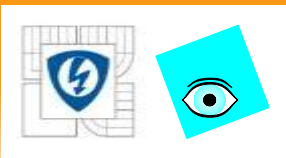


Statistická analýza jednorozměrných dat

Prof. RNDr. Milan Meloun, DrSc.
Univerzita Pardubice, Pardubice

31.ledna 2011

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.



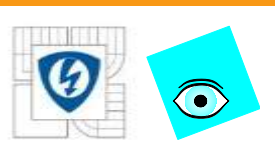
Kapitola 6.1

STATISTICKÁ ZÁVISLOST

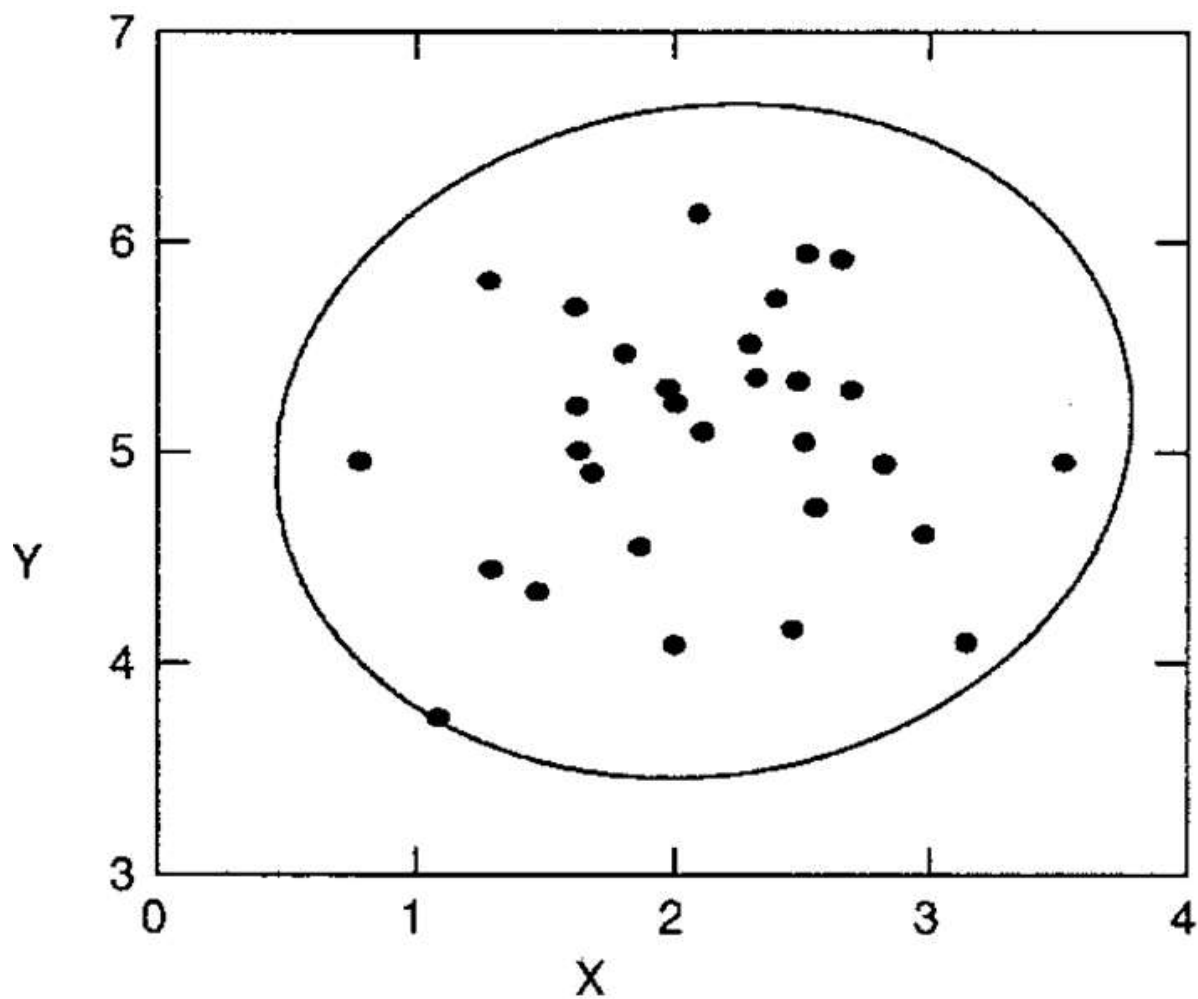
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





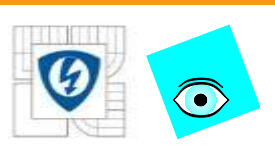
Nekorelovaná data



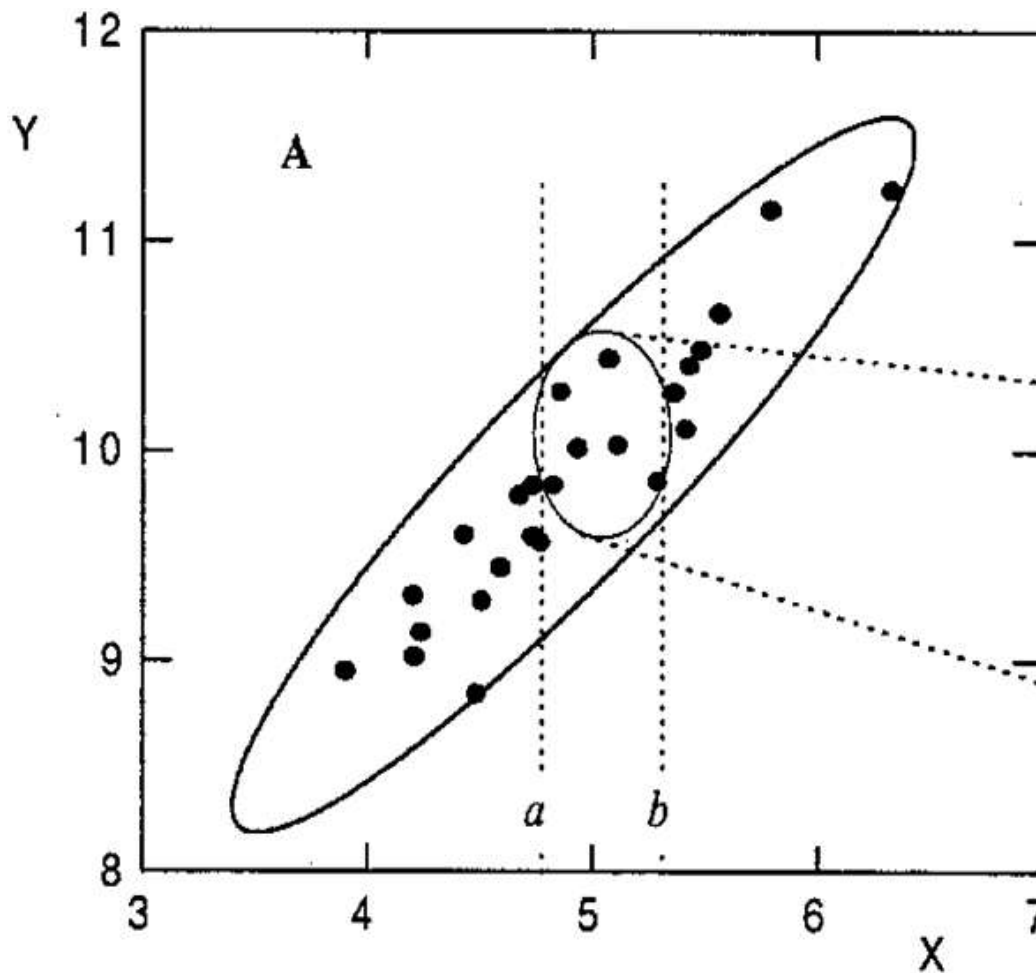
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

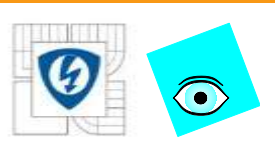




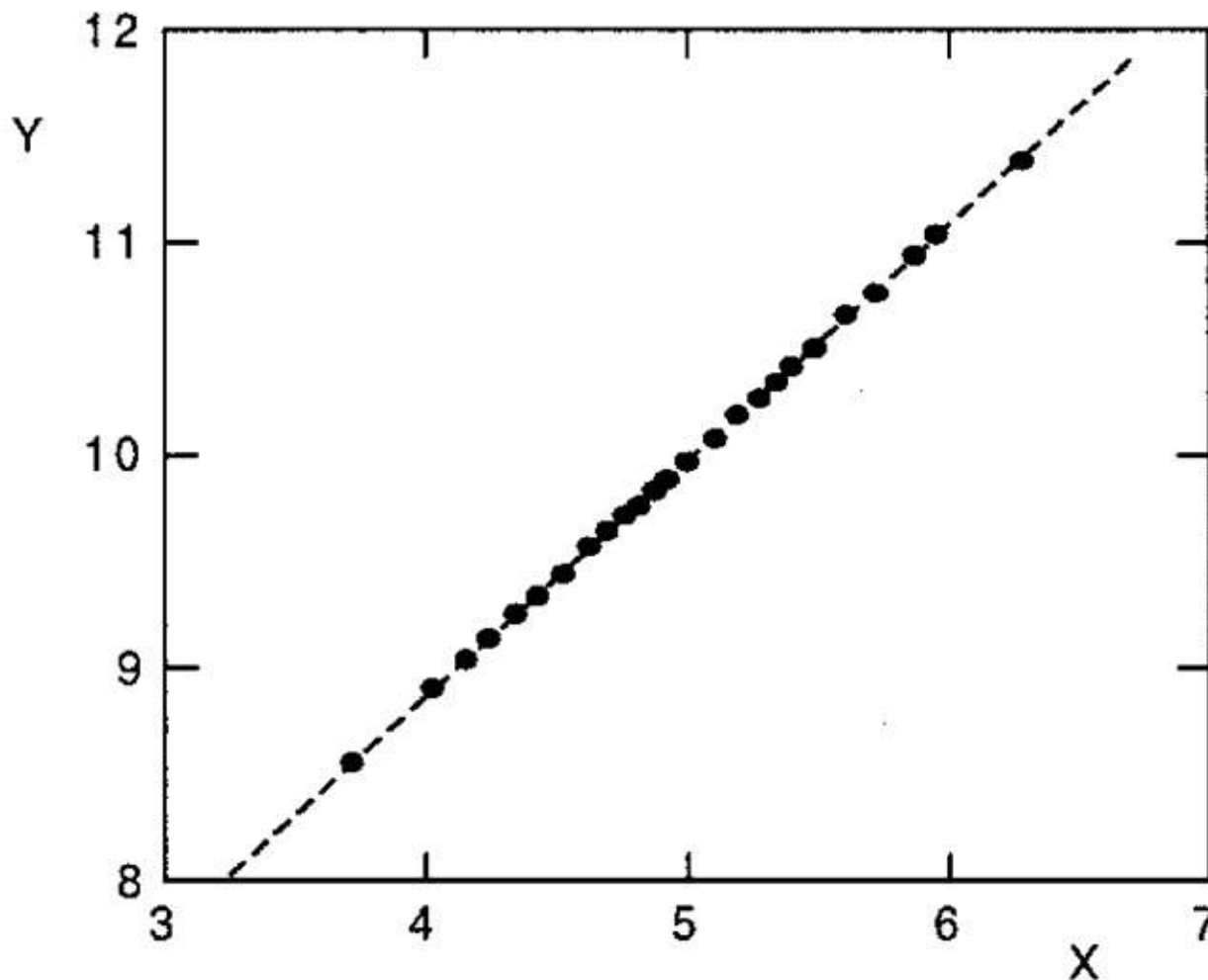
Pozitivní korelace



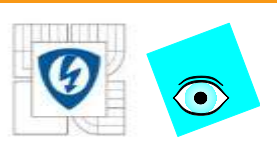
Pokud však budeme měřit data v příliš malém intervalu, nemusí se závislost vůbec prokázat!!



Lineární závislost



Když je jedna proměnná násobkem druhé, je pak možné jednu proměnnou z analýzy vyloučit, a to **bez ztráty informace**.



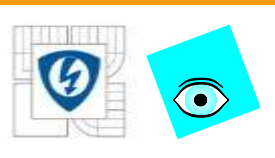
Cíl regresní analýzy

Cílem regresní analýzy je nalezení vhodného modelu studované závislosti tak, že se **snažíme nahradit**

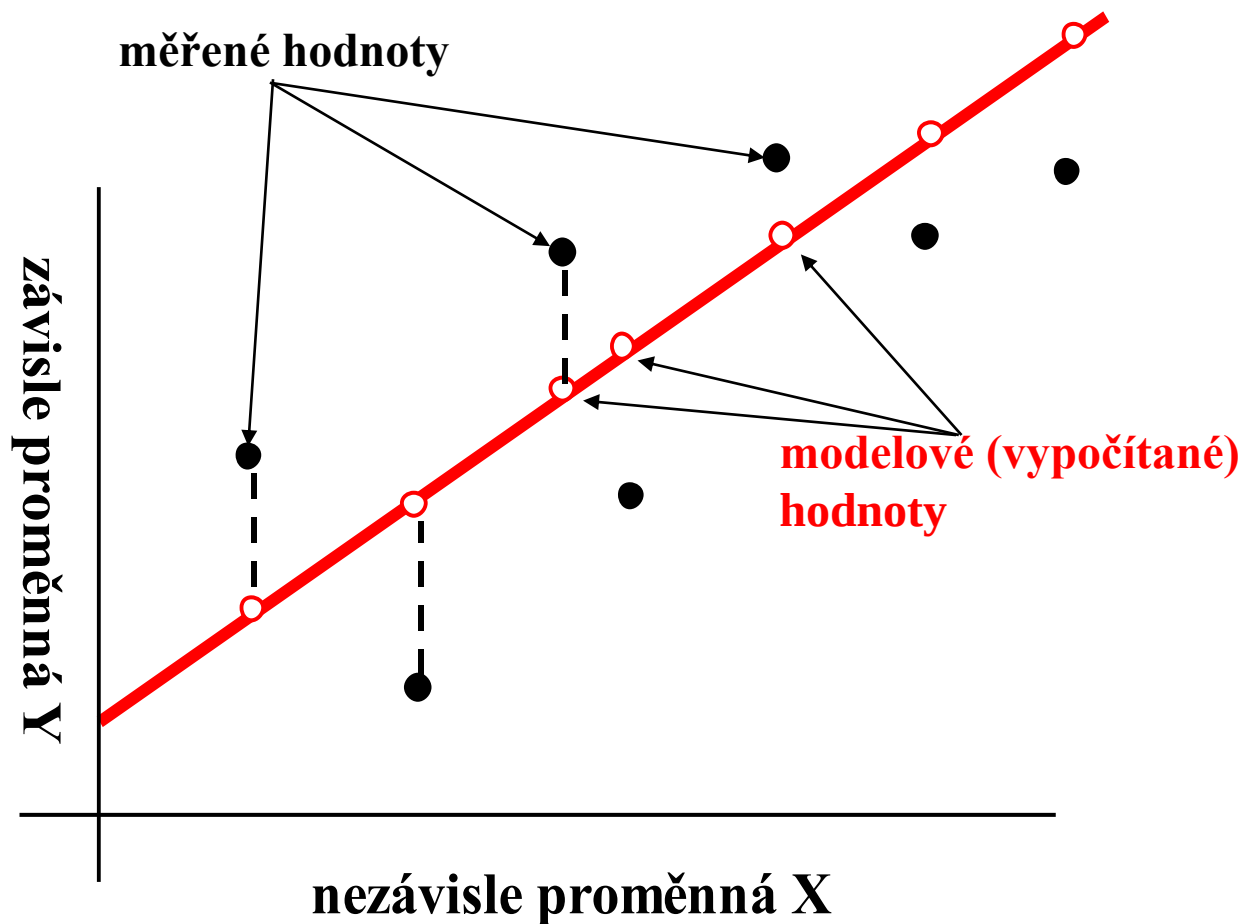
každou měřenou (experimentální) hodnotu
závisle proměnné y_{exp}

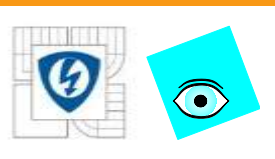
hodnotou vypočtenou (predikovanou) y_{vyp}

čili **hodnotou** ležící na spojitě funkci (modelu) **nezávisle**
proměnné x .



