6. Lineární regresní modely

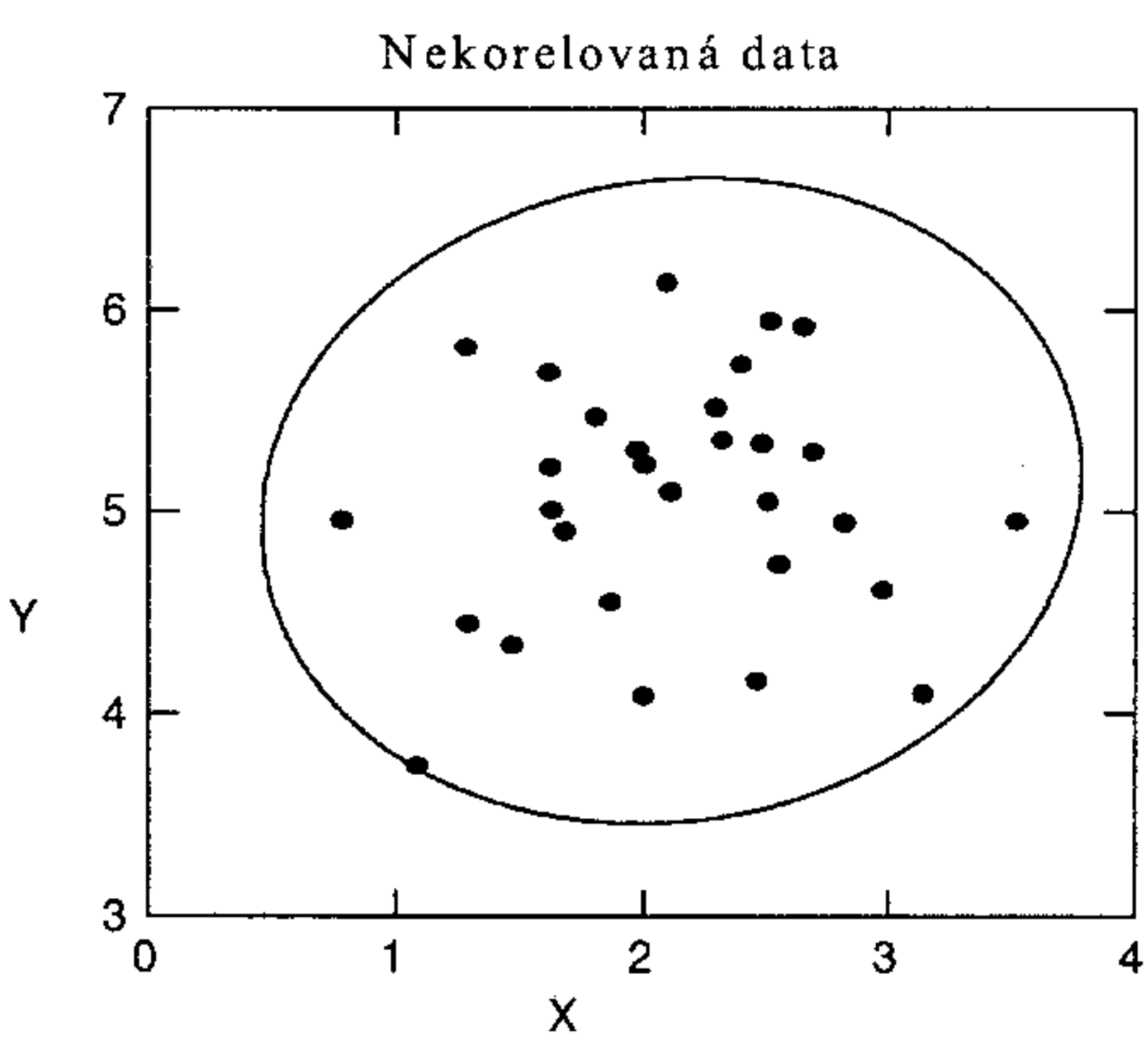
6.1 Jednoduchá regrese a validace
6.2 Testy hypotéz v lineární regresi
6.3 Kritika dat v regresním tripletu
6.4 Multikolinearita a polynomy
6.5 Kritika modelu v regresním tripletu
6.6 Kritika metody v regresním tripletu
6.7 Lineární a nelineární kalibrace
7. Korelační modely

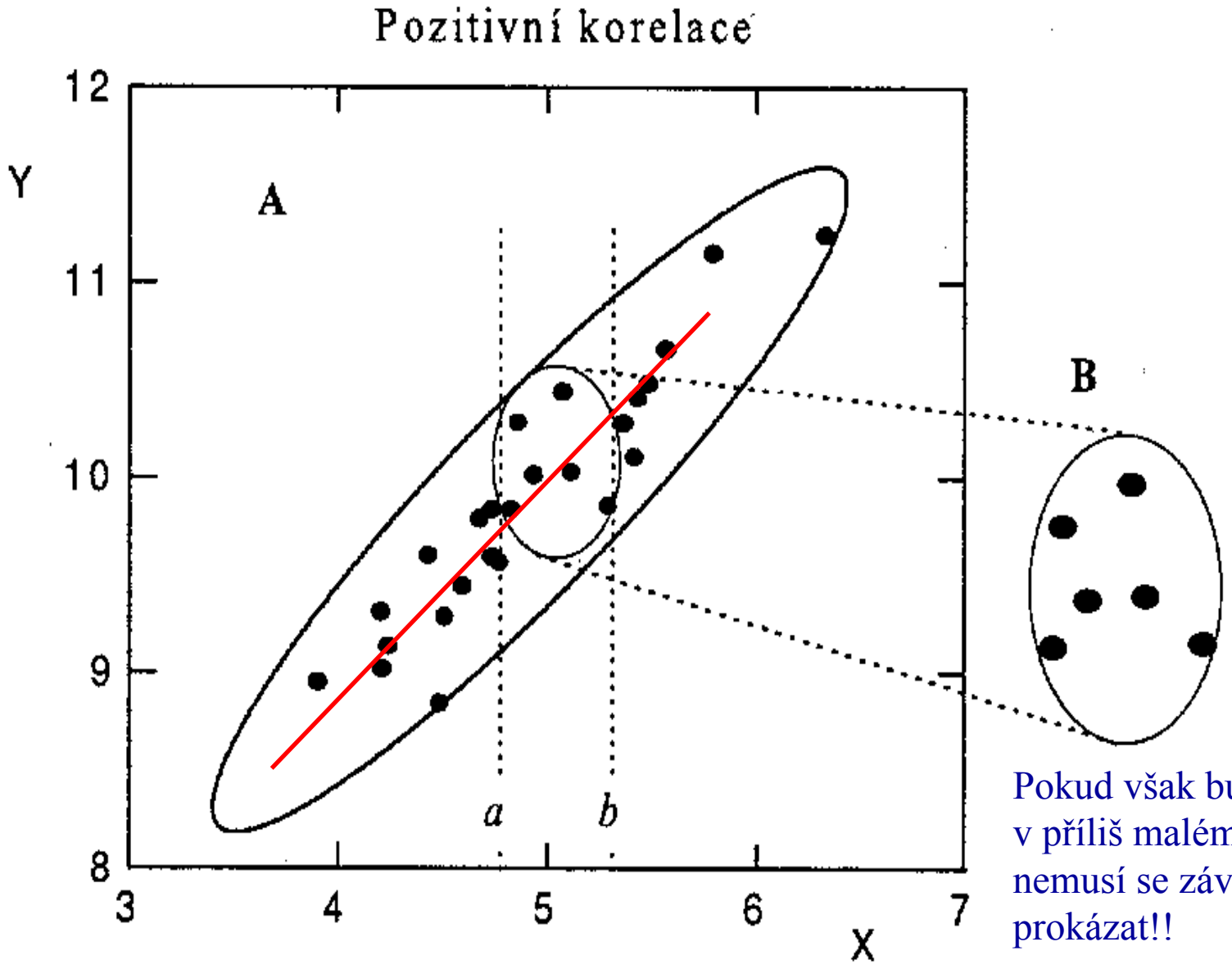
DRUHY STATISTICKÝCH ZÁVISLOSTÍ

- **Korelace** popisuje vliv změny úrovně jednoho znaku na změnu úrovně jiných znaků a platí pro **kvantitativní (měřené) znaky;**

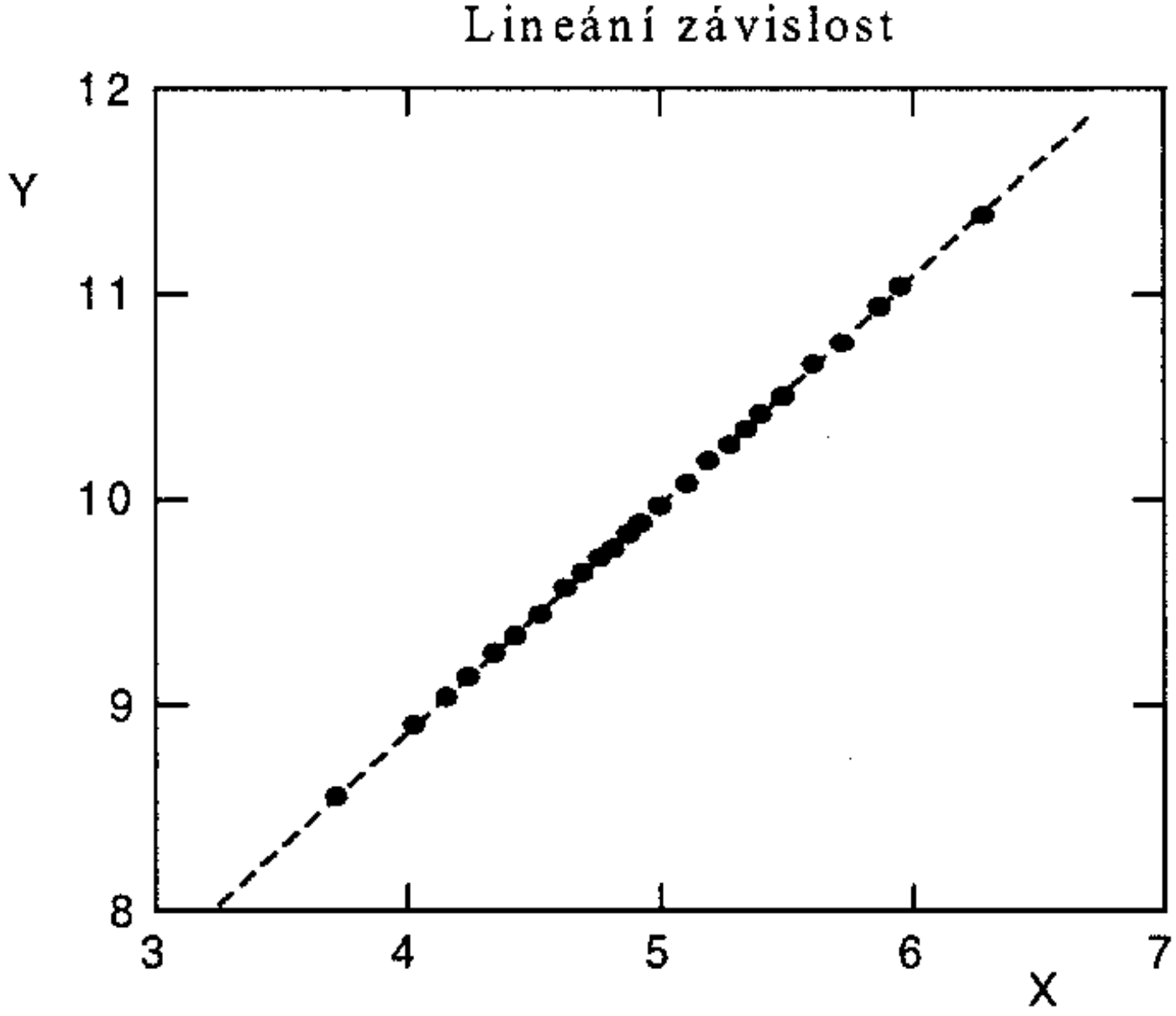
- **Kontingence** popisuje závislost **kvalitativních** (slovních) znaků, které mají více jak dvě alternativy možných znaků (např. druh dřeviny, národnost, apod.);

