b_2) = 0,0386$ , odtud směrodatná odchylka je  $s(b_2) = 0,1965$ .

Po nalezení regresního modelu a rozptylů odhadů regresních koeficientů píšeme obvykle výsledné řešení tak, že pod regresní koeficienty do závorek uvádíme příslušné směrodatné odchylky (též tzv. *standardní chyby*).

$$\hat{Y} = -0,78 + 0,60 x_2 + 0,90 x_3$$

$(0,4584) \quad (0,2267) \quad (0,1965)$

# TEST VÝZNAMNOSTI REGRESNÍCH KOEFICIENTŮ

- Při výpočtu regresních koeficientů  $b_1, b_2, \dots, b_k$  se stává, že mezi koeficienty jsou až řádové rozdíly, např.  $b_1 = 200$  a  $b_2 = 0,02$ .
- V takových případech stojíme před problémem, zda má smysl zařadit např.  $b_2$  do regresní funkce.
- K objektivnímu posouzení významnosti regresních koeficientů lze použít test statistické významnosti regresních koeficientů.

# Struktura testu

- 1) Nulová hypotéza:  $H_0: \beta_i = 0$ , alternativní hypotéza  $H_1: \beta_i \neq 0$
- 2) Testové kritérium

$$T = \frac{b_i}{s(b_i)}$$

- Kde  $b_i$  je odhad parametru  $\beta_i$ ,  $s(b_i)$  je směrodatná odchylka odhadu  $b_i$ .
- 3) Kritická hodnota  $t_{n-k}(\alpha)$
- 
- 4) Porovnáme  $T$  a  $K$ : Je-li  $|T| > K$ , zamítá se  $H_0$  a přijme se alternativní hypotézu  $H_1$ , podle které vypočítaný koeficient je možné považovat za nenulový, neboli statisticky významný a je proto důvod pro jeho zařazení do regresní funkce. V opačném případě přijímáme  $H_0$  a parametr považujeme za statistický nevýznamný.

# Příklad - řešení

- 1) Nulová hypotéza:  $H_0: \beta_i = 0$ , alternativní hypotéza  $H_1: \beta_i \neq 0$
- 2) Testové kritérium

$$T_1 = \frac{b_1}{s(b_1)} = \frac{0,60}{\sqrt{0,0514}} = 2,65,$$

$$T_2 = \frac{b_2}{s(b_2)} = 4,58,$$

- 3)  $t_{n-p}(\alpha) = t_{10-3}(0,05) = 2,365$ .
- 4) Protože  $T_1 > 2,365$  a také  $T_2 > 2,365$ , jsou oba regresní koeficienty statisticky významné a nenulové, a proto je oba zařadíme do regresní funkce.
-

# INTERVALY SPOLEHLIVOSTI PRO REGRESNÍ KOEFICIENTY

- Intervaly spolehlivosti pro parametry  $\beta_1, \dots, \beta_k$ , tj. intervaly, ve kterých lze očekávat tyto parametry s pravděpodobností  $1-\alpha$ , získáme pomocí vztahu:

$$[b_i - t_{n-p}(\alpha) \cdot s(b_i), b_i + t_{n-p}(\alpha) \cdot s(b_i)],$$

- kde

$b_i$  = odhad parametru  $\beta_i$ ,

$s(b_i)$  = směrodatná odchylka odhadu  $b_i$ ,

$t_{n-p}(\alpha)$  = kritická hodnota Studentova rozdělení,

$n$  = počet bodů,

$p$  = počet parametrů modelu,

$\alpha$  = hladina významnosti,

# TESTOVÁNÍ VHODNOSTI REGRESNÍHO MODELU

- Vhodnost volby regresního modelu (tj. volby nezávisle proměnných) se ověří testem. Test má následující strukturu:

- 1)  $H_0 : \vec{\beta} = \vec{0}.$

$$H_1 : \vec{\beta} \neq \vec{0}.$$

$$T = \frac{S_{\hat{Y}} / (k)}{S_e / (n - k - 1)},$$

- 2) Testové kritérium

$$S_{\hat{Y}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i,$$

$$S_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2.$$

- 3) Kritická hodnota Fischerova rozdělení  $K = F_{k, n-k-1}(\alpha)$
- 4) Je-li  $T \geq K$ , pak se  $H_0$  zamítá. V opačném případě se  $H_0$  nezamítá.



# Příklad - řešení

- Použijeme-li test na náš příklad, obdržíme:

$$T = \frac{0,594/(3-1)}{0,006/(10-3)} = 346,5, K = F_{3,10-3-1}(0,05) = 4,757.$$

- Protože  $T$  překročilo kritickou hodnotu  $K$ , zamítá se  $H_0$  a model se považuje za vyhovující, tj. zamítá se hypotéza o nulovosti všech regresních koeficientů (s výjimkou  $\beta_0$ ). Testové kritérium překročilo kritickou hodnotu výrazně a stalo by se tak i na jednoprocenní hladině významnosti.

Děkuji za pozornost.