Grafické vysvětlení



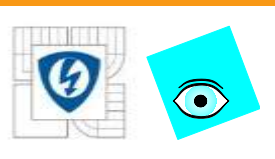


Definice regresního modelu

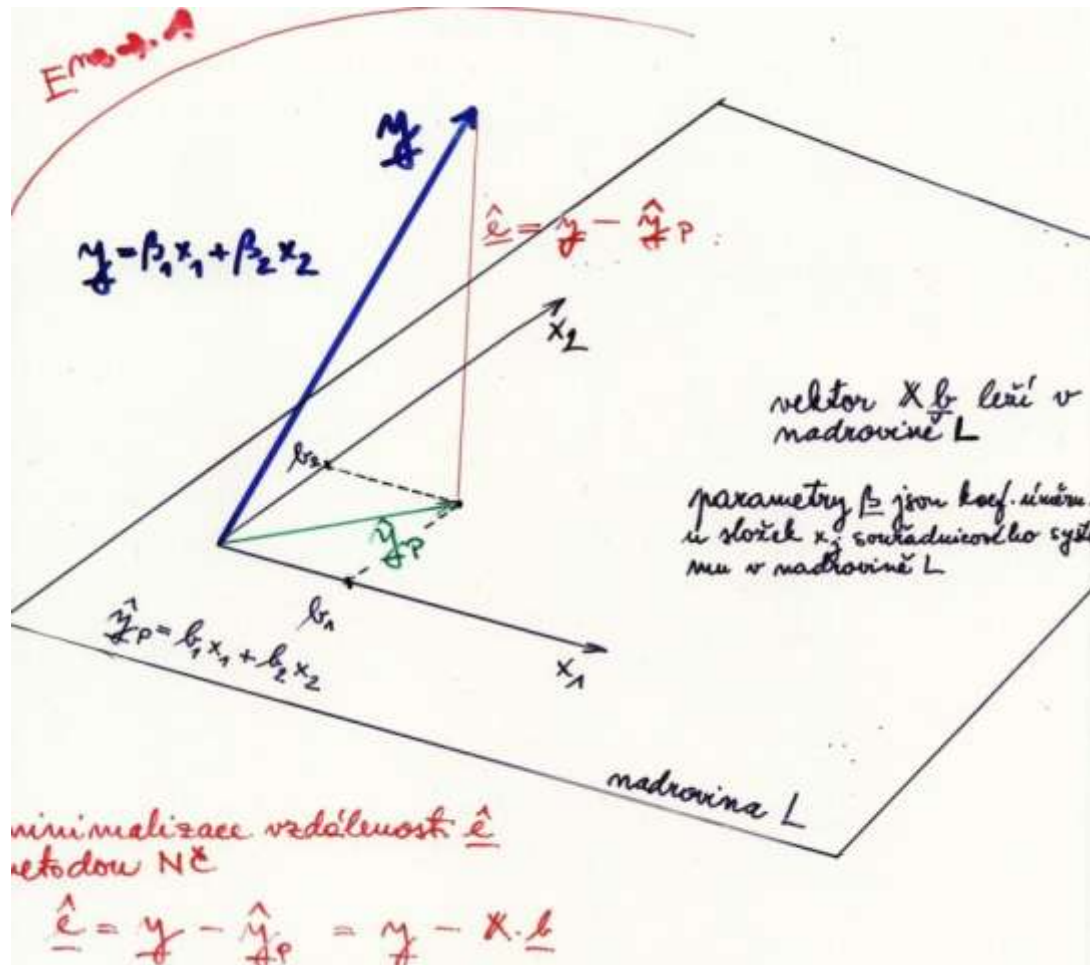
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix}}_{\mathbf{X}} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

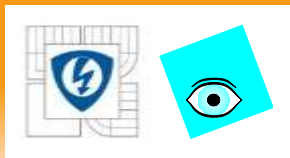
$\mathbf{y}$ závisle proměnná $\mathbf{X}$ nezávisle proměnná $\boldsymbol{\beta}$ regresní parametry $\boldsymbol{\varepsilon}$ náhodná chyba

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

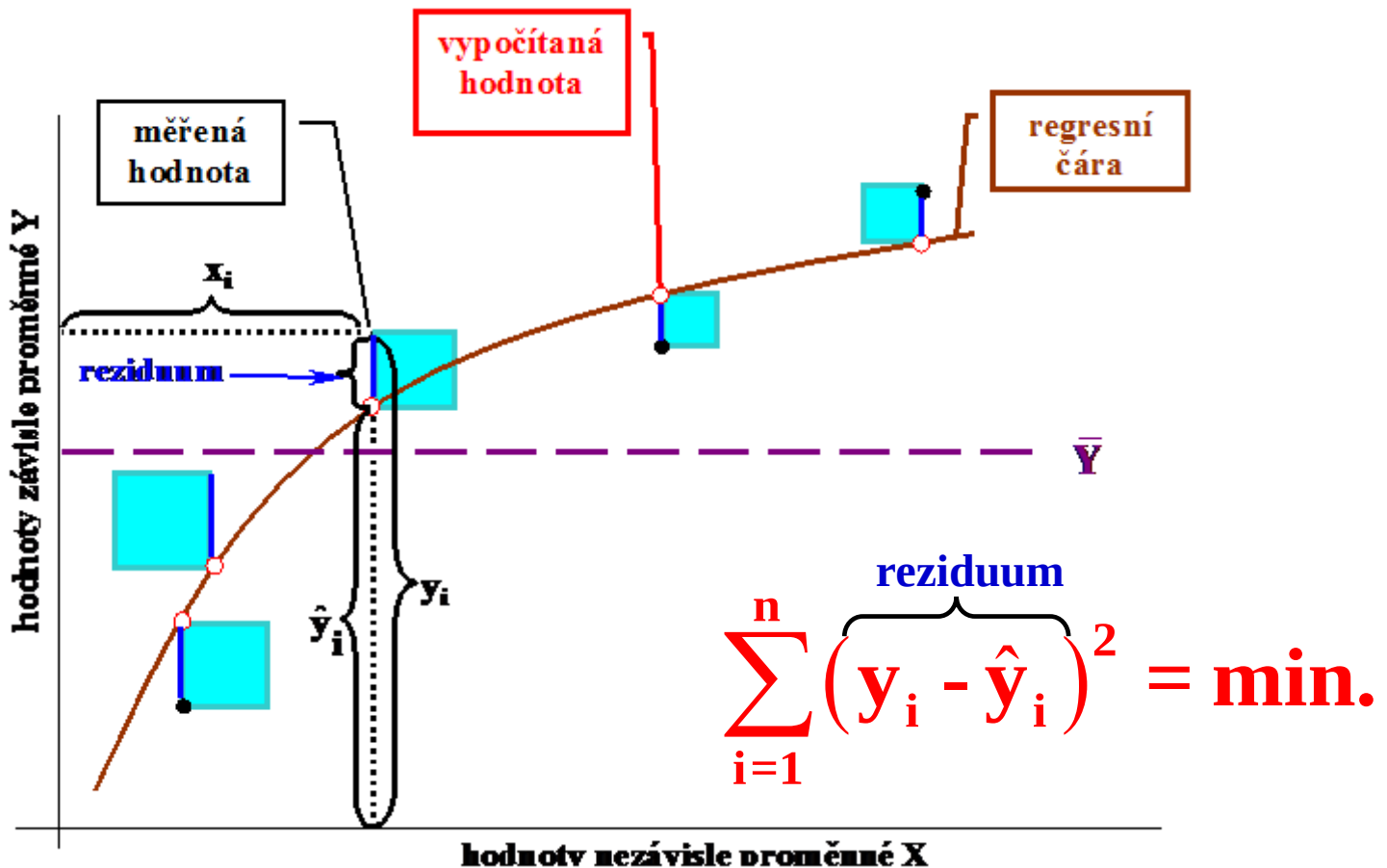


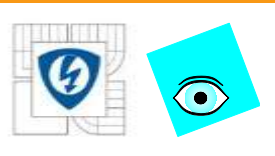
Geometrický význam





Vyčíslení odhadů parametrů regresního modelu metodou nejmenších čtverců

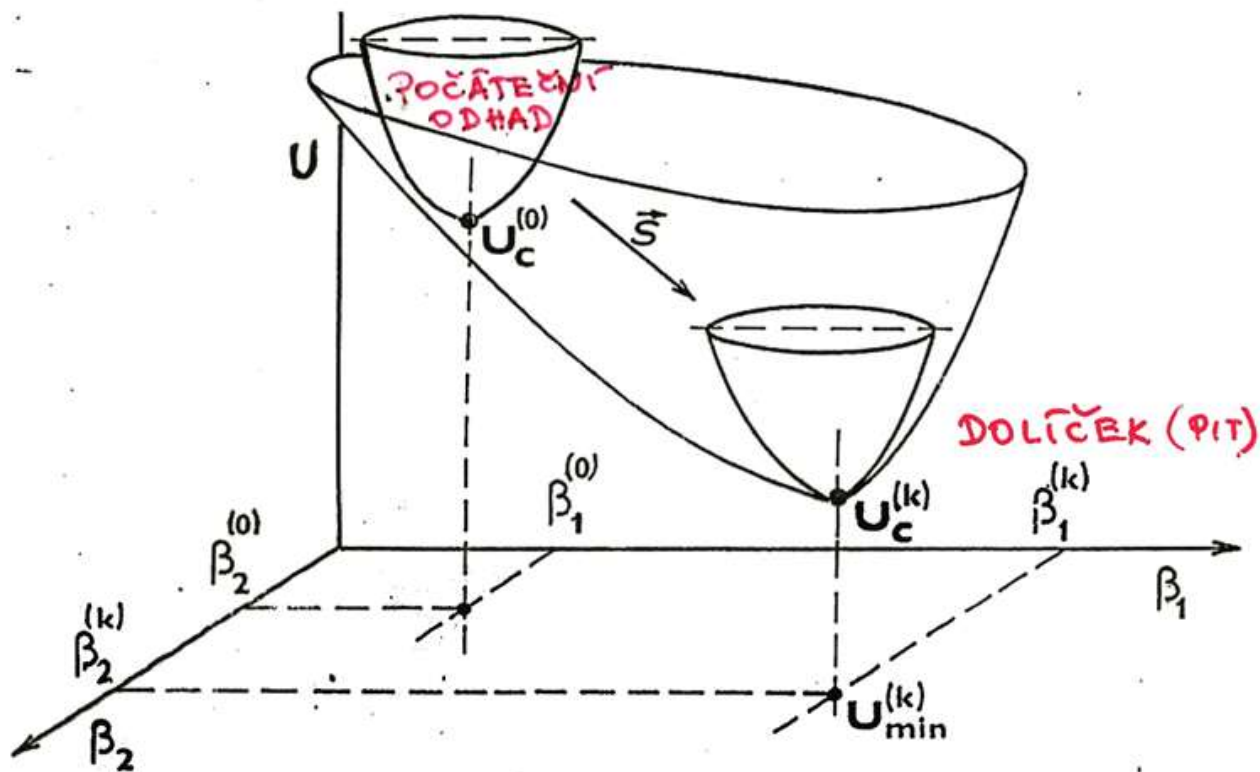


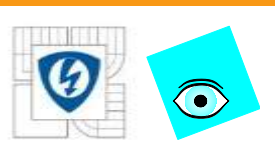


Geometrické znázornění kritéria U

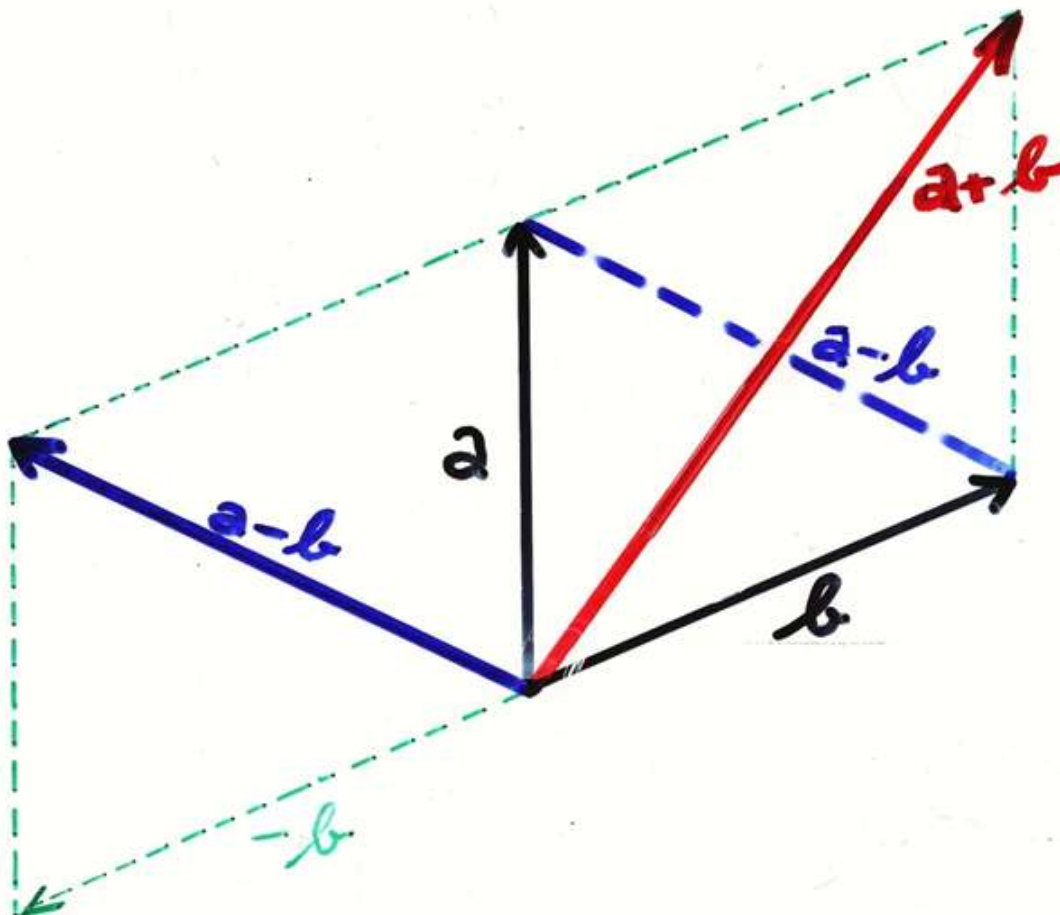
optimalizační problém v $(m+1)$ rozměrném prostoru

$$U = \sum_{i=1}^n w_i (y_{exp} - y_{vyp,i})^2 \approx \text{minimum}, \text{ kde}$$
$$y_{vyp,i} = f(x_i, \beta_1, \dots, \beta_m, p_1, \dots, p_m)$$





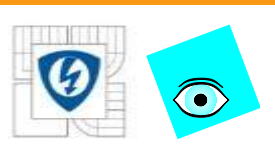
Sčítání a odčítání vektorů



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Metoda projekčních matic

y vektor proměnné

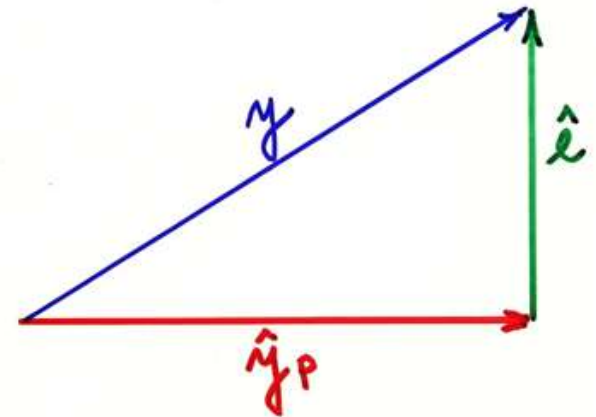
y_P vektor predikce

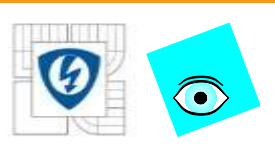
$\hat{e}$ vektor reziduí

Cílem: minimální délka vektoru $\hat{e} = y - y_P$

délku vektoru vyjádříme $D = \sqrt{\langle \hat{e}, \hat{e} \rangle} = \sqrt{\sum_{i=1}^n \hat{e}_i^2}$

Odhady modelových parametrů β minimalizují čtverec délky vektoru $D^2 = \sum_{i=1}^n [y_i - \sum_{j=1}^m x_{ij} b_j]^2 = \sum_{i=1}^n (y_i - \hat{y}_{P,i})^2$





Důsledek

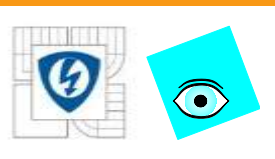
Vektor reziduí $\hat{e}$ je kolmý na všechny sloupce matice X ,
a proto odpovídající skalární součiny jsou nulové

$$\langle x_j, \hat{e} \rangle = \sum_{i=1}^n x_{ij} \hat{e}_i = 0, \quad j = 1, \dots, m$$

a maticově $X^T \hat{e} = 0$

a dosazením $\hat{e} = y - Xb$ bude

$$X^T y = X^T X b$$



Důsledek

Odhad b minimalizující vzdálenost D má tvar

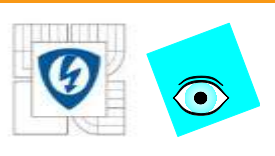
$$b = (X^T X)^{-1} X^T y$$

projekční matice H , pro kterou platí

$$y_P = Hy$$

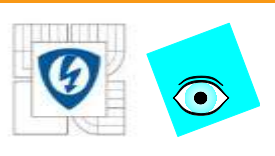
a pomocí vektoru b

$$y = Xb = X(X^T X)^{-1} X^T y$$



Vlastnosti projekční matice H

1. Projekční matice $H = X(X^T X)^{-1}X^T$ promítne libovolný vektor V do roviny L .
2. Pokud vektor V již leží v rovině L , platí $HV = V$.
3. Pokud je však vektor V kolmý na rovinu L , platí $HV = 0$, kde 0 je nulový vektor.