- **Asociace** popisuje závislost **kvalitativních** (slovních) znaků, které mají pouze dvě alternativy možných znaků (např. pohlaví, odpovědi typu ano/ne, …).


Nekorelovaná data
7
6
5
4
3
Y
0
1
2
3
4
X



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


Lineání závislost
12
11
10
9
8
Y
3
4
5
6
7
X

Cíl regresní analýzy

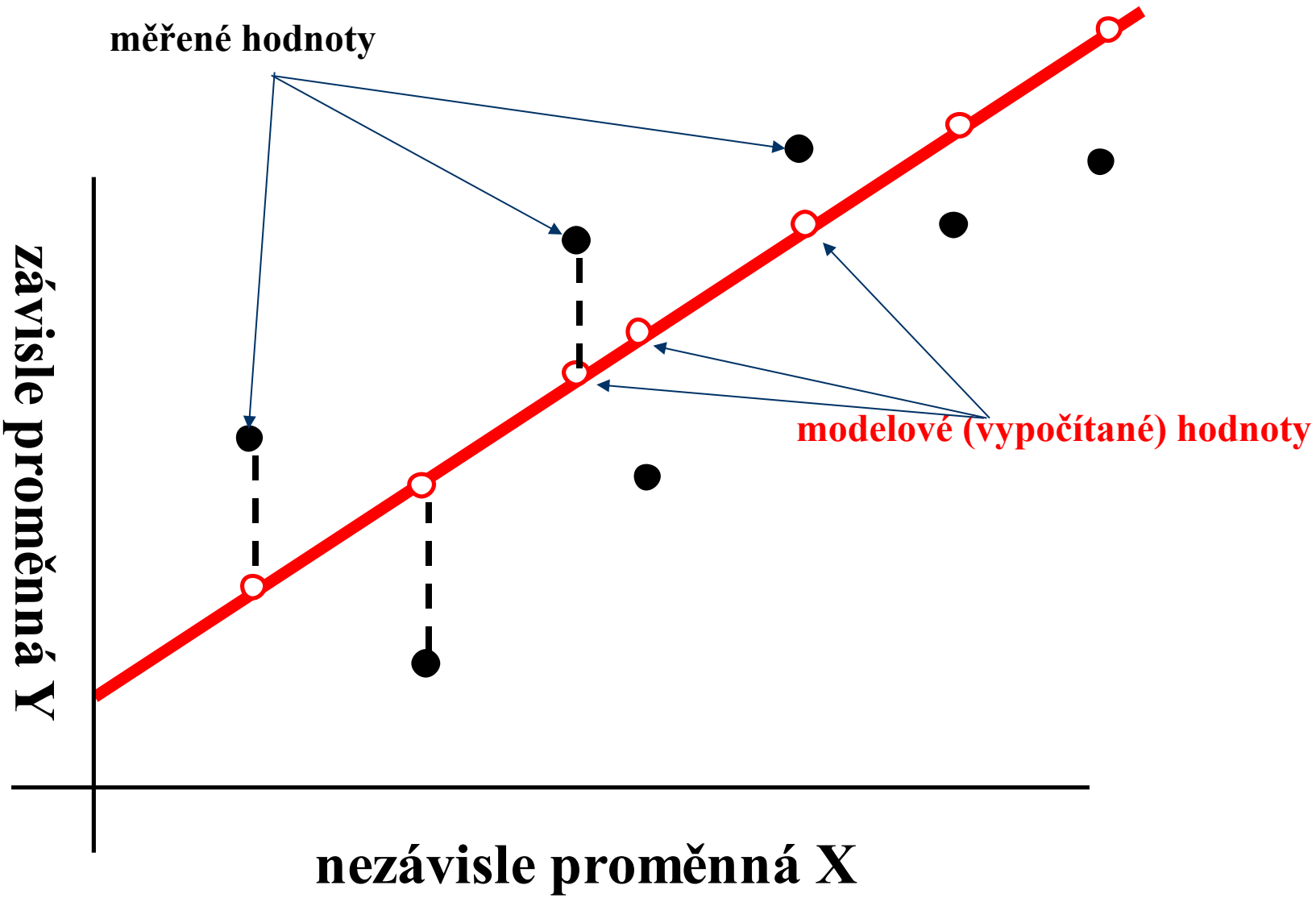
Cílem regresní analýzy je nalezení vhodného modelu studované závislosti tak, že se **snažíme nahradit**

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

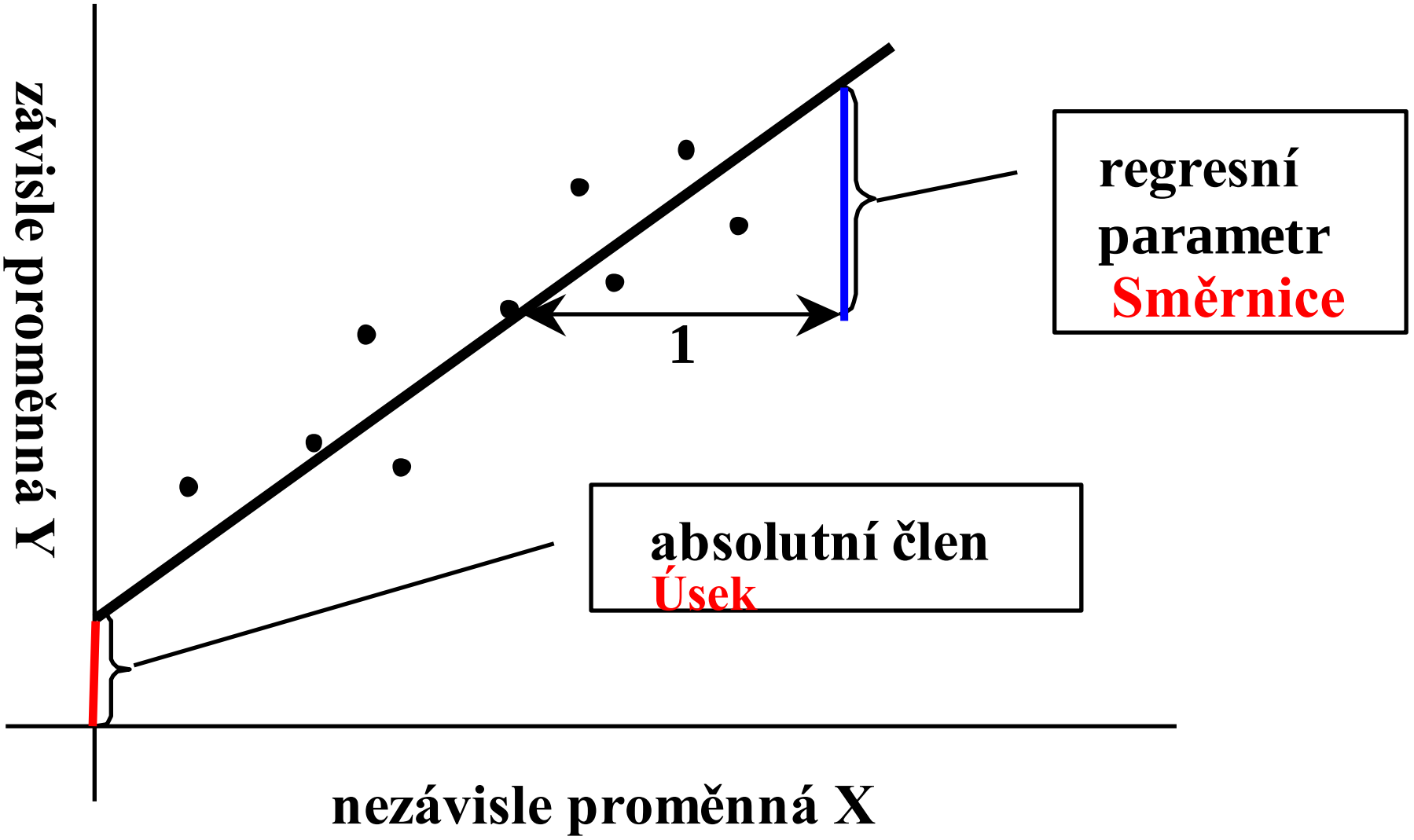
hodnotou vypočtenou (predikovanou) y_{vyp}