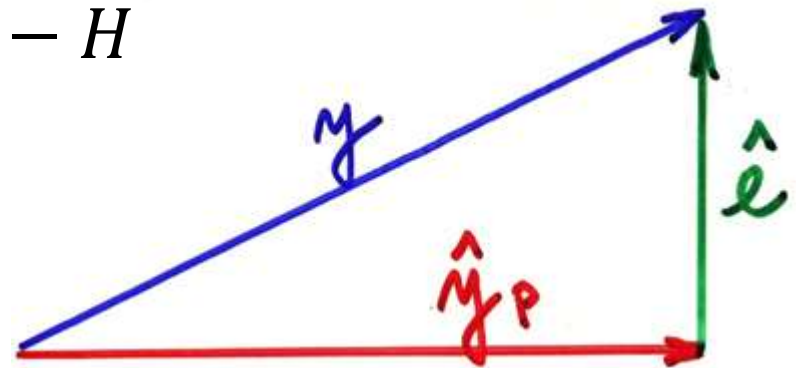


Vlastnosti projekční matice P

Projekční matice P pro kolmou projekci do nadroviny $L^\perp$ kolmé na nadrovinu L , má tvar

$$P = E - H$$

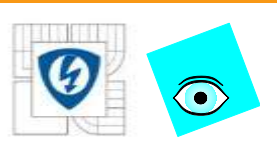
kde E je jednotková matice



Závěr: Rozklad vektoru y do dvou složek

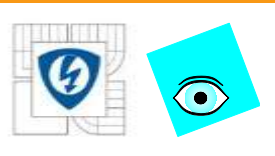
$$y = Hy + Py = \hat{y}_P + \hat{e}$$

Geometricky: vektor y byl rozložen na dva kolmé vektory



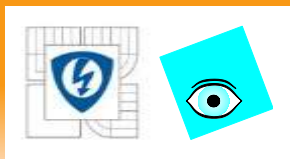
Předpoklady metody nejmenších čtverců

1. Parametry β mohou nabývat libovolných hodnot. Omezení jsou pouze fyzikálního smyslu.
2. Model je lineární v parametrech β ; platí při tom aditivní model měření $y = X\beta + \varepsilon$
3. Matice X je nenáhodná, nastavitelných hodnot nezávisle proměnných. Má hodnot m :
 - (a) žádné dva sloupce x_i a x_k nejsou kolineární (čili paralelní).
 - (b) $(X^T X)$ je pak symetrická regulární matice.
 - (c) Rovina L je m rozměrná a vektory $X b$ jsou jednoznačné.



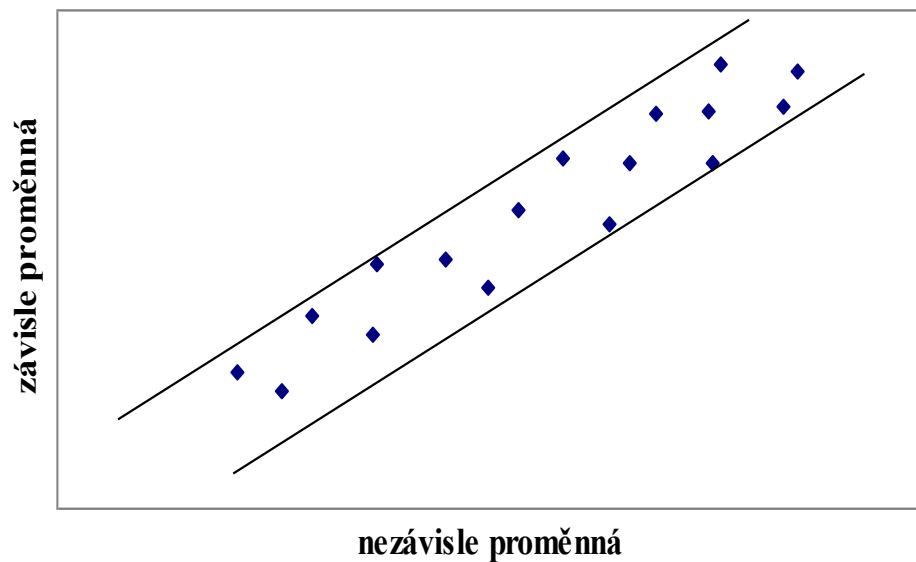
Předpoklady metody nejmenších čtverců

4. Náhodné chyby ε_i mají nulovou střední hodnotu $E(\varepsilon_i) = 0$. Je-li $E(\varepsilon_i) = K$, je nutno zavést absolutní člen a pak bude $E(\varepsilon_i) = 0$.
5. Náhodné chyby ε_i mají konstantní rozptyl, homoskedasticita, $E(\varepsilon_i^2) = \sigma^2$.
6. Náhodné chyby ε_i jsou vzájemně nekorelované, $\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0$.
7. Náhodné chyby mají normální rozdělení $\varepsilon \approx N(0, \sigma^2)$.

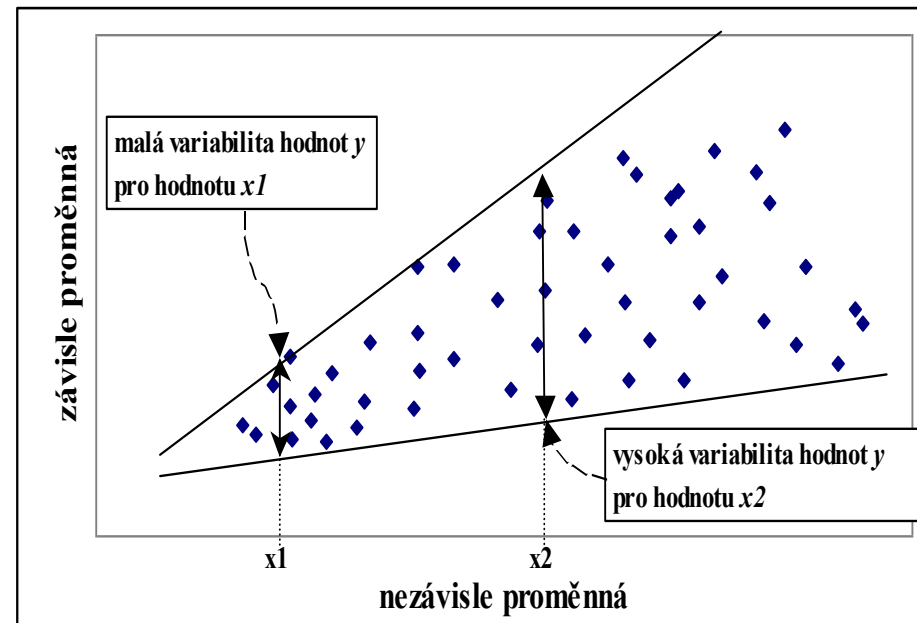


Homoskedasticita vs. heteroskedasticita

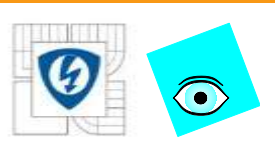
Homoskedasticita znamená, že hodnoty závisle proměnné y mají pro všechny hodnoty nezávisle proměnné x konstantní rozptyl, čili konstantní variabilitu-proměnlivost.



homoskedasticita



heteroskedasticita



Test homoskedasticity

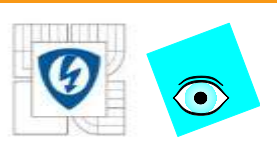
Pomocí testu trendu reziduí

$$D = \sum_{i'=+}^n [R(|\hat{e}_i|) - i]^2$$

$$\rho_s = 1 - \frac{6}{n^3 - n} \cdot D$$

testujeme významnost Spearmanova korelačního koeficientu ρ_s

$$t_R = \frac{\rho_s \cdot \sqrt{n - 2}}{\sqrt{1 - \rho_s^2}}$$

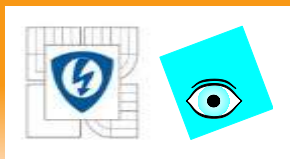


Pomocí Cookova - Weisbergova testu

Vycházíme z předpokladu, že rozptyl naměřené hodnoty y_i je určitou funkcí proměnné x_i β (např. exponenciální funkcí)

$$S_f = \frac{[\sum_{i=1}^n (y'_i - \bar{y}') e_i^2]^2}{1 \cdot \sigma^4 \sum_{i=1}^n (y'_i - \bar{y}')$$

Pravidlo: pokud v datech není heteroskedasticita, potom platí, že $S_f < \chi^2(1)$.

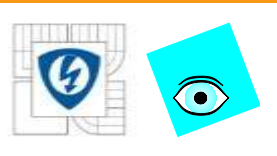


Využití vícenásobného korelačního koeficientu R

$$R = \sqrt{\frac{TSC}{CSC}} = \sqrt{1 - \frac{RSC}{CSC}}$$
$$= \cos \alpha = \sqrt{1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

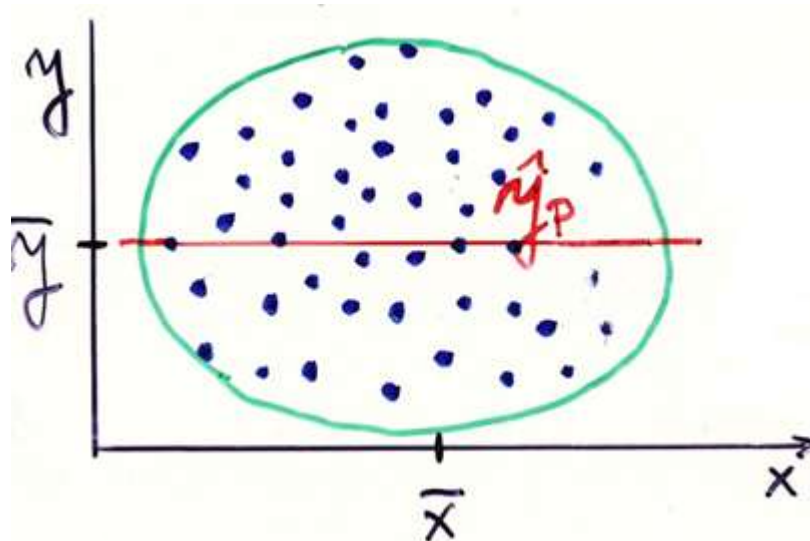
Diskuze:

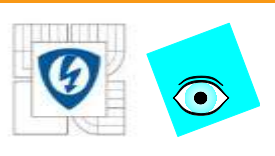
(a) $\hat{R} \rightarrow 0$, pak platí $H_0: \hat{y}_P = J \cdot \bar{y}$ čili $H_0: \beta = 0$



Využití vícenásobného korelačního koeficientu R

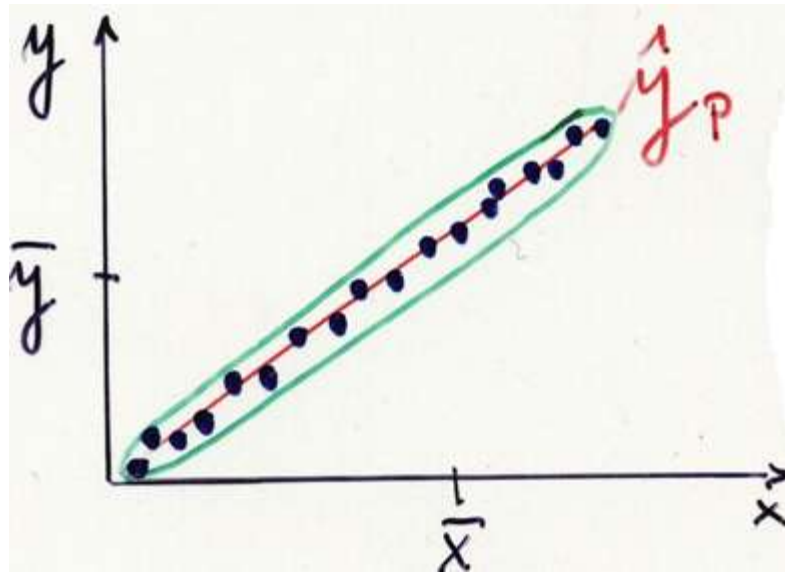
$\hat{R} \rightarrow 0, \beta = 0$, všechny regresní koeficienty jsou rovny nule





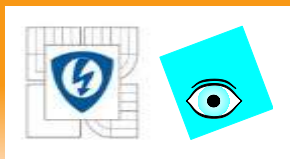
Využití vícenásobného korelačního koeficientu R

(b) $\hat{R} \rightarrow 1$, pak platí $H_A: \hat{y}_P = X \cdot \beta$ čili $H_A: \beta \neq 0$



$R \rightarrow 1, \beta \neq 0$ všechny regresní koeficienty jsou různé od nuly

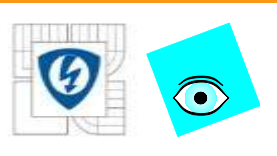
Testování R : $H_0: \beta = 0$ čili $\hat{R}^2 = 0$ vs. $H_A: \beta \neq 0$ čili $\hat{R}^2 \neq 0$



Využití vícenásobného korelačního koeficientu R

$$F_R = \frac{(CSC - RSC)(n - m)}{RSC(m - 1)} = \frac{\hat{R}^2(n - m)}{(1 - \hat{R}^2)(m - 1)}$$

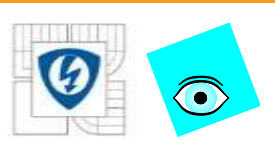
v porovnání s $F_{1-\alpha}(m - 1, n - m)$



Diskuze korelačního koeficientu R

1. R se blíží k nule: parametry regresního modelu kromě absolutního členu jsou rovny nule a platí model M_0 .
2. R se blíží k ± 1 : parametrynejsou rovny nule a platí model M_1 .

U regresních modelů nelze použít R k posouzení linearity ani k posouzení kvality regresního modelu.



Koeficient determinace R^2 (x100%)

Je čtverec vícenásobného korelačního koeficientu. Vyjadřuje procento bodů, popsaných regresním modelem.

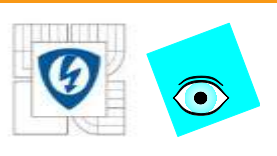
Testování: $H_0: \beta_c = 0$ nebo $H_0: R^2 = 0$

Testační kritérium F-testu

$$F_R = \frac{(CSC - RSC)(n - m)}{RSC(m - 1)} = \frac{\hat{R}^2(n - m)}{(1 - \hat{R}^2)(m - 1)}$$

Je-li $F_R < F_{1-\alpha}(m - 1, n - m)$, je H_0 přijata.

Významnost korelačního koeficientu je shodná s testem významnosti všech regresních koeficientů vyjma absolutního členu.

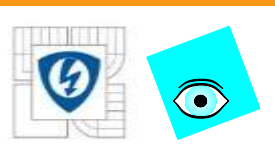


Příklad: Omezení klasické analýzy lineárního modelu

Anscomb⁵ uvádí testační data pro čtyři simulované výběry. Testujte statistickou významnost obou parametrů β_1 a β_2 a proveďte grafickou analýzu reziduí.

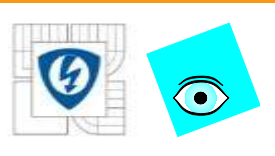
Data: čtyři simulované výběry vykazují stejné charakteristiky

$$b_1 = 0.5, b_2 = 3.0, D(b_1) = 0.0139, D(b_2) = 1.2656.$$



Pokračování příkladu

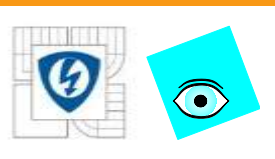
Výběr		A	B	C	D	
Proměnná	x	y	y	y	x	y
1	10	8.04	9.14	7.46	8	6.58
2	8	6.95	8.14	6.77	8	5.76
3	13	7.58	8.74	12.74	8	7.71
4	9	8.81	8.77	7.11	8	8.84
5	11	8.33	9.26	7.81	8	8.47
6	14	9.96	8.10	8.84	8	7.04
7	6	7.24	6.13	6.08	8	5.25
8	4	4.26	3.10	5.39	19	12.50
9	12	10.84	9.13	8.15	8	5.56
10	7	4.82	7.26	6.42	8	7.91
11	5	5.68	4.74	5.73	8	6.89



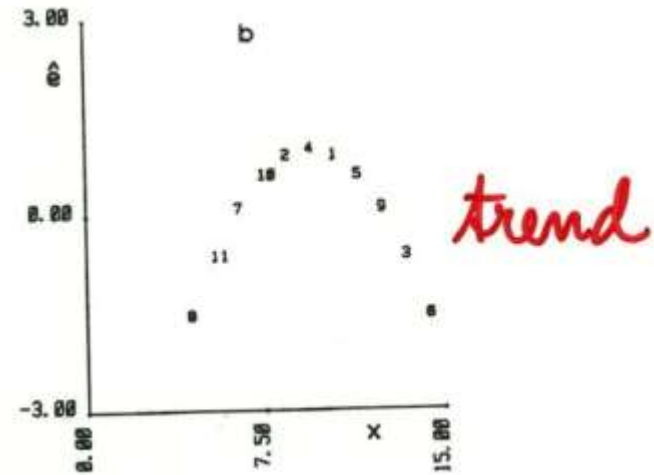
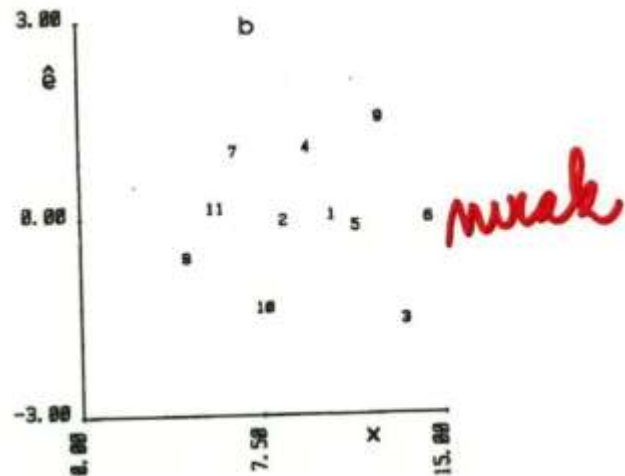
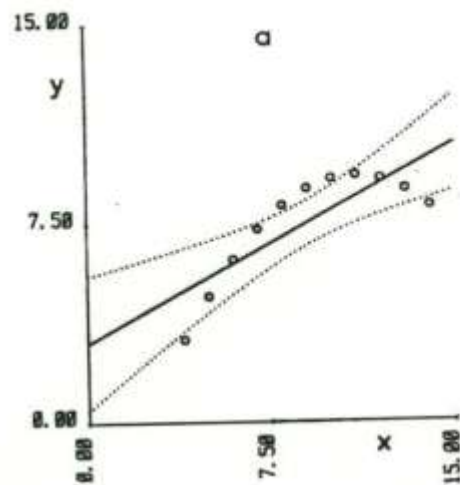
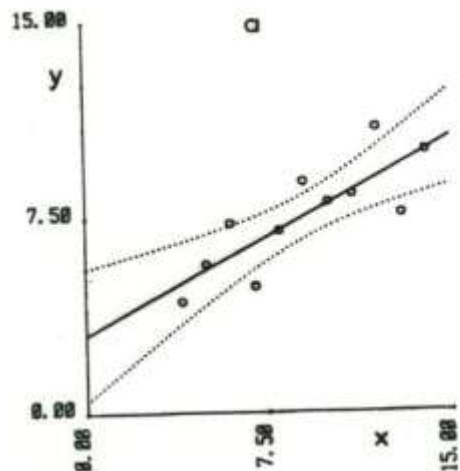
Pokračování příkladu

Řešení: Lineární regresní model $E\left(\frac{y}{x}\right) = \beta_1 x + \beta_2$ má pro všechny výběry

- a) stejné odhady parametrů: $b_1 = 0.5, b_2 = 3.0, D(b_1) = 0.0139, D(b_2) = 1.2656,$
- b) stejné testační charakteristiky významnosti parametrů: $T_1 = 2.667$ a $T_2 = 4.241,$
- c) stejné testační charakteristiky: $F_R = 17.97, \hat{R}^2 = 0.66, \hat{\sigma} = 1.237$ ukazují, že β_1 a β_2 jsou významně odlišné od nuly.



Průběhy



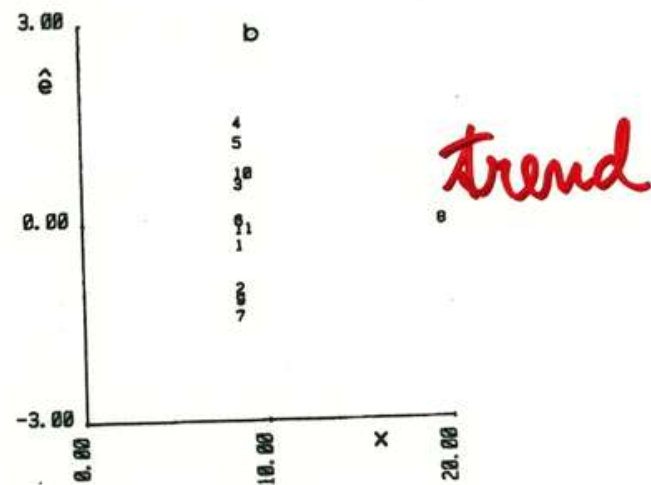
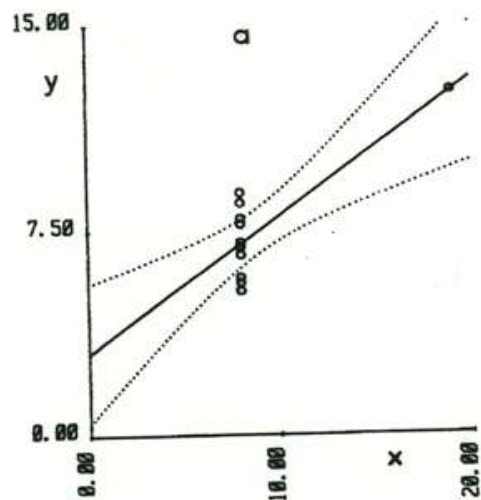
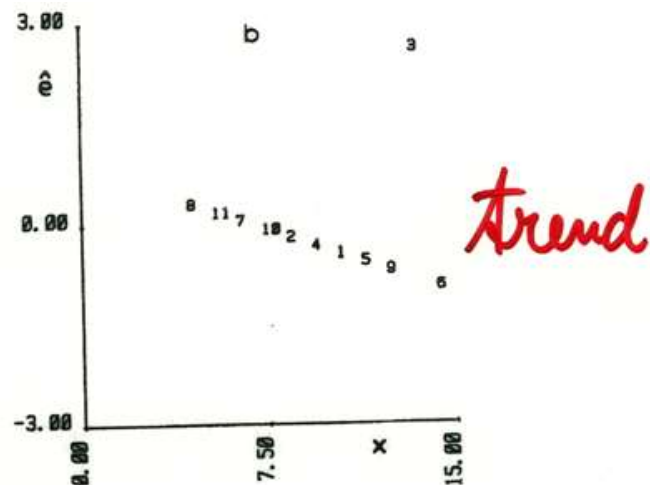
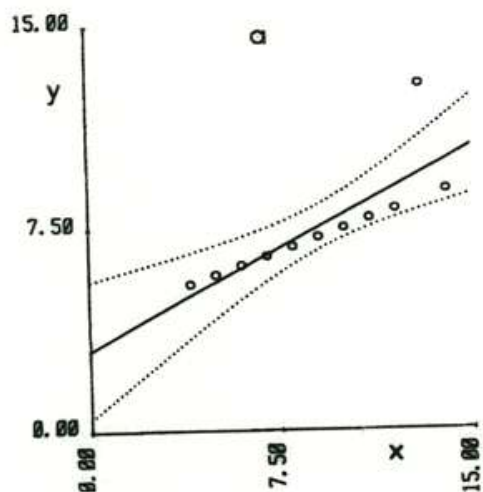
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



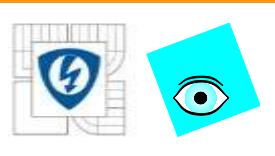


Průběhy



31.1.2011

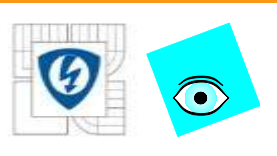
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



Pokračování příkladu

Závěr:

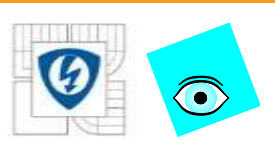
Neshodu modelu s daty indukuje grafická analýza reziduí.



Konstrukce intervalů spolehlivosti

Bodové odhady parametrů β mají pro praxi menší význam.

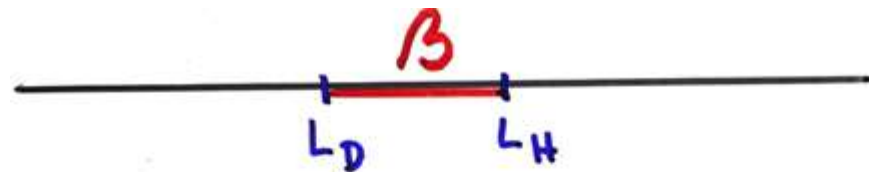
Intervaly spolehlivosti (konfidenční intervaly), ve kterých leží teoretická hodnota β se zvolenou pravděpodobností $1 - \alpha$.



Interval spolehlivosti parametrů β

Pro hladinu významnosti $\alpha = 0.10, 0.05$ nebo 0.01 odpovídají 90%ní, 95%ní nebo 99%ní intervaly spolehlivosti.

$\alpha = 0.10$ odpovídá 90%ní jistota

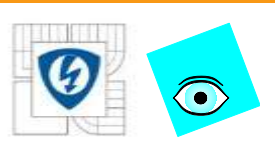


$\alpha = 0.05$ odpovídá 95%ní jistota



$\alpha = 0.01$ odpovídá 99%ní jistota





Interval spolehlivosti parametrů β

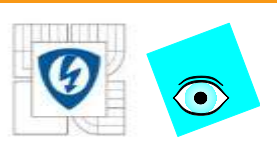
Při zanedbání korelace mezi parametry lze odvodit **100(1 - α)%ní jednoduché intervaly spolehlivosti** parametru β_j dle

$$b_j - t_{1-\frac{\alpha}{2}}(n-m)\hat{\sigma}\sqrt{c_{jj}} \leq \beta_j \leq b_j + t_{1-\frac{\alpha}{2}}(n-m)\hat{\sigma}\sqrt{c_{jj}}$$

kde c_{jj} je j -tý diagonální prvek matice $(X^T X)^{-1}$,

$t_{1-\frac{\alpha}{2}}(n-m)$ je kvantil Studentova rozdělení s $(n-m)$ stupni volnosti.

Pro korelované parametry jsou jednoduché intervaly spolehlivosti příliš úzké.



Extrémní intervaly spolehlivosti

Jsou extrémny na elipsoidech spolehlivosti dle

$$b_j - p\sqrt{c_{jj}} \leq \beta_j \leq b_j + d\sqrt{c_{jj}}$$

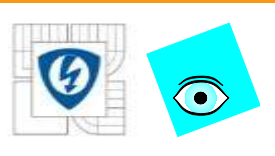
Elipsoid spolehlivosti pro q regresních parametrů:
jedná se o posledních q složek vektoru β .

100(1 - α)%ní hraniční elipsoid spolehlivosti platí

$$(b_2 - \beta_2)^T D_2^{-1} (b_2 - \beta_2) = q \hat{\sigma}^2 F_{1-\alpha}(q, n - m)$$

kde

- matice D_2 rozměru $(q \times q)$ vznikla z matice $(X^T X)^{-1}$ vynecháním prvních $(m - q)$ sloupců a $(m - q)$ řádků,
- symbol β_2 značí vektor posledních q složek vektoru β



Interval spolehlivosti predikce $\hat{y}_{P,i}$

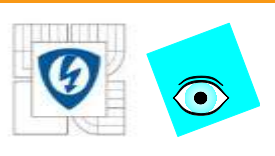
$100(1 - \alpha)\%$ ní interval spolehlivosti predikce $\hat{y}_{P,i}$ v místě x_0 :

$$b - t_{1-\frac{\alpha}{2}}(n - m)\hat{s}_{P0} \leq x_0^T \beta \leq x_0^T \beta + t_{1-\frac{\alpha}{2}}(n - m)$$

kde $\hat{s}_{P0}^2$ je rozptyl predikce

$$D(\hat{y}_{P0}) = s_{P,0}^2 = \hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0$$

$100(1 - \alpha)\%$ ní pásy spolehlivosti (resp. konfidenční pásy): jsou nejužší v těžišti nezávisle proměnných $x_{0j} = \bar{x}_j$.

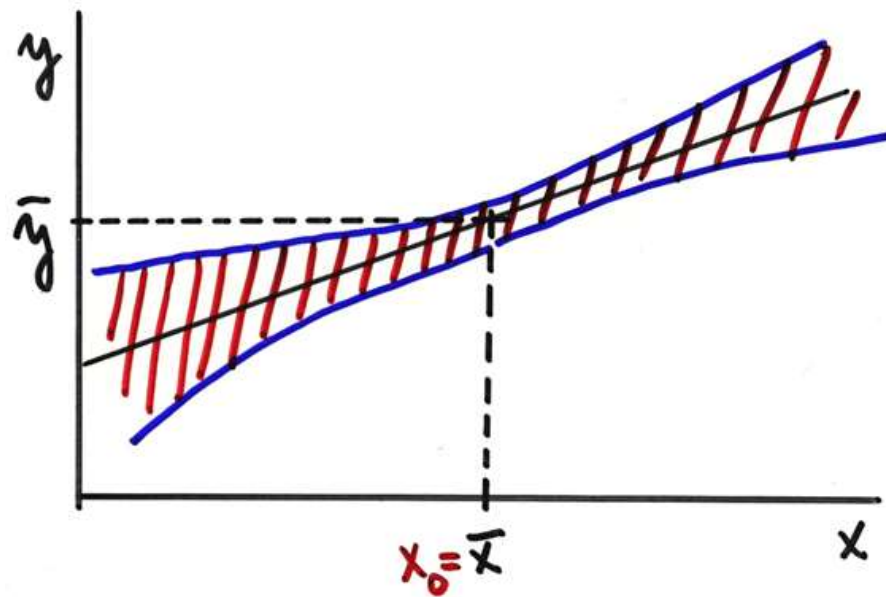


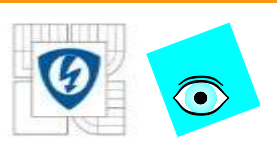
Scheffého metoda

S pravděpodobností $(1 - \alpha)$ leží teoretická, čili „pravdivá“ hodnota $x^T \beta$ v oblasti pásů

$$x^T b \pm \sqrt{m F_{1-\alpha}(m, n - m)} \hat{\sigma} \sqrt{x^T (X^T X)^{-1} x}$$

a konfidenční pásy se nazývají pásy Workinga-Hottellinga.





Interval spolehlivosti regresních parametrů

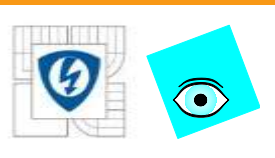
Vyjadřuje interval na číselné ose, ve kterém se s pravděpodobností $1 - \alpha$ vyskytuje neznámý parametr β základního souboru

$$\beta_j = b_j \pm t_{\frac{\alpha}{2}, n-m} \cdot s_{b_j}$$

Pravidlo: Pokud **IS** obsahuje **nulu** což znaží, že dolní hranice je záporná a horní kladná, je testovaný parametr β_j **statisticky nevýznamný**.

Směrodatné odchylky pro úsek a a směrnici b přímky:

$$s_a = \frac{s_{yx}}{\sqrt{n-2}} \cdot \sqrt{1 + \frac{\bar{x}^2}{s_x^2}} \quad s_b = \frac{s_{xy}}{s_x \sqrt{n-2}}$$



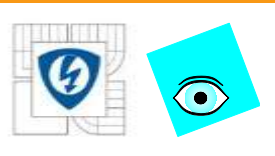
Interval spolehlivosti modelových hodnot

JEDNA HODNOTA REGRESNÍHO
MODELU (tyto hodnoty platí jen
pro jeden konkrétní výběr, ze
kterého byly vypočítány)

IS jedné
modelové
hodnoty

horní hranice IS

Plocha, ve které se s pravděpodobností
 $1 - \alpha$ nacházejí všechny možné modely
vypočítané z jakéhokoliv výběru
pocházejícího z daného základního
souboru



IS modelových hodnot

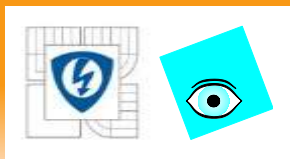
Pro model přímky:

Směrodatná odchylka reziduí

$$\mu_{y'} = y'_i \pm t_{\frac{\alpha}{2}, n-2} \cdot \frac{\sigma}{\sqrt{n-2}} \cdot \sqrt{1 + \frac{n(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Modelová hodnota

Polovina IS modelu přímky



Test významnosti regresního modelu – co testujeme?

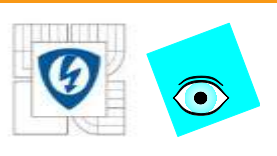
$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_mx_m$$

Testuje se významnost odhadů jednotlivých parametrů:

Test: Jestliže je testovaný odhad parametru statisticky nevýznamný, pak jeho příslušná proměnná x_j **nepřispívá** ke zpřesnění odhadu závisle proměnné y a tato proměnná x_j je v modelu zbytečná.

Testuje se model jako celek:

Test: zda příslušná kombinace všech nezávisle proměnných statisticky významně zpřesní odhad závisle proměnné y oproti použití pouhého průměru všech hodnot y .