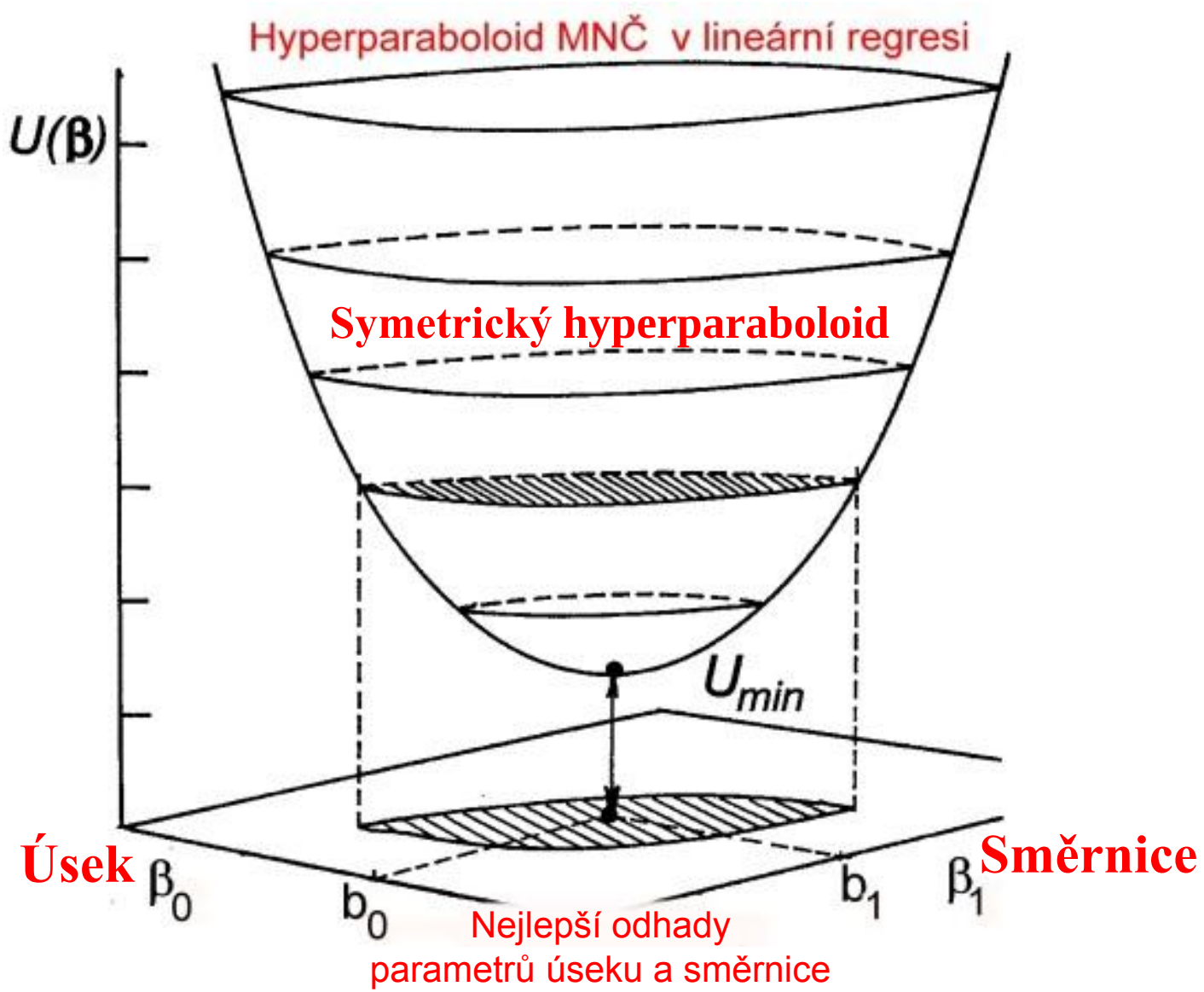
čili **hodnotou** ležící na spojité funkci (modelu) **nezávisle proměnné** x .

Grafické vysvětlení cíle regresní analýzy



Grafické vysvětlení regresního modelu:





Účelová funkce U dosáhne minima pro nejlepší odhady parametrů úseku a směrnice

Popis závislostí

Příklad: Český hydrometeorologický ústav v Praze měřil na stanicích s různou nadmořskou výškou průměrnou roční teplotu půdy. Údaje jsou uvedeny v následující tabulce.

Existuje mezi oběma proměnnými nějaká závislost?

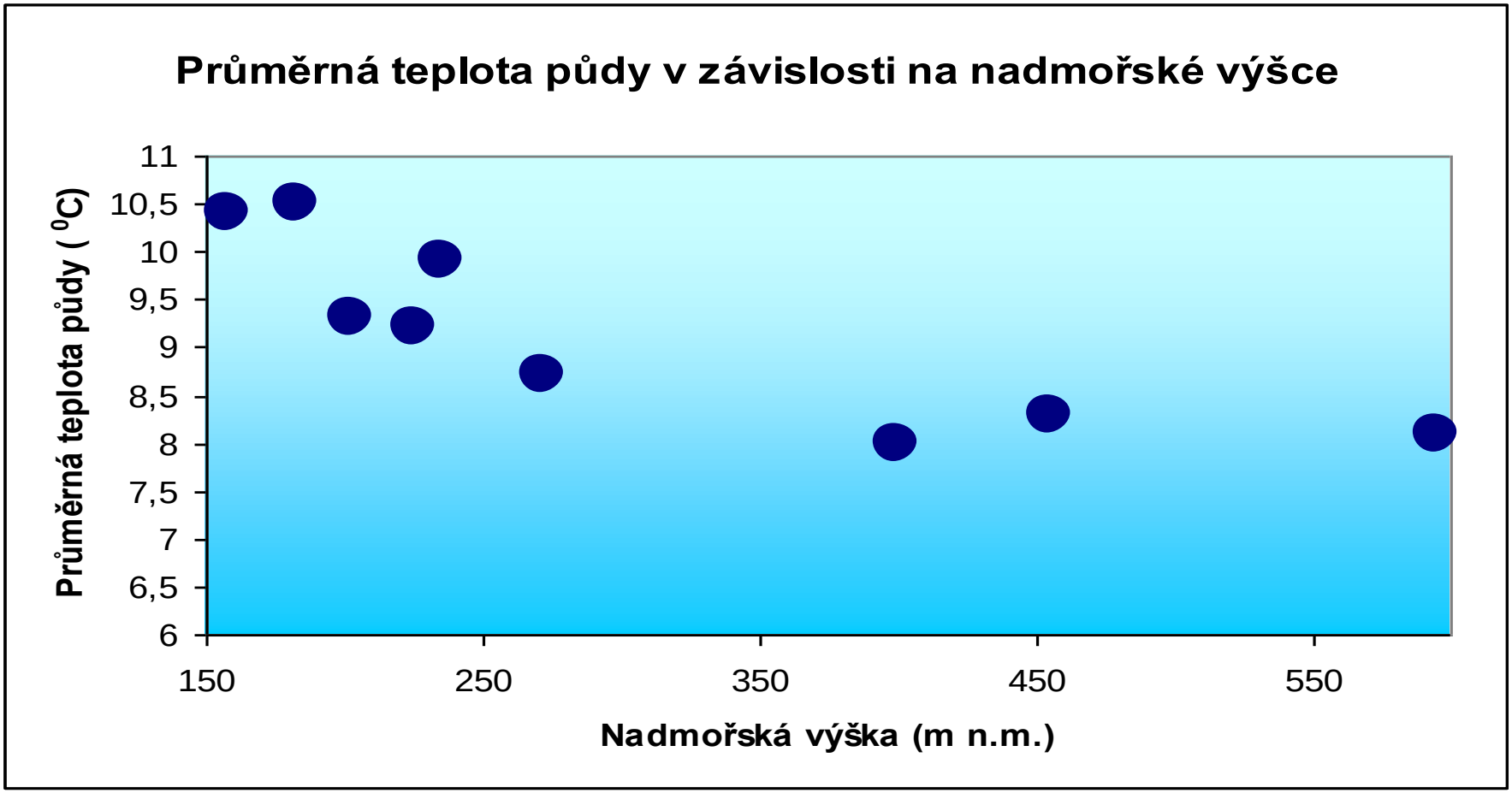
Data:

Nadmořská výška v m n.m.	158	183	203	225	235	272	400	455	595
Průměrná teplota půdy ^{0}C	10,4	10,5	9,3	9,2	9,9	8,7	8	8,3	8,1

Příklad: Český hydrometeorologický ústav v Praze měřil ...

Jednoduché metody k popisu závislosti vystihuje korelační koeficient r_{yx} = -0,835 a také rozptylový bodový graf

Jedná se o silnou závislost, ne však deterministickou, protože kromě výšky zde působí na teplotu i jiné faktory. S rostoucí výškou průměrná teplota půdy klesá.



Regresní model

Zjednodušené zobrazení reality.

Závislost popisuje pomocí rovnice (a v grafu určitou křivkou).

Např. pomocí přímky – lineární závislost:

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

Deterministická složka – vliv vysvětlující

Náhodná složka – všechny ostatní proměnné (nepopsané) vlivy

Deterministická složka η

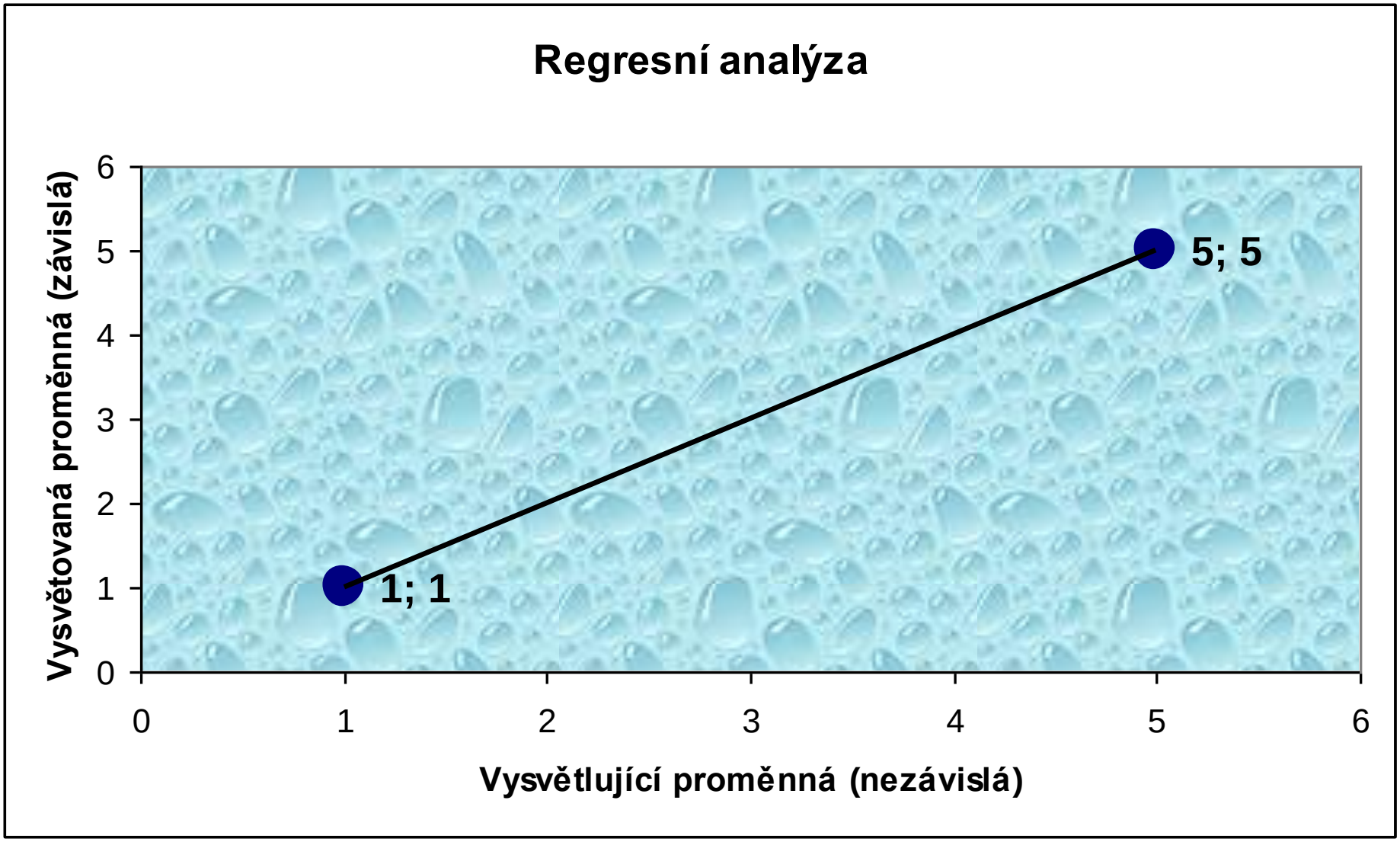
Popisuje závislost mezi hlavními (pozorovanými) proměnnými. Je vyjádřena konkrétní matematickou funkcí.

Náhodná složka ε

Popisuje závislost vysvětlované proměnná na neznámých nebo nepozorovaných proměnných a popisuje i vliv náhody.
Vyjadřuje se pravděpodobnostní funkcí.

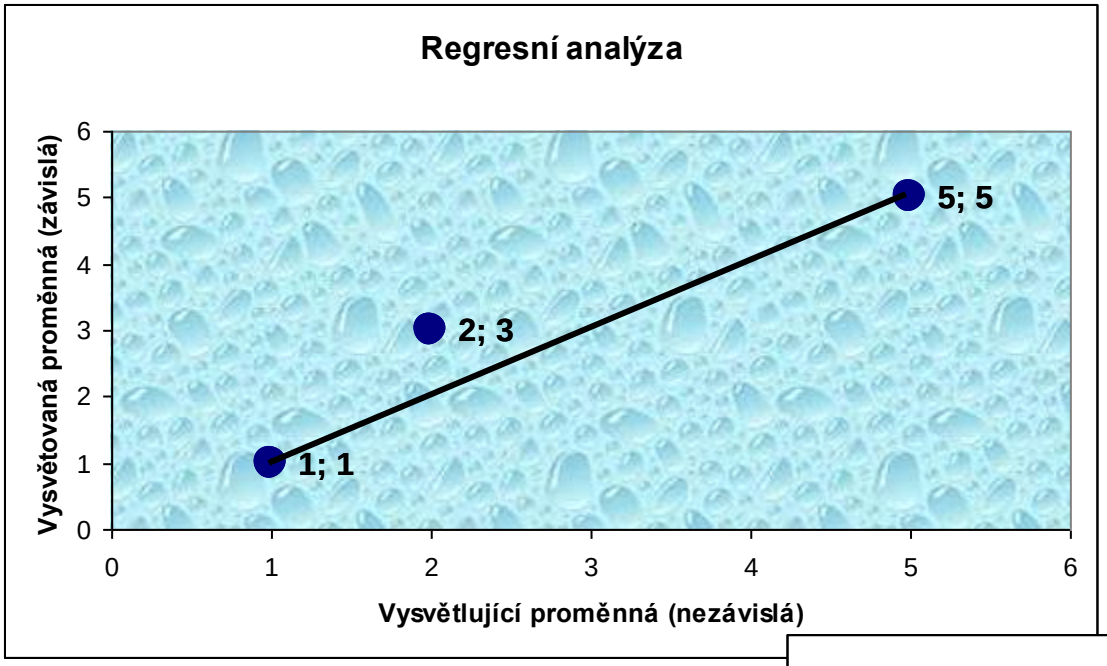
Která regresní přímka je ta správná??

Pokud jsou pouze dva body, je to jejich spojnice. To ovšem není úloha pro statistiku.



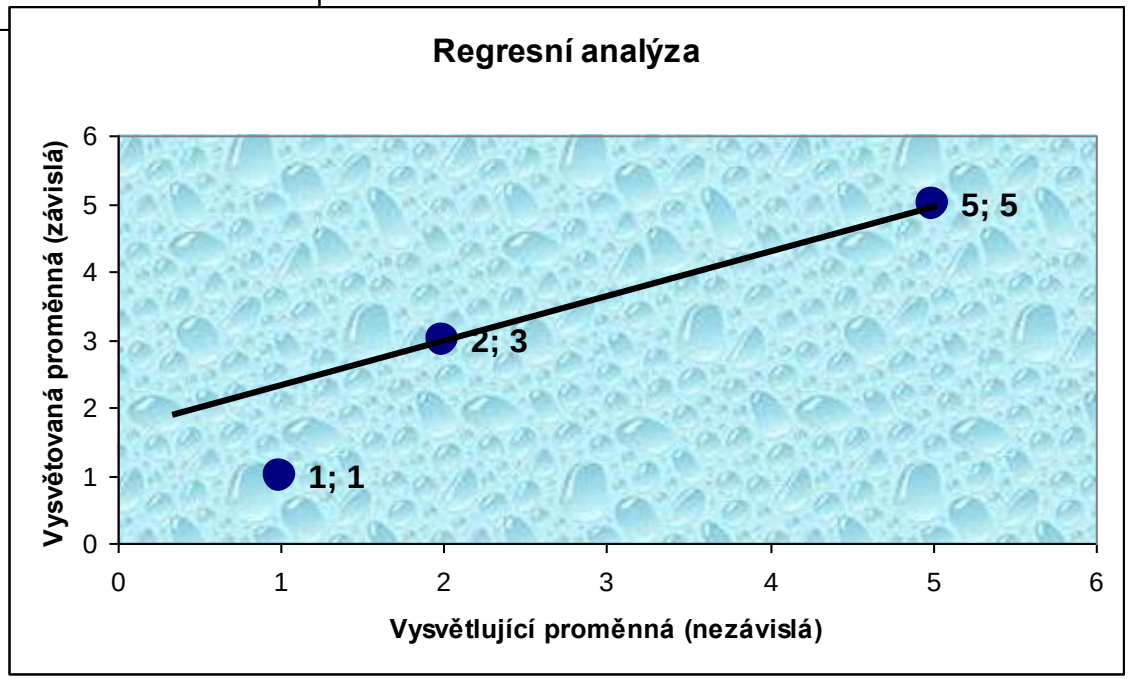
Která regresní přímka je ta správná?

Pokud je více bodů, je to již problém.



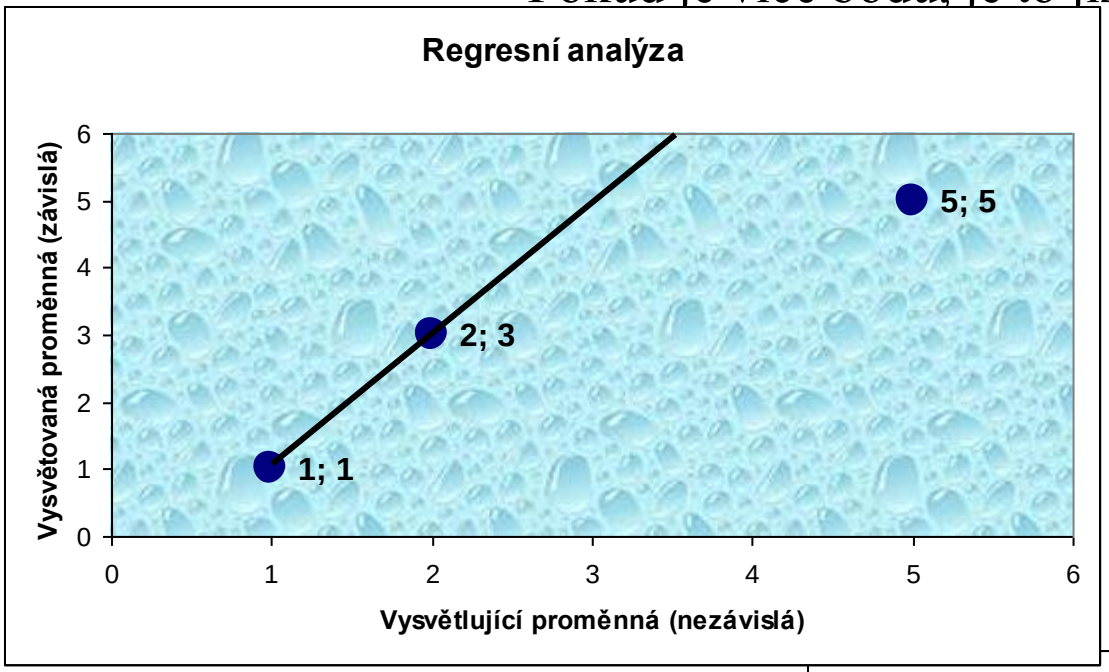
Spojuje přímka jiné dva body?