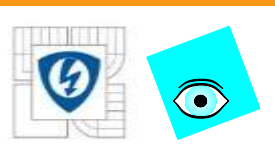


Příklad: Validizace nové analytické metody

Proved'te validizaci nové analytické metody porovmáním jejich výsledků y vůči standardům x .

- (a) určete odhady b_1 a b_2 ,
- (b) zkonstruujte 95%ní interval spolehlivosti úseku a směrnice,
- (c) 95%ní elipsoid spolehlivosti,
- (d) 95%ní interval spolehlivosti predikce v těžišti.

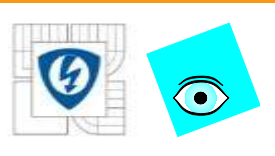
Data: $n = 24, m = 2$



Pokračování příkladu

Obsah látky určený standardní metodou (x) a novom metodou (y) v g

x	40.2	43.8	47.6	50.7	56.8	81.3	83.3	97.1	102.5	118.7
y	48.9	39.1	42.6	56.9	70.3	71.5	97.6	99.9	105.2	102.3
x	129.4	184.8	287.5	295.4	420.3	421.3	427.9	566.1	608.5	640.7
y	106.8	162.9	234.0	303.4	388.8	391.1	369.3	611.6	580.2	643.3
x	692.8				705.2			714.4		881.4
y	596.6				612.6			633.5		669.8



Pokračování příkladu

(a) odhady b_1 a b_2 :

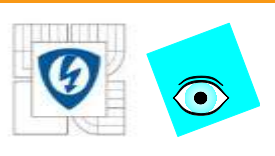
- odhad úseku $b_2 = 14.73 (\pm 12.61)$,
- odhad směrnice $b_1 = 0.868 (\pm 0.0302)$,
- koeficient determinace $\hat{R}^2 = 0.974$,
- odhad směrodatné odchylky reziduí $\hat{\sigma} = 39.54$.

(b) interval spolehlivosti úseku $H_0: \beta_2 = 0$ vs. $H_A: \beta_2 \neq 0$

$$b_2 - t_{1-\frac{\alpha}{2}}(22)\sqrt{D(b_2)} \leq \beta_2 \leq b_2 + t_{1-\frac{\alpha}{2}}(22)\sqrt{D(b_2)}$$

a dosazením

$$14.73 - 2.08 \cdot 12.61 \leq \beta_2 \leq 14.73 + 2.08 \cdot 12.61$$



Pokračování příkladu

vyjde

$$-11.499 \leq \beta_2 \leq 40.959$$

Závěr testování: interval spolehlivosti úseku zahrnuje nulu, takže H_0 je přijata a úsek β_2 lze považovat za nulový.

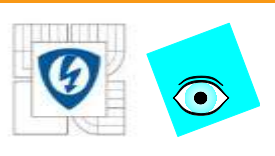
Interval spolehlivosti směrnice $H_0: \beta_1 = 1$ vs.

$$H_A: \beta_1 \neq 1$$

$$0.868 - 2.08 \cdot 0.0302 \leq \beta_1 \leq 0.868 + 2.08 \cdot 0.0302$$

a vyčíslením

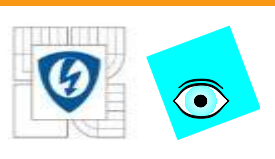
$$0.805 \leq \beta_1 \leq 0.93$$



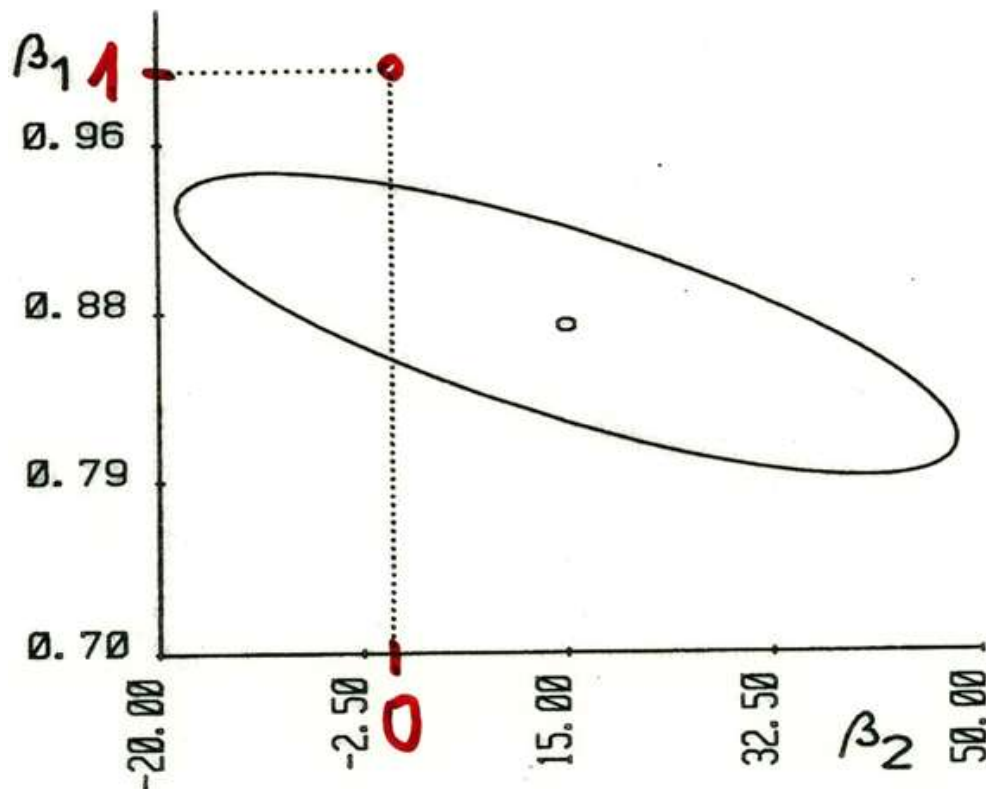
Závěr testování:

Interval spolehlivosti neobsahuje jedničku, takže H_0 je zamítnuta a směrnici β_1 nelze považovat za jednotkovou.

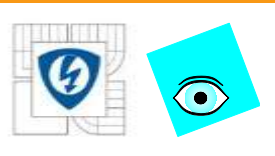
(c) Konstrukce 95%ní elipsy spolehlivosti a bod $\beta_1 = 1$ a $\beta_2 = 0$: $H_0: \beta_2 = 0$ a $\beta_1 = 1$ vs. $H_A: \beta_2 \neq 0$ a $\beta_1 \neq 1$



Pokračování příkladu



Závěr testování: 95%ní elipsa spolehlivosti neobsahuje bod $\{\beta_1 = 1; \beta_2 = 0\}$ a proto H_0 je zamítnuta.



Pokračování příkladu

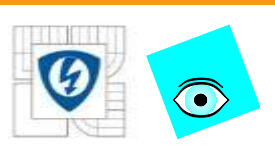
(d) Interval spolehlivosti predikce v těžišti:

Pro případ regresní přímky a $x_i = \bar{x} = 320.7$ je psát

$$D(\hat{y}_P) = \hat{s}_{P,0}^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{n(\bar{x} - \bar{x})}{D} \right) = \frac{\hat{\sigma}^2}{n}$$

Dosazením

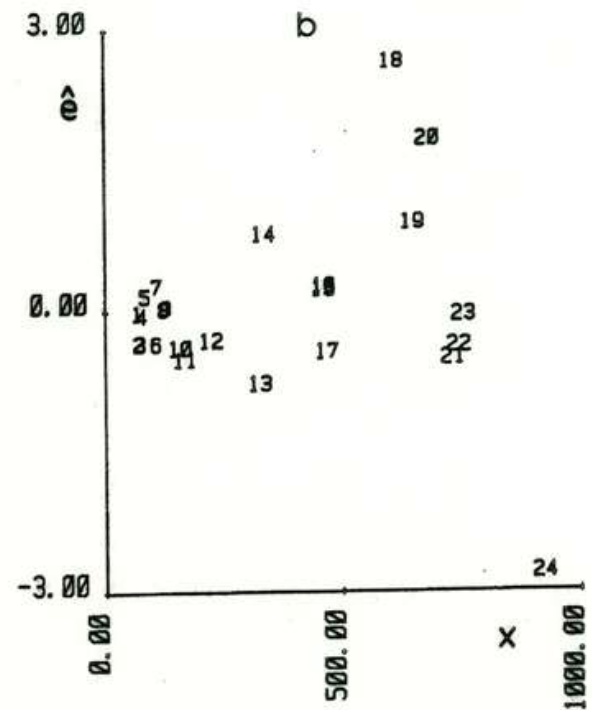
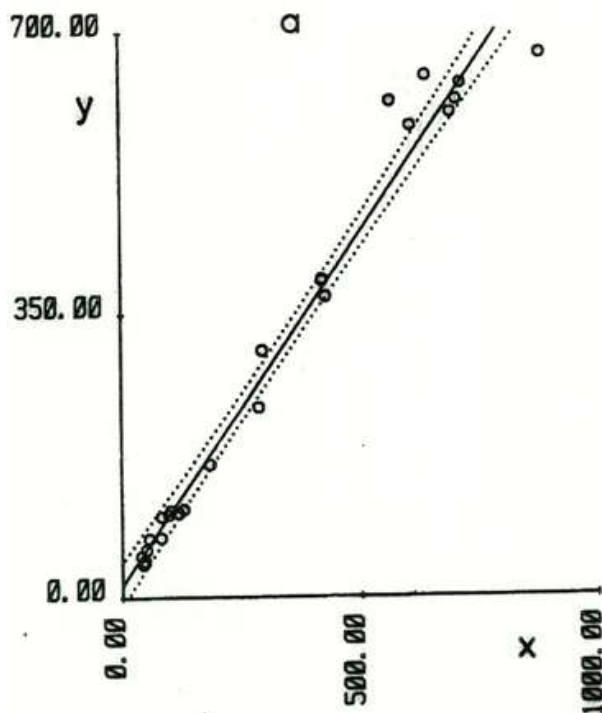
$$\begin{aligned} 14.73 + 0.868 \cdot 320.7 - \frac{2.08 \cdot 39.54}{\sqrt{24}} &\leq \eta_{\bar{x}} \leq \\ &\leq 14.73 - 0.868 \cdot 320.7 + \frac{2.08 \cdot 39.54}{\sqrt{24}} \end{aligned}$$

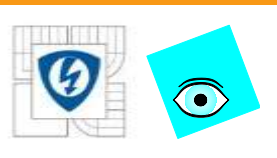


Pokračování příkladu

Tedy $276.89 \leq \eta_{\bar{x}} \leq 309.89$

Závěr testování: intervaly spolehlivosti predikce v těžišti
vzchází poměrně široký.





Závěr:

Intervaly spolehlivosti indikují, že úsek regresní přímky lze považovat za nulový $\beta_2 = 0$, zatímco směrnice β_1 je výrazně odlišná od jedničky.

Nová analytická metoda vede k odlišným výsledkům od standardní.



Kapitola 6.2

PODSTATNÉ TESTY VÝZNAMNOSTI V KORELAČNÍ A REGRESNÍ ANALÝZE

31.1.2011

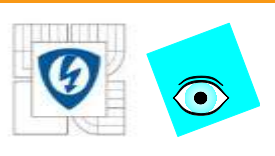
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Podstatné testy významnosti v korelační a regresní analýze

- test významnosti korelačního koeficientu
 - test významnosti modelu jako celku
 - test významnosti jednotlivých regresních parametrů
 - test shody lineárních regresních modelů
- a mnoho dalších testů



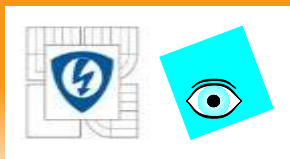
Test významnosti R

Test významnosti odpovídá, zda je korelace R mezi výběrovými proměnnými natolik silná, abychom ji mohli považovat za prokázanou i pro základní soubor ρ .

Pro párový R : $t_R = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ $t_{\alpha, n-2}$ n je počet hodnot výběru

Pro násobný R : $F_R = \frac{R^2(n-m)}{(1-R^2)(m-1)}$ $t_{\alpha, n-m}$ m je počet proměnných

Pro parciální R : $t_R = \frac{R\sqrt{n-k-2}}{\sqrt{1-R^2}}$ $t_{\alpha, n-k-2}$ k je počet vyloučených proměnných



Test významnosti regresního modelu – co testujeme?

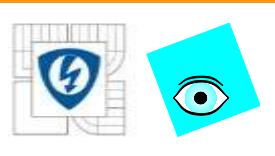
$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_mx_m$$

Testuje se významnost odhadů jednotlivých parametrů:

Test: Jestliže je testovaný odhad parametru statisticky nevýznamný, pak jeho příslušná proměnná x_j **nepřispívá** ke zpřesnění odhadu závisle proměnné y a tato proměnná x_j je v modelu zbytečná.

Testuje se model jako celek:

Test: zda příslušná kombinace všech nezávisle proměnných statisticky významně zpřesní odhad závisle proměnné y oproti použití pouhého průměru všech hodnot y .

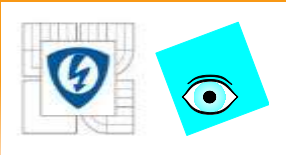


Test významnosti regresních parametrů

$H_0: \beta_j = 0$, tj. j -tý regresní parametr je nevýznamný

$$t = \frac{b_j - \beta_j}{s_b} \quad \text{pro } \beta_j = 0 \quad \longrightarrow \quad t = \frac{b_j}{s_b}$$

Pokud platí, že $|t| > t_{\alpha_2; n-m}$, potom je j -tý regresní parametr statisticky významný a příslušná proměnná musí zůstat v modelu.



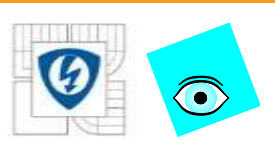
Test shody regresních modelů

Shoda dvou empirických modelů

$H_0: \beta_{j,1} = \beta_{j,2}$, tj. regresní koeficienty obou modelů jsou v základním souboru shodné.

Vycházíme z testování shody regresních parametrů dvou lineárních modelů $y_1 = X_1\beta_1 + \varepsilon_1$ a $y_2 = X_2\beta_2 + \varepsilon_2$.

Při tomto testu využijeme tzv. **složeného modelu**, tj. oba porovnávané výběry sloučíme do jednoho a také pro něj stanovíme parametry stejného modelu jako pro oba dílčí výběry



Testy složených hypotéz

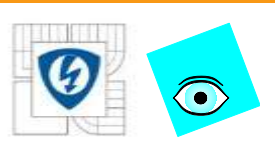
$H_0: \beta_2 = 0$, proti $H_0: \beta_2 \neq 0$

kde β_2 je posledních q prvků vektoru β , užijeme regresní model v děleném tvaru.

$$y = [X_1 \quad X_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

kde X_1 je $n \times (m - q)$, obsahující vysvětlující proměnné, jejichž regresní koeficienty nejsou v testovaném vektoru β_2 .

X_2 je $n \times q$, obsahující vysvětlující proměnné, jejichž regresní koeficienty jsou v testovaném vektoru β_2 .



Testy složených hypotéz

a) Pokud platí H_0 , je $\hat{y}_{P,1} = X_1 b_1$, kde

$$b_1 = (X_1^T X_1)^{-1} X_1^T y$$

a odpovídající reziduální součet čtverců je

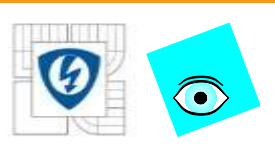
$$RSC_1 = (y - \hat{y}_{P,1})^T (y - \hat{y}_{P,1})$$

b) Pokud platí H_A , je $\hat{u}_P = Xb$, kde

$$b = (X^T X)^{-1} X^T y$$

a odpovídající reziduální součet čtverců je

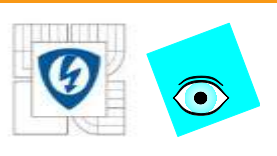
$$RSC = (y - \hat{y}_P)^T (y - \hat{y}_P)$$



Testy složených hypotéz

Rozdíl $(RSC_1 - RSC)$ odpovídá zvýšení reziduálního součtu a testuje se

$$F_1 = \frac{(RSC_1 - RSC)(n - m)}{RSC \cdot q} \quad \text{vs.} \quad F_{1-\alpha}(q, n - m)$$



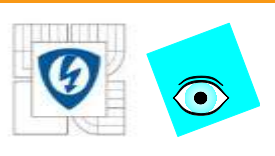
Příklad – Simultánní test složené hypotézy u LB zákona

U přímky LB zákona testujte $H_0: \beta_2 = 0, \beta_1 = 0.148$ proti $H_A: \beta_2 \neq 0, \beta_1 \neq 0.148$. Nesprávný přístup spočívá v odděleném testování dvou nulových hypotéz $H_0: \beta_2 = 0, H_0: \beta_1 = 0.148$.

Řešení: Dosazením

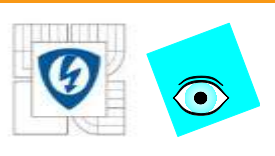
$$T_2 = \frac{|1.46 \cdot 10^{-4} - 0|}{0.00398} = 0.037$$

$$T_1 = \frac{|0.1459 - 0.148|}{0.000908} = 2.214$$



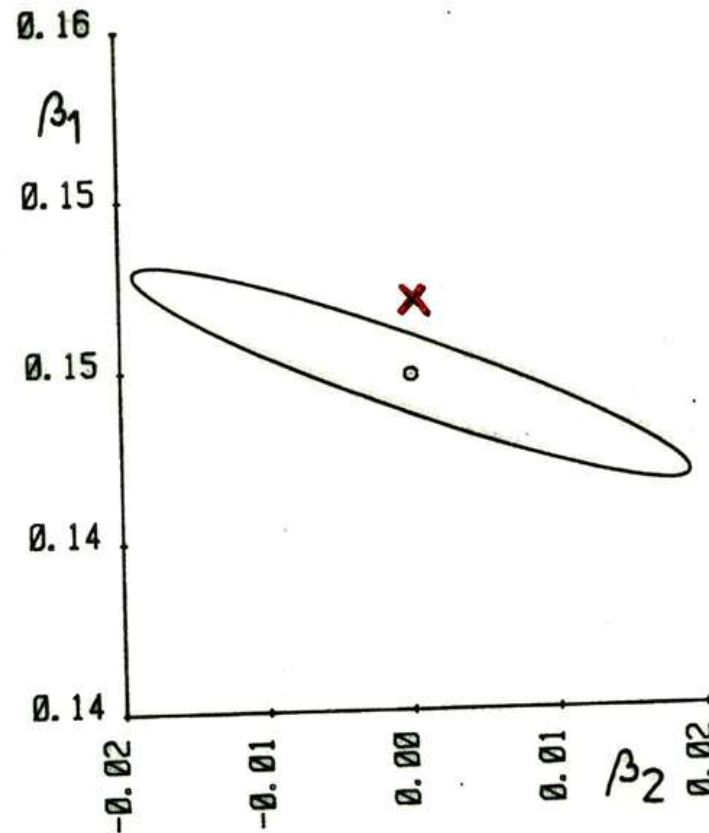
Pokračování příkladu

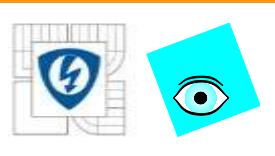
1. Obě kritéria jsou menší než $t_{0.975}(4) = 2.7764$ a testování vede k nesprávnému závěru přijmutí hypotézy $H_0: \beta_2 = 0, \beta_1 = 0.148$.
2. Správnější přístup spočívá v simultánním testování složené hypotézy $H_0: \beta_2 = 0, \beta_1 = 0.148$.



Konstrukce 95%ní oblasti spolehlivosti

Křížkem je vyznačen bod $\beta_2 = 0, \beta_1 = 0.148$





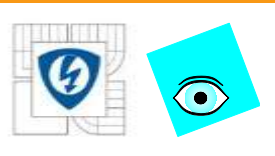
Postup

- a) $RSC = 5.12 \cdot 10^{-5}$ s využitím skutečných odhadů
 $b_1 = 0.1459$ a $b_2 = 1.461 \cdot 10^{-1}$
- b) $RSC_1 = 5.3476 \cdot 10^{-4}$ s využitím předpokládaných
hodnot parametrů $\beta_2 = 0, \beta_1 = 0.148$.

Dosazením do testačního kritéria bude

$$F_1 = \frac{(5.347 \cdot 10^{-4} - 5.12 \cdot 10^{-5}) \times 4}{5.12 \cdot 10^{-5} \times 2} = 18.89$$

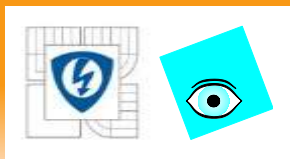
Jelikož $F_1 > F_{0.95}(2,4) = 6.944$, je $H_0: \beta_2 = 0, \beta_1 = 0.148$ zamítnuta.



95%ní konfidenční elipsoid parametrů β_1 a β_2 ukazuje, že bod $\beta_{1,0} = 0.148, \beta_{2,0} = 0$ leží mimo 95%ní oblast spolehlivosti.

Závěr:

Zásadně **nelze nahradit** simultánní testování složené hypotézy dílčími testy oddělených hypotéz.



Příklad: Validizace simultánním testem složené hypotézy

Testujte složenou hypotézu $H_0: \beta_2 = 0, \beta_1 = 1$ u validizace nové analytické metody proti $H_A: \beta_2 \neq 0, \beta_1 \neq 1$.

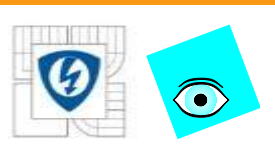
Řešení:

$RSC = 3440$ a dosazením $\beta_{1,0} = 1$ a $\beta_{2,0} = 0$ vyjde $RSC_1 = 8221$.

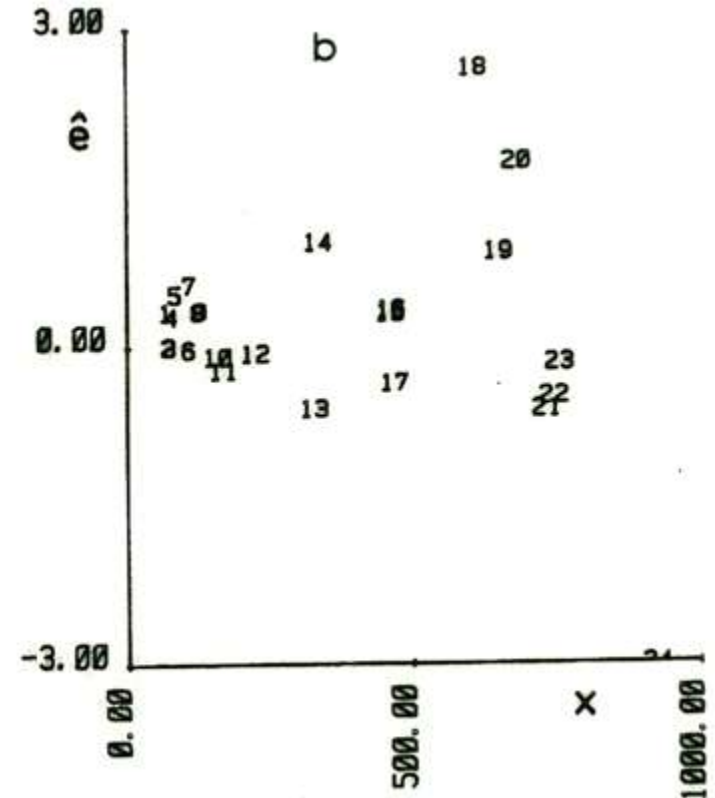
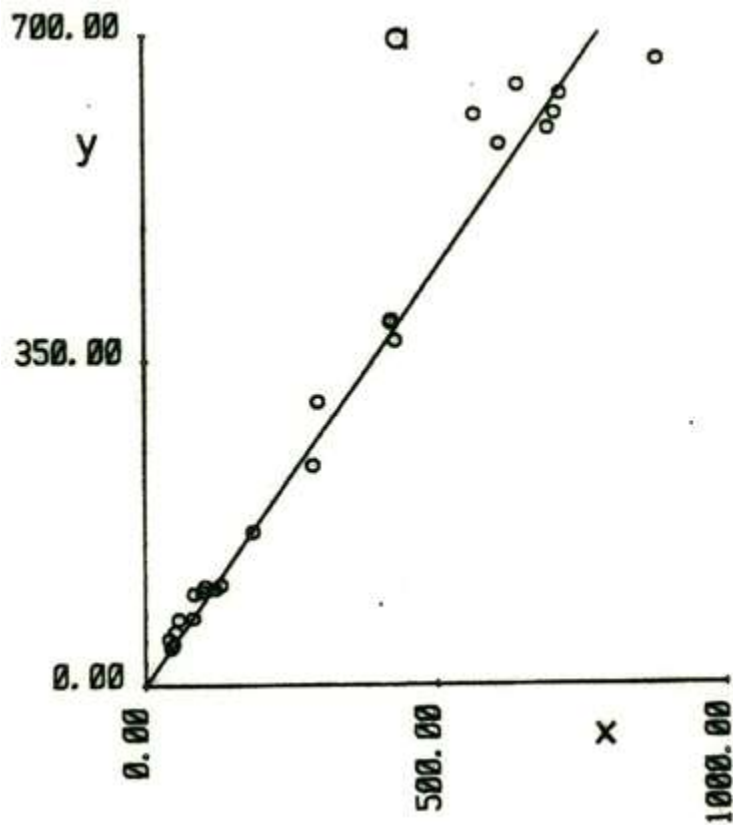
Testační kritérium

$$F_1 = \frac{(8220 - 3440) \times 22}{3440 \times 2} = 15.28$$

je větší než $F_{0.95}(2.22) = 3.44$, takže H_0 je zamítnuta.



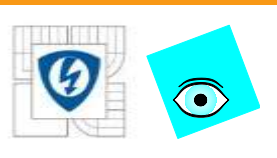
Pokračování příkladu



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

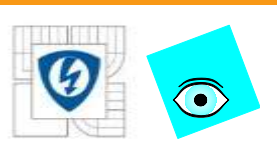




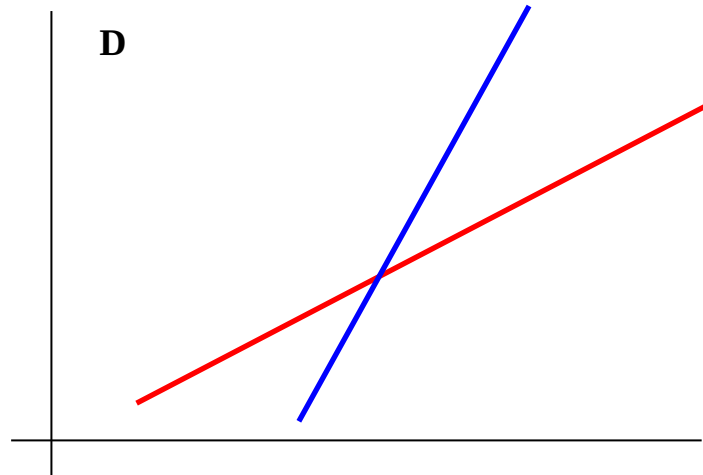
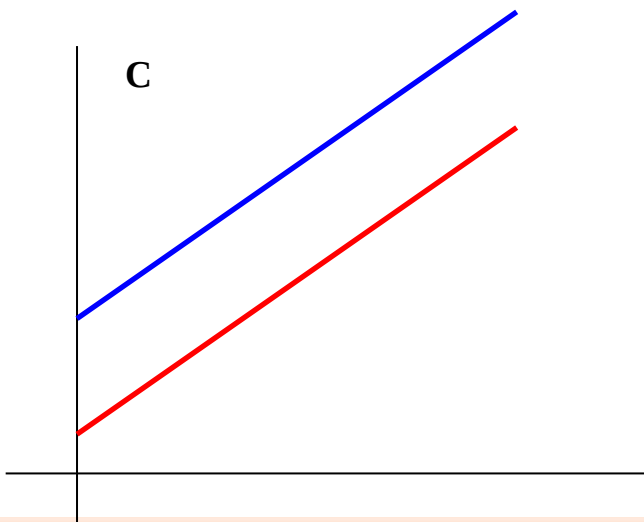
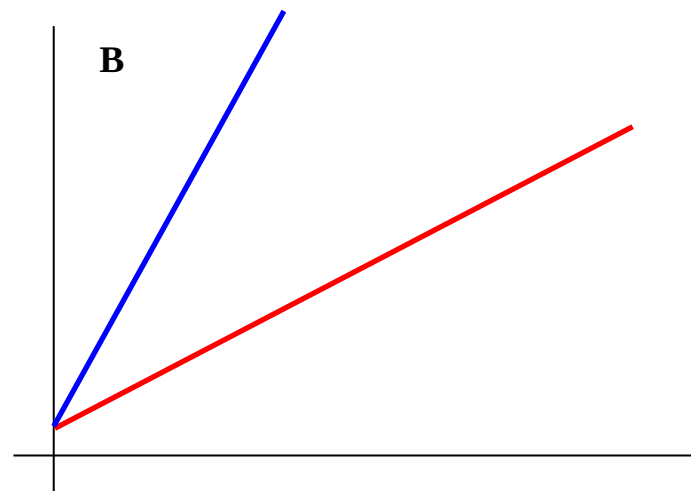
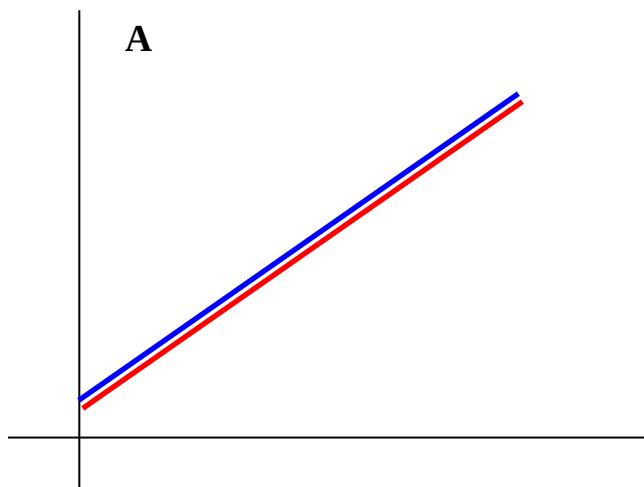
Pokračování příkladu

Závěr:

I simultánní test složené hypotézy $H_0: \beta_2 = 0, \beta_1 = 1$ potvrdil, že nová analytická metoda není shodná se standardní metodou.



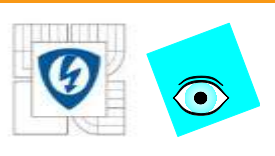
Porovnání regresních přímek



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





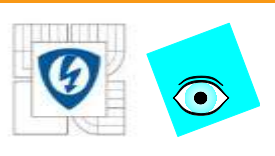
Porovnání regresních přímk

Porovnáním M navržených regresních modelů

$$y_{ij} = \beta_{2j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \quad j = 1, \dots, M; i = 1, \dots, n_j$$

Postup:

1. krok: vyčíslení $\beta_{2j}, \beta_{1j}, \hat{\sigma}_j^2, \quad j = 1, \dots, M$
2. krok: test homskedasticity $\hat{\sigma}_j^2 = \hat{\sigma}^2, j = 1, \dots, M$
3. krok: testování, zda-li
 - a) regresní přímky mají společný průsečík,
 - b) regresní přímky mají společnou směrnici,
 - c) regresní přímky jsou totožné.

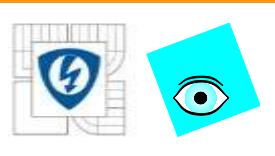


Test homoskedasticity (Bartlettův test)

$H_0: \sigma_j^2 = \sigma^2, j = 1, \dots, M$, se stupni volnosti $\nu_j = (n_j - m)$, pro přímku $m = 2$

Celkový počet stupňů volnosti $V = \sum_{j=1}^M \nu_j$

$$L = 1 + \frac{\sum_{j=1}^M \nu_j^{-1} - V^{-1}}{3M - 3}$$



Sdružený odhad rozptylu

$\hat{\sigma}_c^2 = \sum_{j=1}^M \frac{v_j \hat{\sigma}_j^2}{V}$ je vážená kombinace odhadů rozptylu jednotlivých modelů.

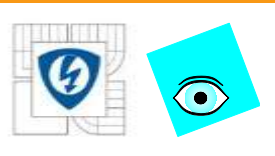
Testační kritérium Bartlettova testu

$$B = \frac{V \ln \hat{\sigma}_c^2 - \sum_{j=1}^M v_j \ln \hat{\sigma}_j^2}{L}$$

Test: Je-li $B < \chi_{1-\alpha}^2(M-1)$, je H_0 přijmata (homoskedasticita).

K porovnání dvou přímek (skupin bodů), $M = 2$,

stačí testační kritérium $F_2 = \frac{\max(\hat{\sigma}_1^2, \hat{\sigma}_2^2)}{\min(\hat{\sigma}_1^2, \hat{\sigma}_2^2)}$



Test homogeneity úseků

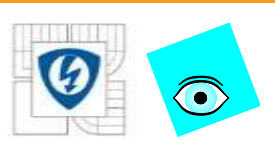
$$H_0: \beta_{21} = \beta_{22} = \dots = \beta_{2j} = \dots = \beta_{2M} = \beta_{2c},$$

Sdružený odhad úseku β_{2c} je vážená kombinace odhadů jednotlivých úseků

$$b_{2c} = \frac{\sum_{j=1}^M w_{Bj} b_{2j}}{\sum_{j=1}^M w_{Bj}}$$

kde j -tý váhový koeficient w_{Bj} úseku j -té přímky

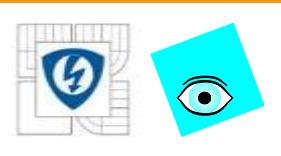
$$w_{Bj} = \frac{n_j \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{n_j} x_{ij}^2}$$



Testační statistika F-testu

$$F_I = \frac{\frac{1}{M-1} \sum_{j=1}^M w_{B,j} (b_{2j} - b_{2c})^2}{\frac{1}{n-2M} \sum_{j=1}^M \sum_{i=1}^{n_j} \hat{e}_{ij}^2}$$

kde $n = \sum_{j=1}^M n_j$



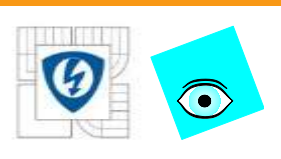
Rezidua $\hat{e}_{ij}$ jsou určována z jednotlivých průmek a platí

$$\sum_{j=1}^M \sum_{i=1}^{n_j} \hat{e}_{ij}^2 = \sum_{j=1}^M RSC_j ,$$

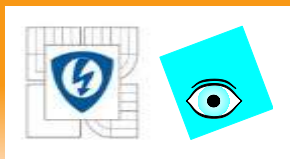
kde RSC_j je reziduální součet čtverců v j-té skupině.