Spojuje přímka krajní body?



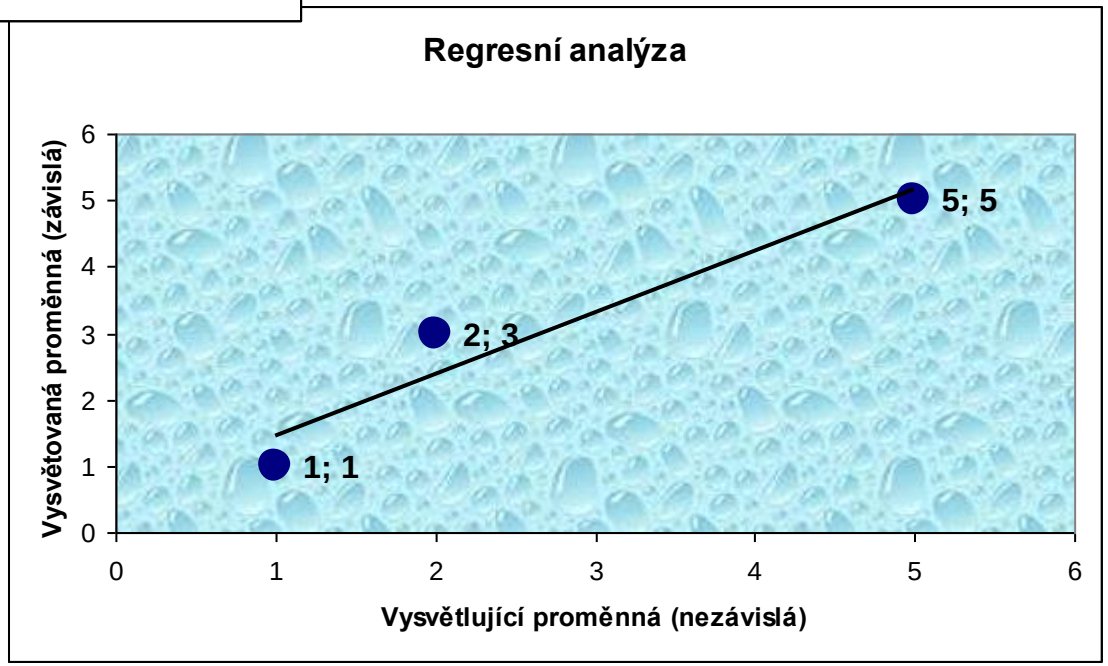
Která regresní přímka je ta správná?

Pokud je více bodů, je to již problém.



Prochází přímka mezi body?

Spojuje přímka jiné dva body?

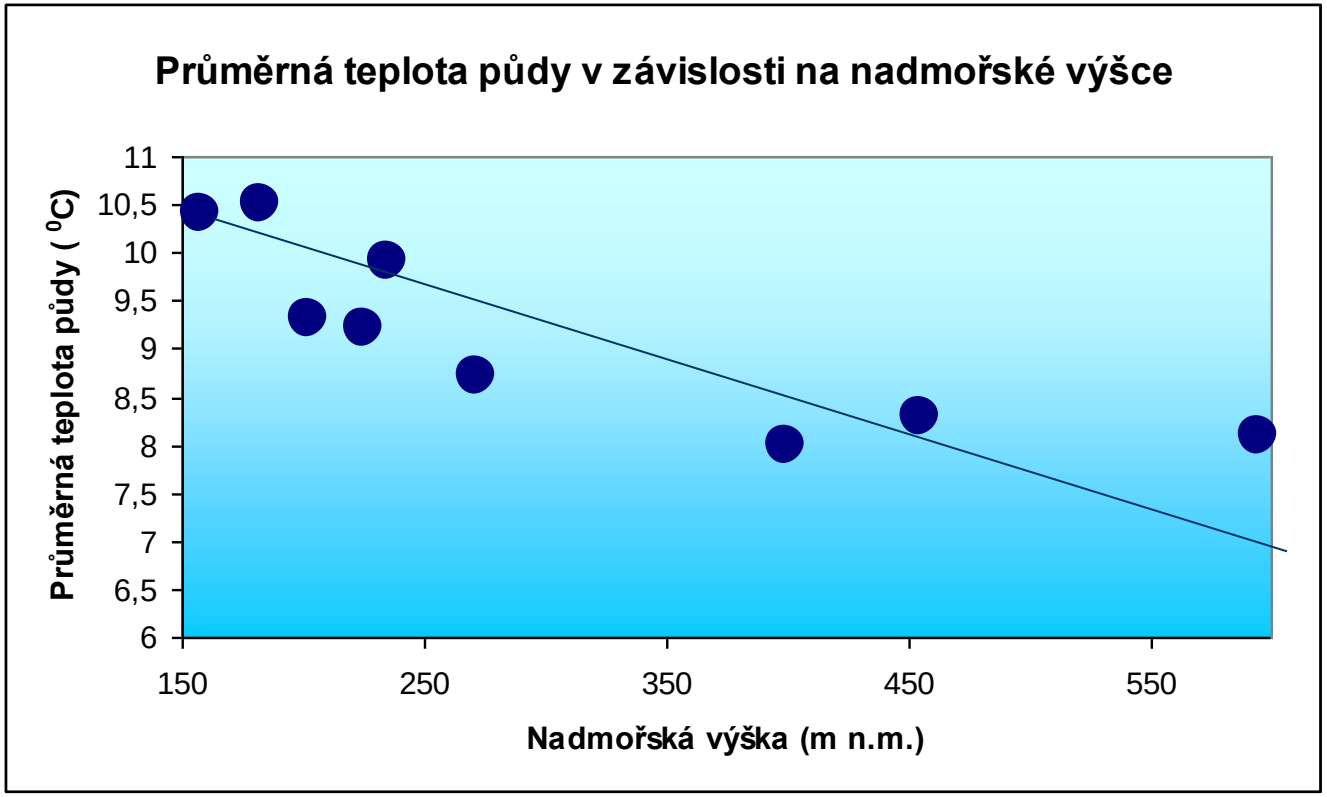


Příklad: Český hydrometeorologický ústav v Praze měřil ...

Pokud body proložíme přímkou, hovoříme o tzv. *regresní přímce*.

Pokud by všechny body ležely na přímce, šlo by o model pouze s deterministickou složkou η.

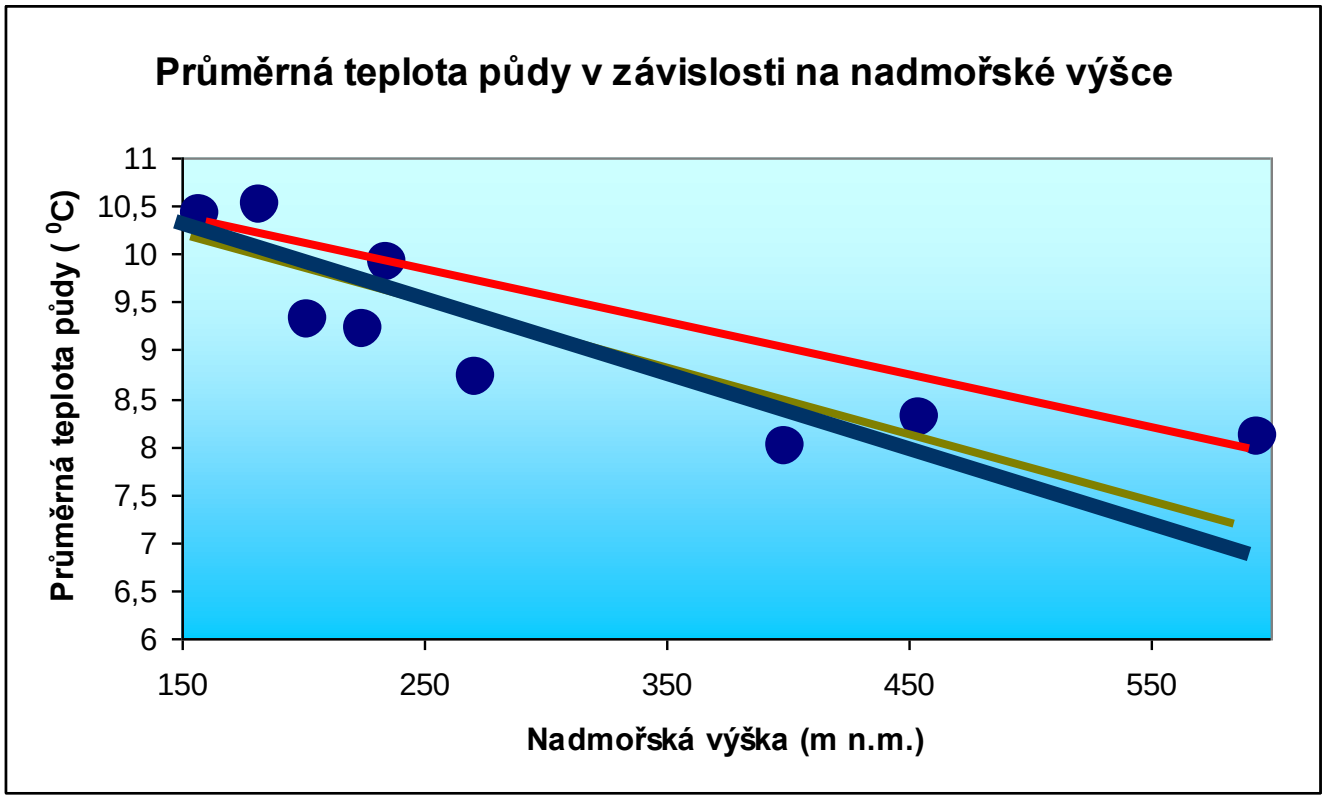
Body však leží i mimo – v modelu je deterministická složka η i náhodná složka ε.



Lze vložit přímku jen tak od oka - zelená.

Lze spojit krajní body – červená.