Test: Je-li $F_I < F_{1-\alpha}(M - 1, n - 2M)$, je H_0 přijata a všechny přímky mají stejný úsek.

Nejlepším odhadem společného úseku je b_{2c} a jeho rozptyl je



$$D(b_{2c}) = \frac{\hat{\sigma}^2}{\sum_{j=1}^M w_{Bj}} = \frac{\frac{1}{n - 2M} \sum_{j=1}^M \sum_{i=1}^{n_j} \hat{e}_{ij}^2}{\sum_{j=1}^M w_{Bj}}$$



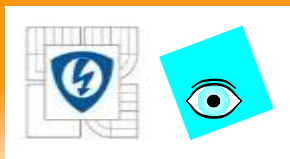
Test homogeneity směrníc (test rovnoběžnosti regresních přímk)

$$H_0: \beta_{11} = \beta_{12} = \dots = \beta_{1j} = \dots = \beta_{1M} = \beta_{1c},$$

Sdružený odhad celkové směrnice β_{1c} je vážená kombinace jednotlivých směrníc β_{1j} dle

$$b_{1c} = \frac{\sum_{j=1}^M w_{Sj} b_{1j}}{\sum_{j=1}^M w_{Sj}}$$

$$\text{kde } w_{Sj} = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$



Test homogeneity směrnic (test rovnoběžnosti regresních přímek)

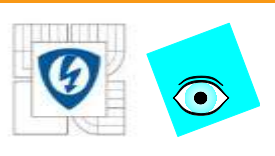
Testační statistika F-testu

$$F_S = \frac{\frac{1}{M-1} \sum_{j=1}^M w_{S,j} (b_{1j} - b_{1c})^2}{\frac{1}{n-2M} \sum_{j=1}^M \sum_{i=1}^{n_j} \hat{e}_{ij}^2}$$

Test: Je-li $F_S < F_{1-\alpha}(M-1, n-2M)$, H_0 je přijata a regresní přímky jsou rovnoběžné.

Nejlepším odhadem celkové směrnice je b_{1c} a její

rozptyl je $D(b_{1c}) = \frac{\frac{1}{n-2M} \sum_{j=1}^M \sum_{i=1}^{n_j} \hat{e}_{ij}^2}{\sum_{j=1}^M w_{Sj}}$



Test shody regresních přímek

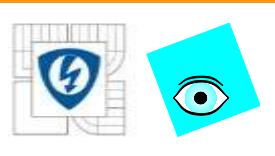
$$H_0: \beta_{2j} = \beta_{2c}, \beta_{1j} = \beta_{1c}, j = 1, \dots, M,$$

je kombinací předchozích testů

$$\text{Testační statistika } F_A = \frac{\frac{RSC_K - RSC_C}{2M - 2}}{\frac{RSC_C}{n - 2M}}$$

$$\text{kde } RSC_C = \sum_{j=1}^M RSC_j$$

Test: Je-li $F_A < F_{1-\alpha}(2M - 2, n - 2M)$, je H_0 přijata a všechny regresní přímky jsou totožné se společným odhadem úseku a b_{2K} směrnice b_{1K} .



Test shody dvou lineárních modelů

test shody parametrů β_1 a β_2 dvou lineárních modelů

$$y_1 = X_1\beta_1 + \varepsilon_1 \quad RSC_1$$

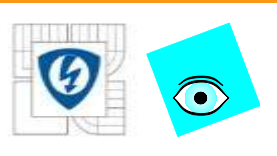
$$y_2 = X_2\beta_2 + \varepsilon_2 \quad RSC_2$$

kde X_1 rozměru $(n_1 \times m)$, y_1 rozměru $(n_1 \times 1)$,

kde X_2 rozměru $(n_2 \times m)$, y_2 rozměru $(n_2 \times 1)$.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

RSC ...



Chowův test shody dvou lineárních modelů

$$H_0: \beta_1 = \beta_2 \text{ vs. } H_A: \beta_1 \neq \beta_2$$

Testační kritérium

$$F_c = \frac{(RSC - RSC_1 - RSC_2)(n - 2m)}{(RSC_1 + RSC_2)(m)}$$

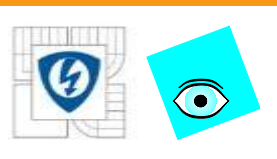
kde $n = n_1 + n_2$.

Testování:

(a) při homoskedasticitě $\sigma_1^2 = \sigma_2^2: F_c \text{ vs. } F(m, n - 2m)$,

(b) při heteroskedasticitě $\sigma_1^2 \neq \sigma_2^2: F_c \text{ vs. } F(m, r)$,

$$\text{kde } r = \frac{[(n_1 - m)\sigma_1^2 + (n_2 - m)\sigma_2^2]^2}{(n_1 - m)\sigma_1^4 + (n_2 - m)\sigma_2^4}$$

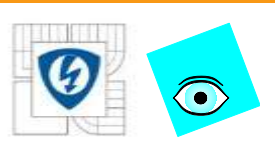


Příklad 6.14 Porovnání výsledků měření ze dvou laboratoří

Stanovení volné enthalpie $-\Delta G_i$ par oxidu boritého v závislosti na teplotě $T[\text{K}]$ bylo provedeno paralelně ve dvou laboratořích. Je možné hodnoty naměřené v obou laboratořích považovat za shodné?

Data: $n = 6$, $m = 2$

T [K]	$-\Delta G [\text{kcal mol}^{-1}]$	
	Laboratoř A	Laboratoř B
1409	34.9	34.9
1441	34.6	33.8
1457	31.9	33.4
1492	33.1	32.4
1569	30.1	30.3
1610	29.3	29.1



Pokračování příkladu

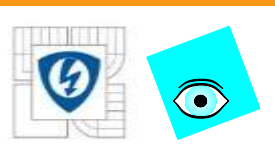
Řešení:

Lineární regresní model pro obě skupiny dat

$$E \left(-\frac{\Delta G}{T} \right) = \beta_{1,A} T + \beta_{2,A}$$

$$E \left(-\frac{\Delta G}{T} \right) = \beta_{1,B} T + \beta_{2,B}$$

Chowův test $H_0: \beta_A = \beta_B$ vs. $H_A: \beta_A \neq \beta_B$ kde $\beta_A = (\beta_{1,A}, \beta_{2,A})^T$ a $\beta_B = (\beta_{1,B}, \beta_{2,B})^T$.

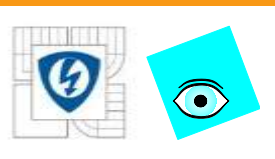


Pokračování příkladu

Laboratoř A: $b_{1,A} = -2.768 \cdot 10^{-2} (\pm 5.25 \cdot 10^{-3})$
 $b_{2,A} = 73.73 (\pm 7.865), \hat{\sigma} = 0.916$
 $RSC_A = 3.358$

Laboratoř B: $b_{1,B} = -2.776 \cdot 10^{-2} (\pm 1.57 \cdot 10^{-4})$
 $b_{2,B} = 73.82 (\pm 0.235), \hat{\sigma} = 0.0274$
 $RSC_B = 2.992 \cdot 10^{-3}$

Laboratoř A + B: (označení C = A + B)
 $b_{1,C} = -2.772 \cdot 10^{-2} (\pm 2.35 \cdot 10^{-3})$
 $b_{2,C} = 73.77 (\pm 3.521), \hat{\sigma} = 0.58$
 $RSC_C = 3.364$



Pokračování příkladu

Dosazením za $RSC = RSC_c$, $RSC_1 = RSC_A$ a $RSC_2 = RSC_B$ do testačního kritéria Chowova testu

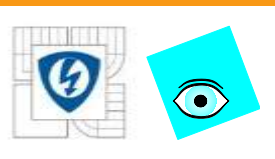
$$F_c = \frac{(RSC - RSC_1 - RSC_2)(n - 2m)}{(RSC_1 + RSC_2)(m)}$$

bude

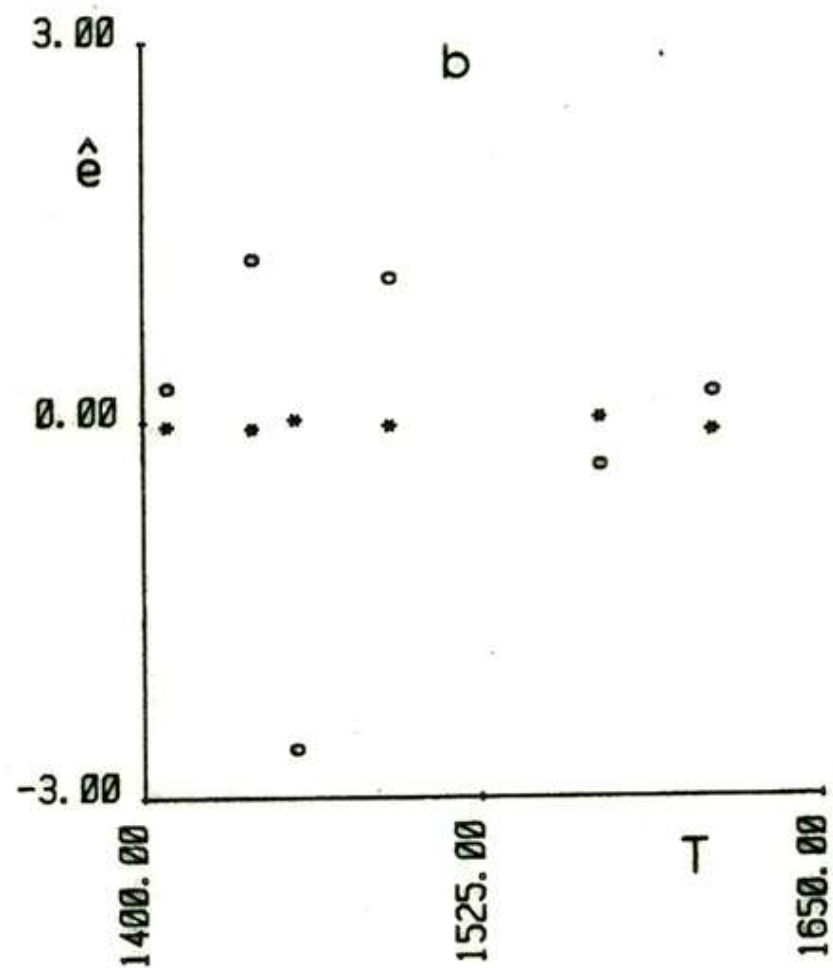
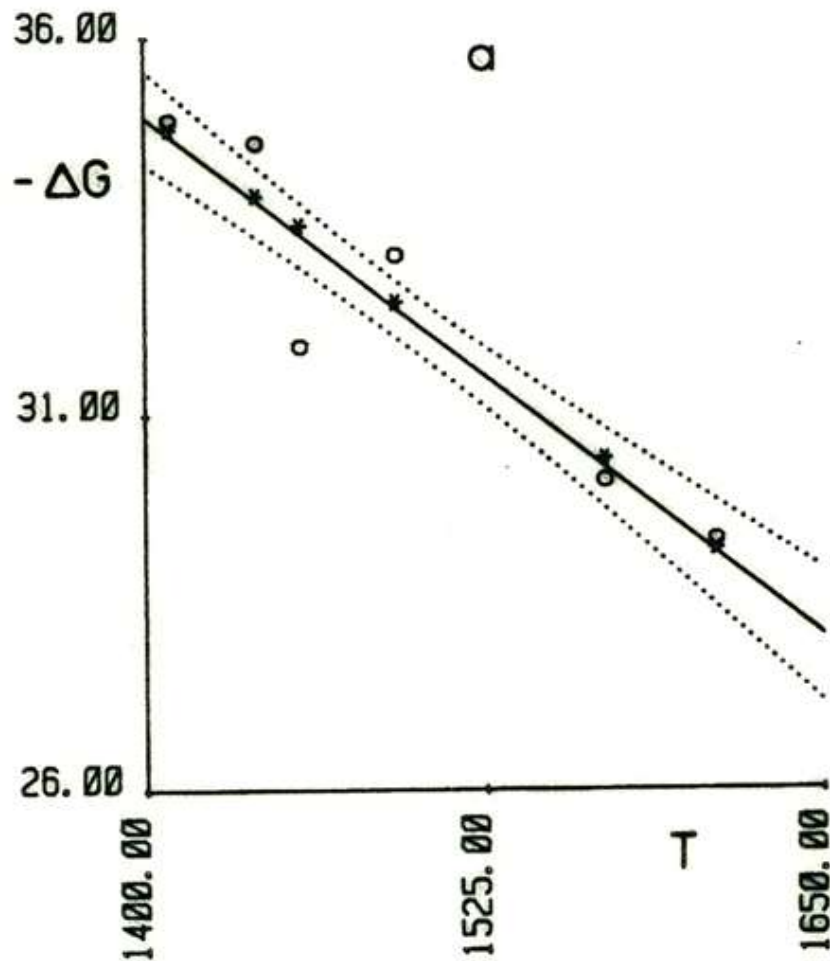
$$F_c = \frac{(3.364 - 3.358 - 0.002992)(12 - 4)}{(3.358 + 0.002992)(2)} = 0.0036$$

Data A se od B liší v rozptylu, budou stupně volnosti dle

$$r = \frac{[4 \times 0.916^2 + 4 \times 0.0274^2]^2}{4 \times 0.916^4 + 4 \times 0.0274^4} = 4.007 \approx 4$$



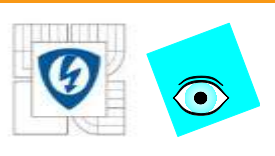
Pokračování příkladu



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





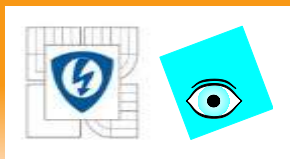
Pokračování příkladu

Test:

Jelikož $F_{0.95}(2,4) = 6.94 > F_c$, je $H_0: \beta_A = \beta_B$ přijata.

Závěr:

Chowovým testem je prokázáno, že výsledky v obou laboratořích A a B jsou shodné.



Příklad 6.15 Porovnání dvou kalibračních přímek

Dva preparáty insulínu A a B byly vyšetřovány s ohledem na snížení úrovně krevního cukru. 11 krysám byl naočkován insulín A a 9 krysám pak insulín B. Porovnejte, zda účinek obou insulínů je shodný.

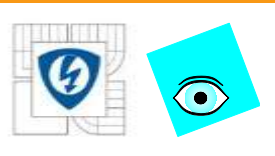
Data: množství insulínu x [μl] a snížení hladiny cukru y [%]

Insulin A: $n_A = 11$

x	120	160	200	240	280	320	360	400	440	480	500
y	17	26	30	27	45	47	48	63	60	69	67

Insulin B: $n_B = 9$

x	160	200	240	280	320	360	400	440	480
y	9	18	17	25	39	45	47	57	61



Pokračování příkladu

Řešení:

Pro insulin A platí

$$\hat{y}_A = 1.808 (\pm 3.504) + 0.1369 (\pm 0.0103)x$$

$$RSC_A = 159.6 \text{ a } \hat{\sigma} = 4.211$$

Pro insulin B platí

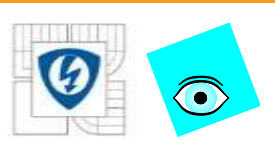
$$\hat{y}_A = -18.67 (\pm 3.535) + 0.1688 (\pm 0.0105)x$$

$$RSC_A = 74.25 \text{ a } \hat{\sigma} = 3.26$$

Pro sloučené obě skupiny $C = A + B$ platí

$$\hat{y}_c = -6.937 (\pm 4.45) + 0.1481 (\pm 0.0132)x$$

$$RSC_c = 821.1$$

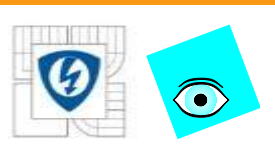


Pokračování příkladu

Dosazením do testačního kritéria

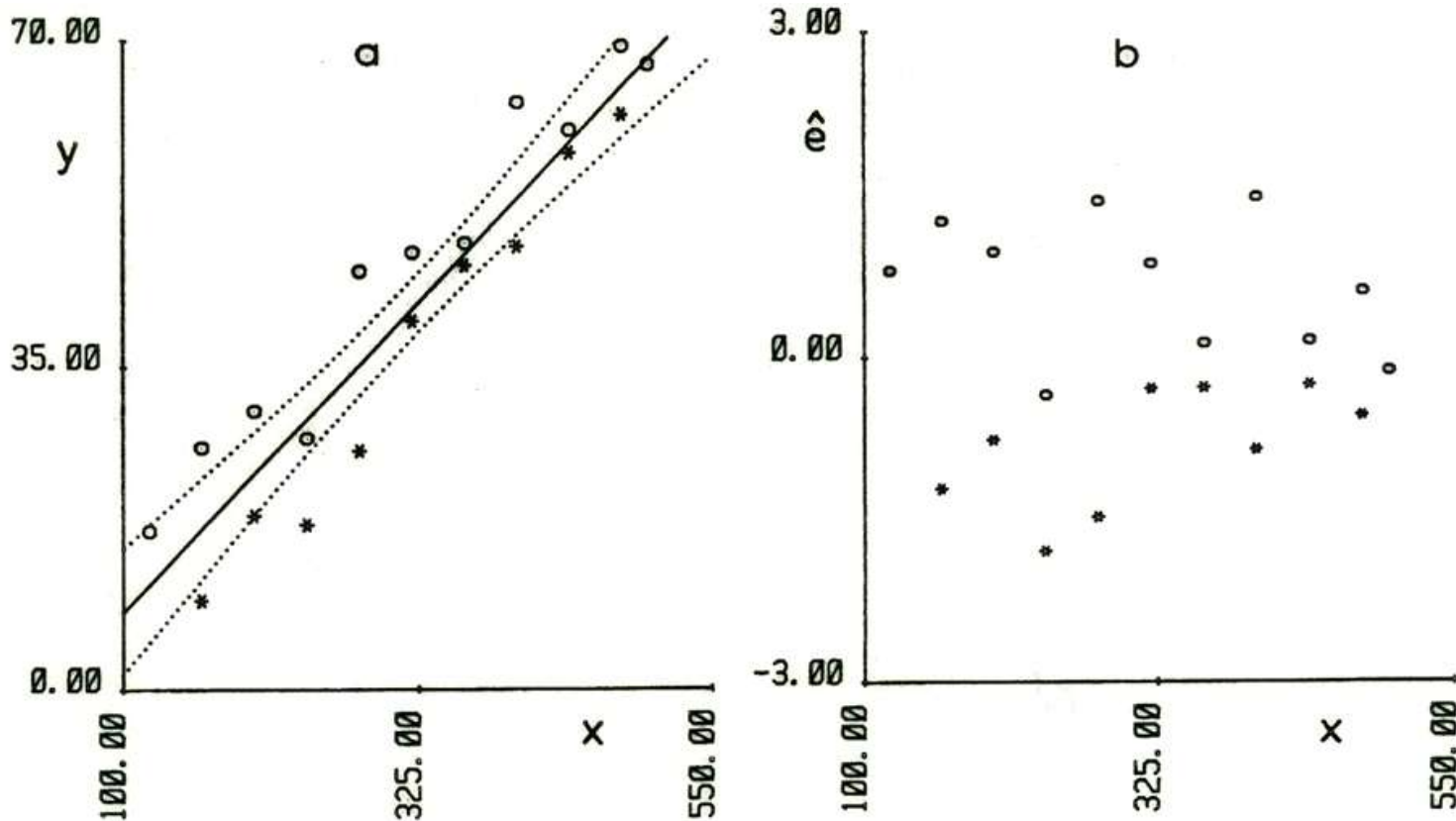
$$F_c = \frac{(821.1 - 159.6 - 74.25)(20 - 4)}{(159.6 + 74.25)^2} = 20.09$$

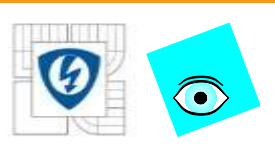
Test: Jelikož $F_{0.95}(2,16) = 3.63 < F_c$, je H_0 zamítnuta a působení obou insulinů je rozličné.



Pokračování příkladu

Závěr: Oba vyšetřované typy insulínu působí odlišně.

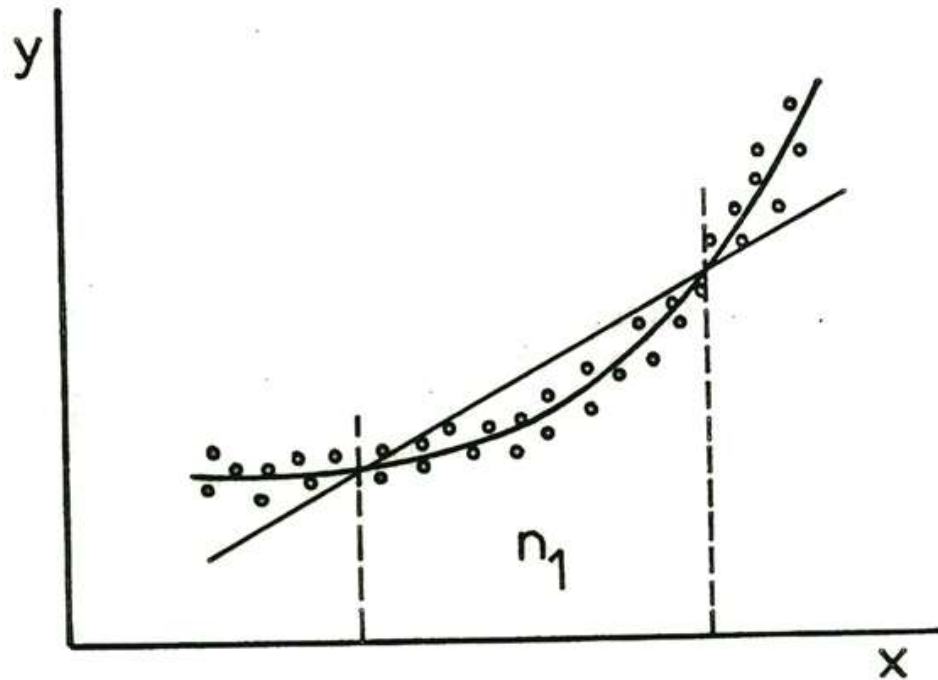


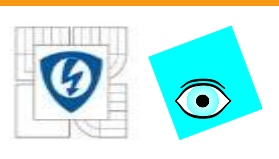


Testy vhodnosti lineárního modelu

1. Test vhodnosti lineárního modelu $f(x, \beta) = X\beta$ dle
Uttsové:

H_0 : lineární model vs. H_A : nelineární model





RSC_1 regresí s využitím n_1 bodů,

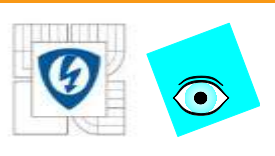
RSC regresí s využitím n bodů.

Testační kritérium

$$F_U = \frac{(RSC - RSC_1)(n_1 - m)}{RSC_1(n - n_1)}$$

Uttsová: volit $n_1 \approx n/2$ a body co nejblíže k těžišti

Test: Je-li $F_U < F_{1-\alpha}(n - n_1, n_1 - m)$, je H_0 přijata.



Test kvadratického členu

$H_0: \beta_2 = 0$ (test významnosti β_2)

$$E(y/x) = \beta_1 x + \beta_2 x^2 + \beta_3$$

RSC_K pro kvadratický model

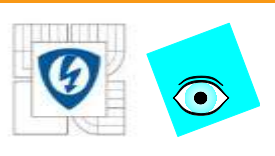
RSC_L pro lineární model

Testační kritérium linearity

$$F_L = \frac{(RSC_L - RSC_K)(n - 3)}{RSC_K(1)}$$

Test:

Je-li $F_L < F_{1-\alpha}(1, n - 3)$, je H_0 přijata.



Linearita testem všech charakteristik

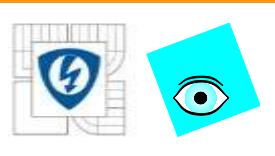
a) střední kvadratická chyba predikce

$$MEP = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T b_{(i)})^2$$

kde $b_{(i)}$ je odhad, určený ze všech bodů kromě i-tého.
Platí vztah

$$MEP = \frac{1}{n} \sum_{i=1}^n \frac{\hat{e}_i^2}{(1 - H_{ii})^2}$$

pro velké n jsou prvky $H_{ii} \approx 0$ a $MEP = \frac{RSC}{n}$.



Linearita testem všech charakteristik

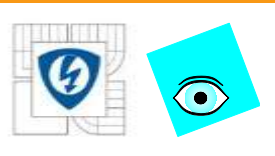
b) Predikovaný koeficient determinace

$$\hat{R}_P^2 = 1 - \frac{n \text{ MEP}}{\sum_{i=1}^n y_i^2 - n\bar{y}}$$

c) Akaikovo informační kritérium

$$AIC = n \ln \left(\frac{RSC}{n} \right) + 2m$$

nejvhodnější model má AIC minimální.



Hodnocení kvality regresního modelu

Střední kvadratická chyba predikce (MEP)

$$MEP = \frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{(1 - H_{ii})^2}$$

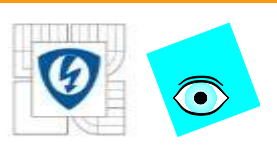
kde e_i^2 je čtverec reziduí modelu a H_{ii} je i -tý diagonální prvek projekční matice H .

Akaikovo informační kritérium (AIC)

$$AIC = n \ln \left(\frac{RSC}{n} \right) + 2m$$

kde RSC je reziduální součet čtverců a m je počet parametrů.

Čím je AIC (MEP) menší, tím je model vhodnější.



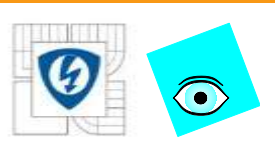
Příklad 6.16 Výběr ze tří polynomických regresních modelů

Regresní analýzou dat vyšetřete, zda místo kvadratického modelu by lépe vyhovoval polynom třetího, nebo pátého stupně.

Řešení:

a) polynom třetího stupně

MEP , R_p^2 a AIC indikují polynom třetího stupně jako nejvhodnější $\hat{y}_p = 860.2(\pm 85.17) - 5.057(\pm 0.485)x + 9.77 \cdot 10^{-3}(\pm 9.19 \cdot 10^{-4})x^2 - 6.146 \cdot 10^{-6}(\pm 5.78 \cdot 10^{-7})x^3$ Přičemž odhady všech tří parametrů vycházejí statisticky významné.

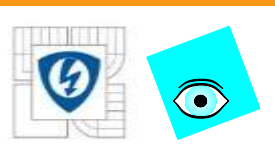


Pokračování příkladu

b) polynom pátého stupně:

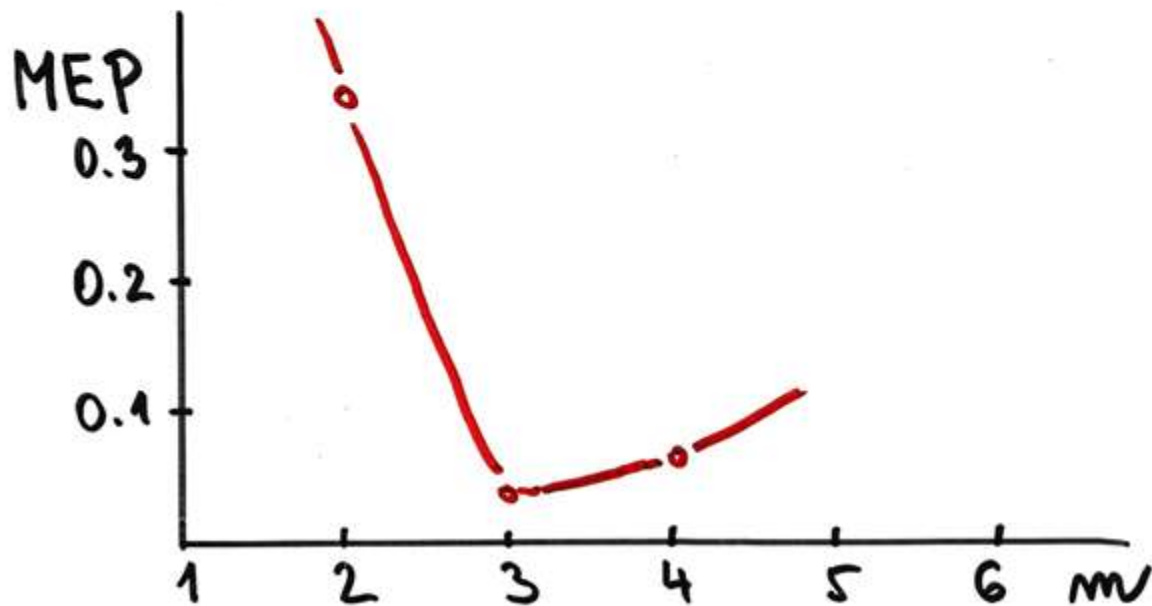
všechny parametry β kromě β_3 vycházejí statisticky nevýznamné, protože se zde projevuje multikolinearita.

Tabulka 6.3 Rozlišení stupně regresního polynomu statistikami MEP , $\hat{R}_p^2$, $\hat{R}^2$ a AIC



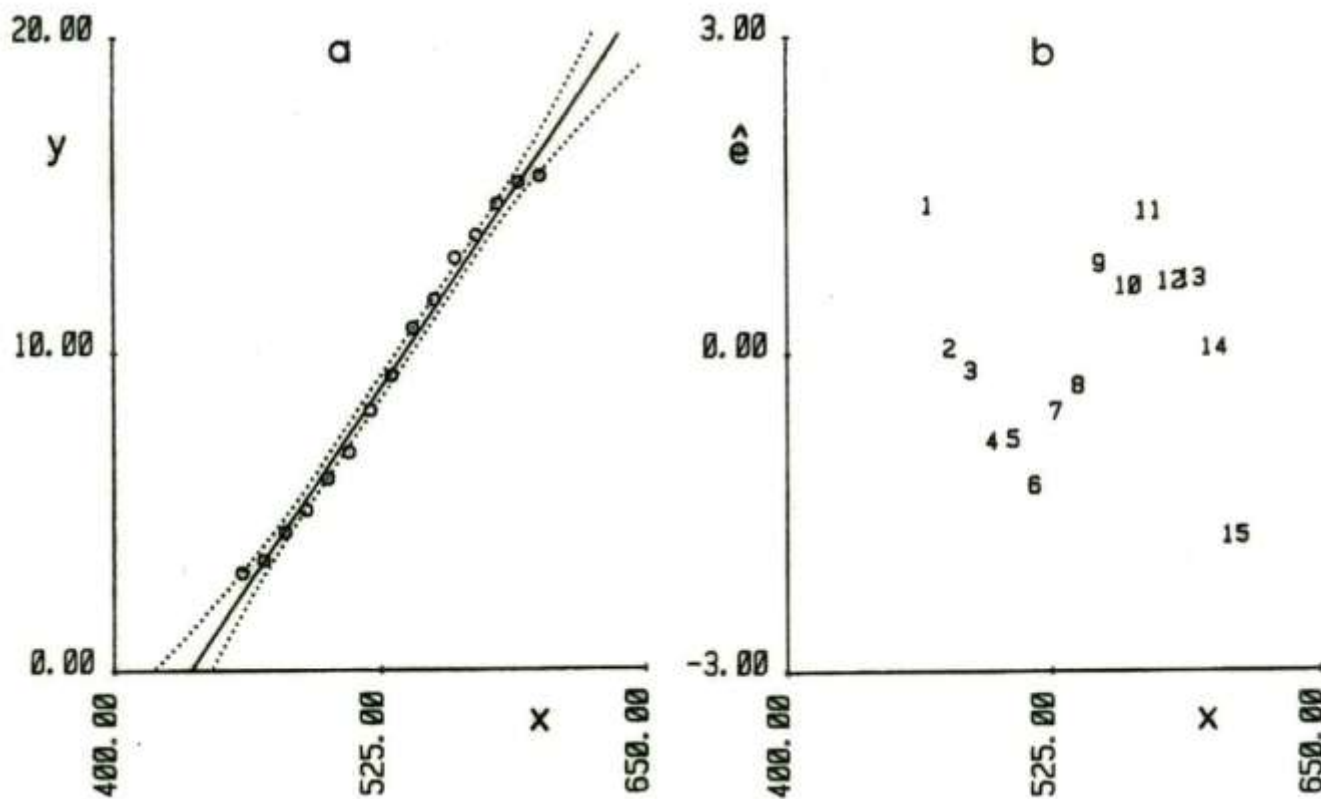
Pokračování příkladu

Stupeň polynomu	MEP	$\hat{R}_p^2$	$\hat{R}^2$	AIC
$m = 2$	0.3502	0.9905	0.9915	-21.65
3	0.0283	0.9992	0.9992	-56.02
5	0.0613	0.9983	0.9997	-55.04





Pokračování příkladu



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