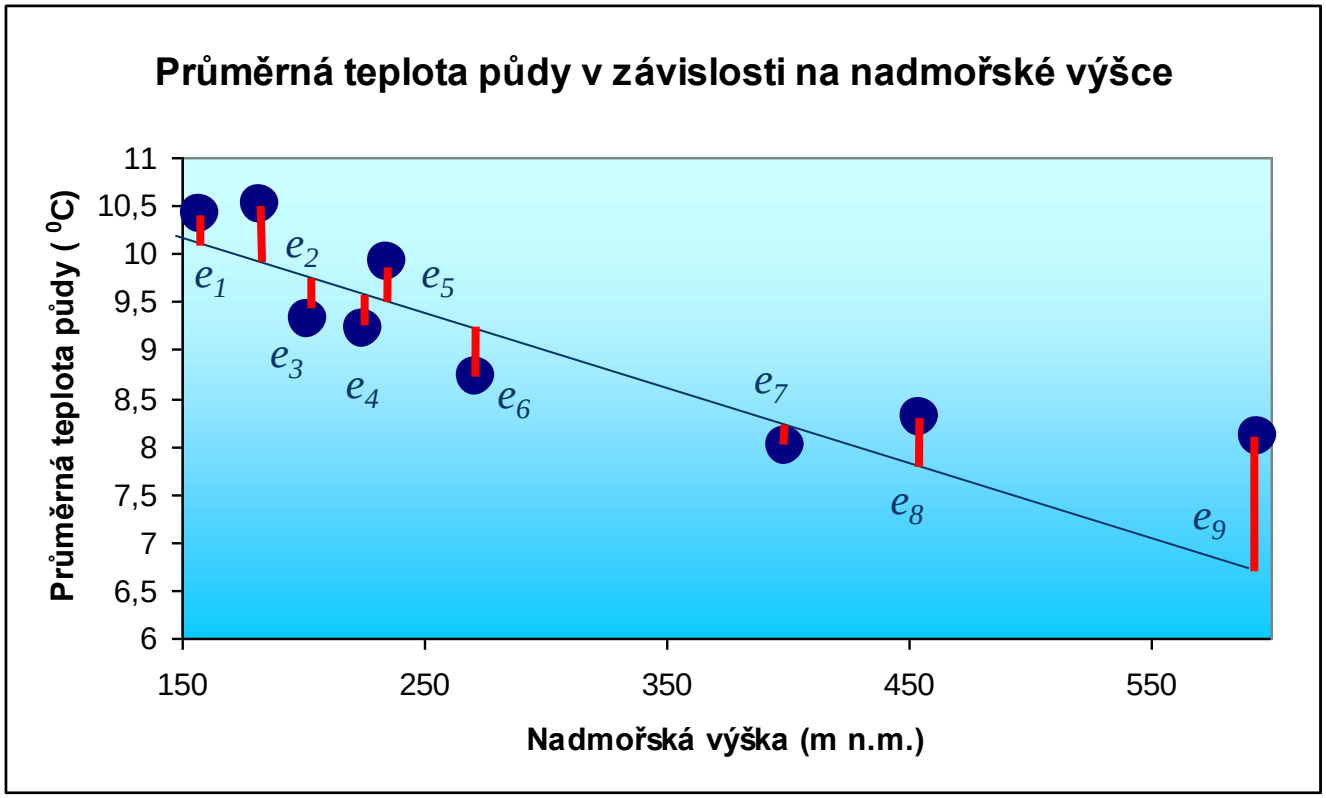
Lze použít nástroje regresní analýzy a nalézt přímku, která prochází ***nejblíže*** všem bodům – černá)



Příklad: Český hydrometeorologický ústav v Praze měřil ...

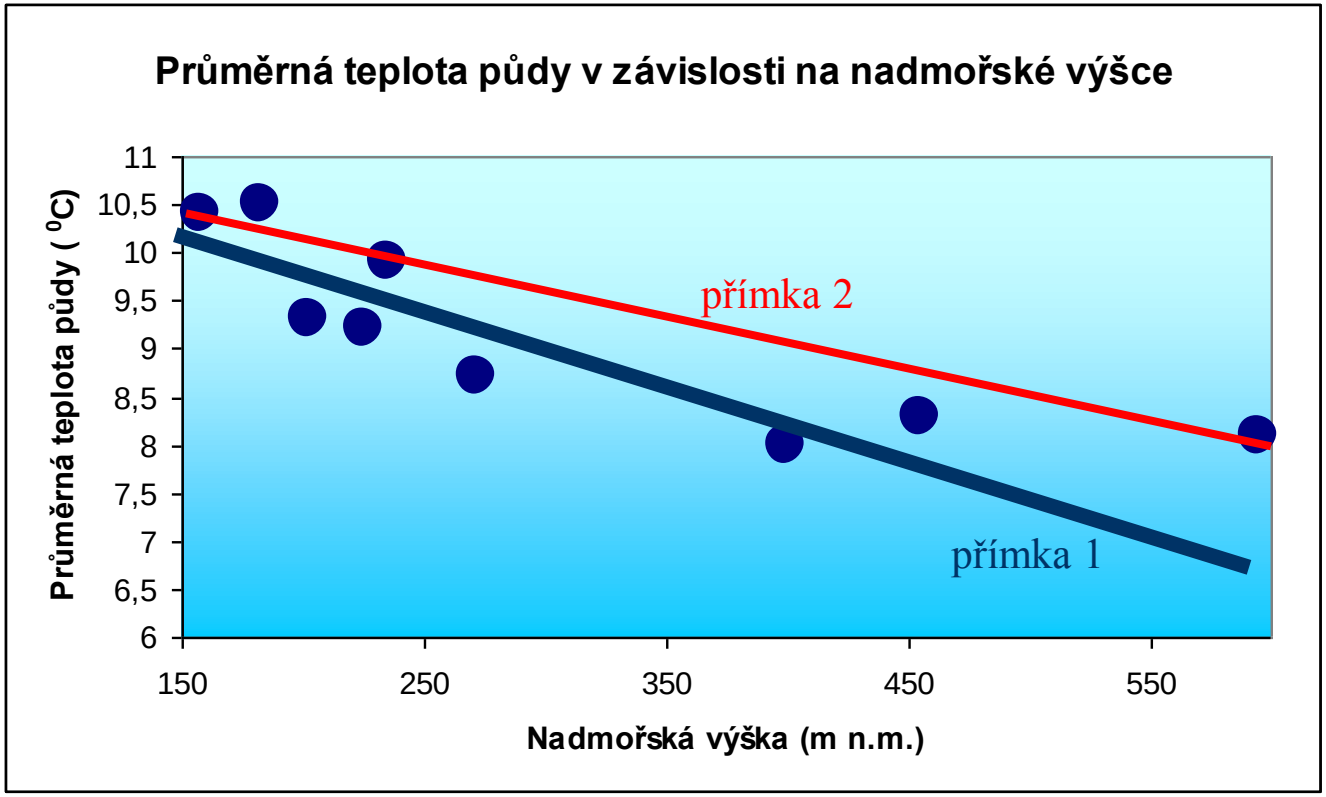
Přímka procházející nejblíže všem bodům je vždy *jen jedna!*
K jejímu nalezení slouží *metoda nejmenších čtverců* (MNČ).

Vybere ze všech možných přímek takovou, pro kterou je součet druhých mocnin (čtverců) odchylek bodů od přímky e_i^2 minimální.



Příklad: Český hydrometeorologický ústav v Praze měřil ...

Přímka označená jako 1 je blíže k bodům, součet čtverců odchylek je menší než u přímky označené jako 2. Přímka 1 je vhodnější.



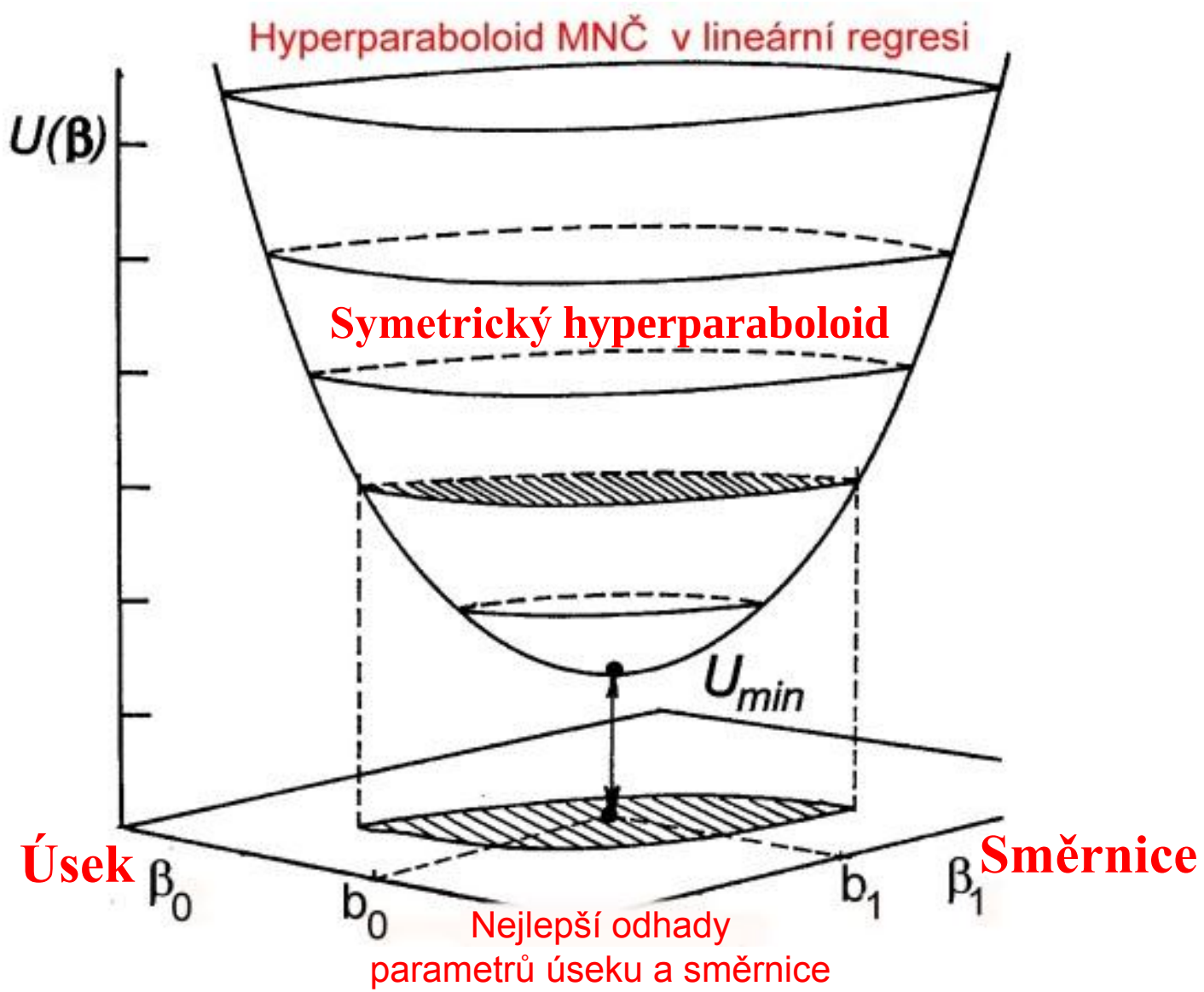
Metoda nejmenších čtverců

Nástroj k určení bodových odhadů parametrů *výběrové regresní přímky*: $\hat{y} = b_0 + b_1 x$

Výběrová je protože je založena pouze na výběrových datech.

Parametry b_0, b_1 jsou *výběrové* (empirické) *regresní parametry*.

Oproti tomu *regresní přímka*: $\eta = \beta_0 + \beta_1 x$, je založena na datech základního souboru, která ale nejsou k dispozici.



Účelová funkce U dosáhne minima pro nejlepší odhady parametrů úseku a směrnice

Metoda nejmenších čtverců

Je založena na řešení *soustavy normálních rovnic* (pro regresní přímku):

$$b_0 n + b_1 \sum x_i = \sum y_i$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i$$

jejichž řešením je:

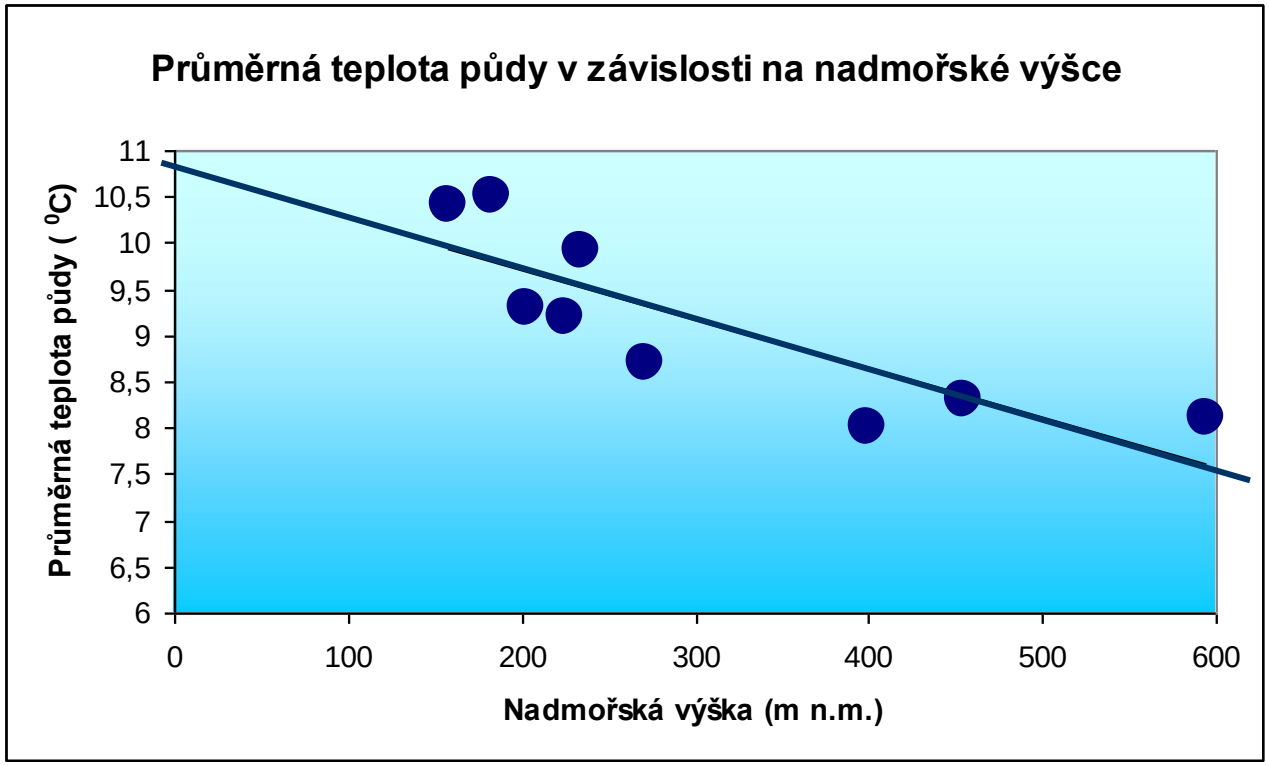
$$b_0 = \frac{\sum y_i \sum x_i - \sum x_i y_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}, b_1 = \frac{n \sum x_i y_i - \sum y_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

Příklad: Český hydrometeorologický ústav v Praze měřil ...

Pomocí metody nejmenších čtverců byla odhadnuta regresní přímka ve tvaru

$\hat{y} = 10{,}795 - 0{,}00541 \cdot x$, kterou lze též zapsat:

průměrná teplota půdy = 10,795 – 0,00541×nadmořská výška

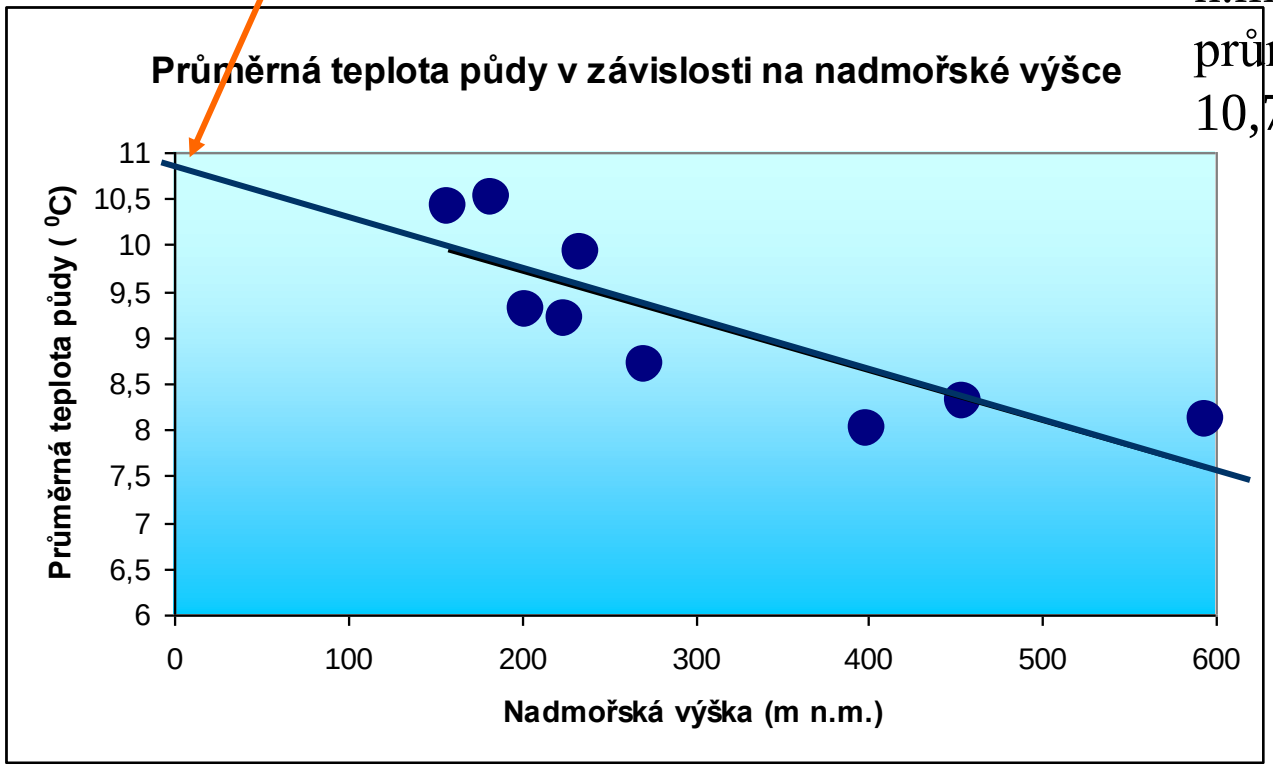


Příklad: Český hydrometeorologický ústav v Praze měřil ...

Pomocí metody nejmenších čtverců byla odhadnuta regresní přímka ve tvaru $\hat{y} = 10{,}795 - 0{,}00541x$.

Parametr
$b_0 = 10{,}795$ je průsečík přímky s osou Y.

V nadmořské výšce 0 metrů n.m. by podle modelu byla průměrná teplota půdy 10,795 ^{0}C.

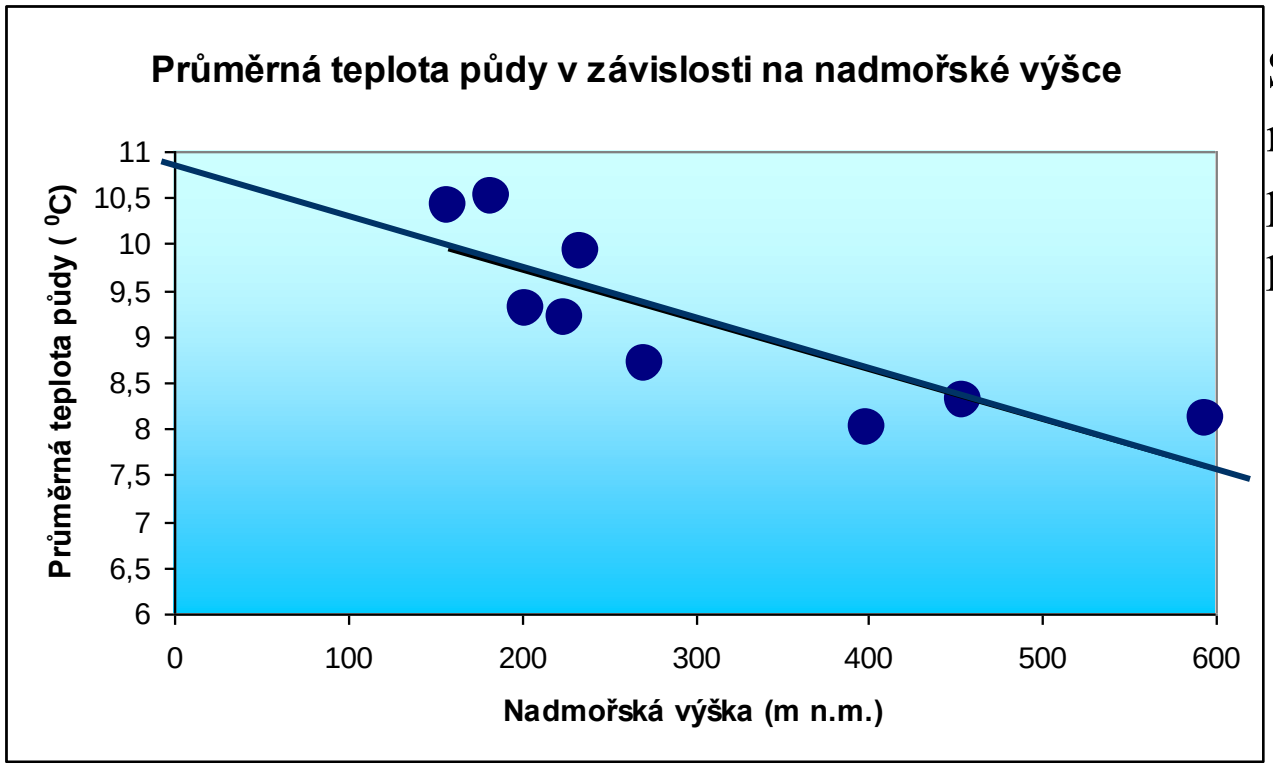


Příklad: Český hydrometeorologický ústav v Praze měřil ...

Pomocí metody nejmenších čtverců byla odhadnuta regresní přímka ve tvaru $\hat{y} = 10{,}795 - 0{,}00541x$.

Parametr $b_1 = 0{,}00541$ je směrnicí přímky a udává její sklon.

Je záporný, protože přímka klesá.



S každým dalším metrem nadmořské výšky klesá průměrná teplota půdy v průměru o 0,00541 °C.

Příklad: Český hydrometeorologický ústav v Praze měřil ...

MS EXCEL: *Nástroje – Analýza Dat - Regrese*

	Koeficienty	*Chyba stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	**10,79504**	0,446866	24,15722	5,3E-08	9,738368	11,85171
Nadmořská výška	**-0,00541**	0,00134	-4,03872	0,004941	-0,00858	-0,00224

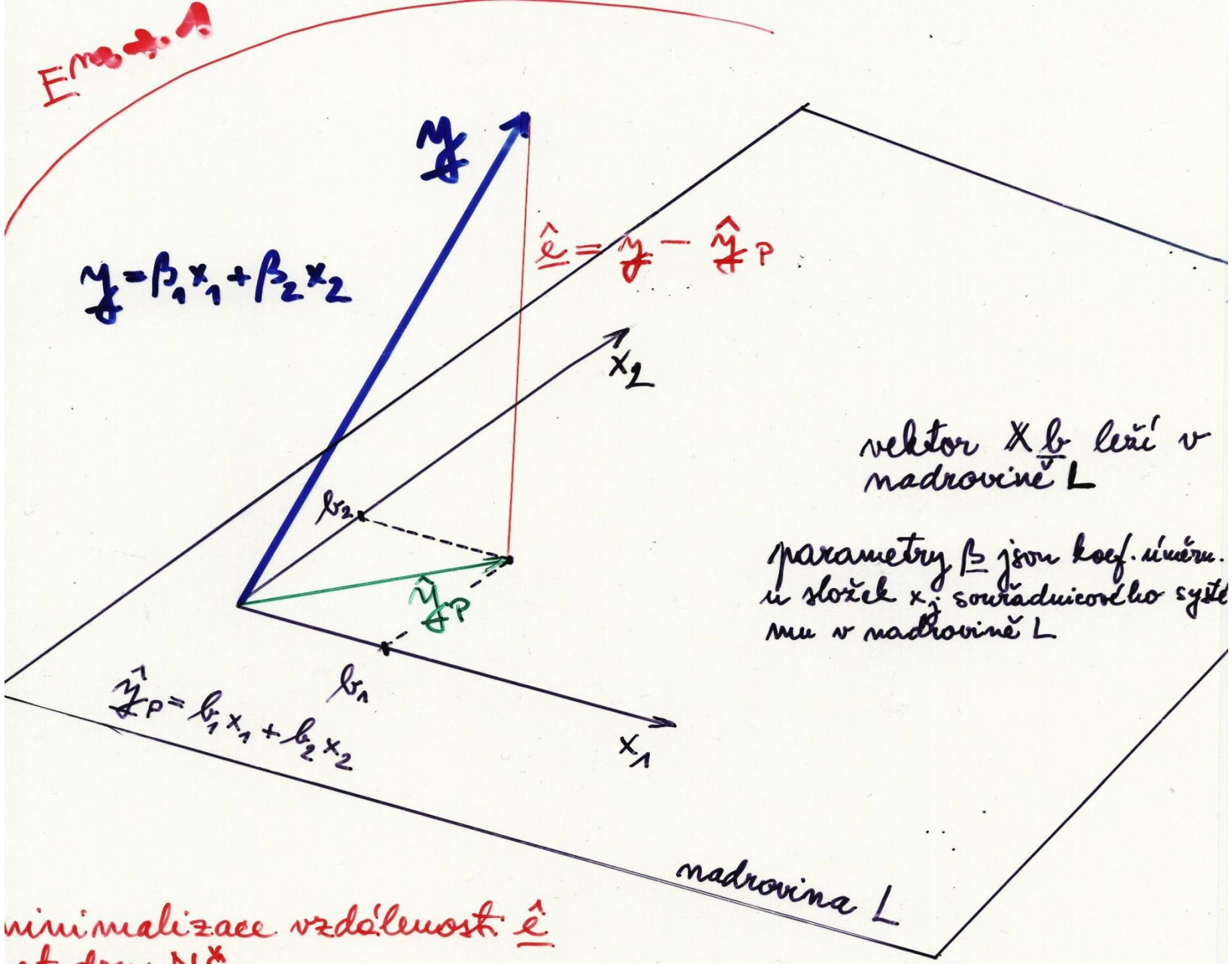
Regresní přímka ve tvaru $\hat{y} = 10{,}795 - 0{,}00541x$

Formulace lineárního regresního modelu

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

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

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



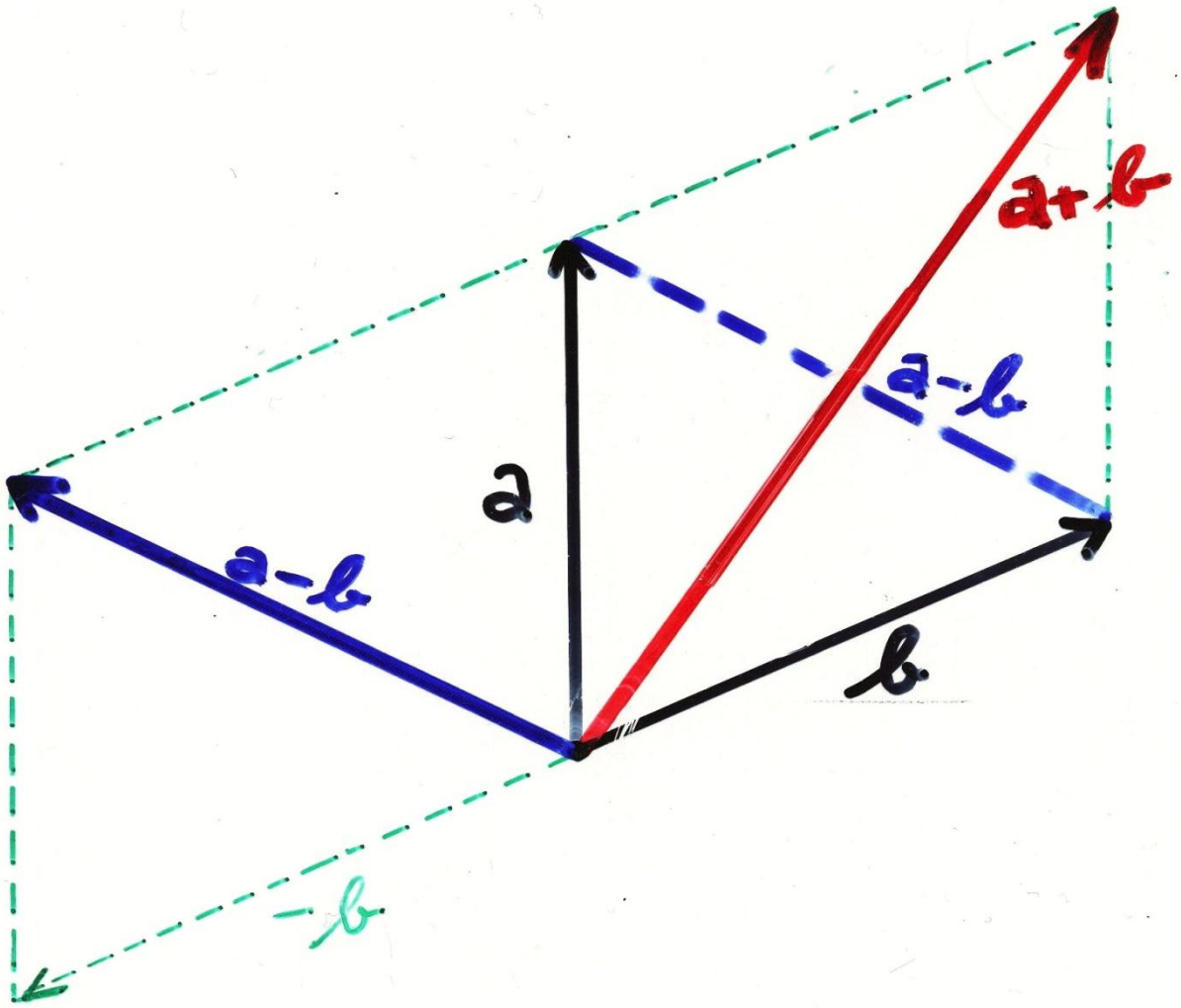
vektor $\underline{b}$ leží v nadrovině L

parametry $\underline{b}$ jsou koef. úměrn. u složek x_j souřadnicového systému v nadrovině L

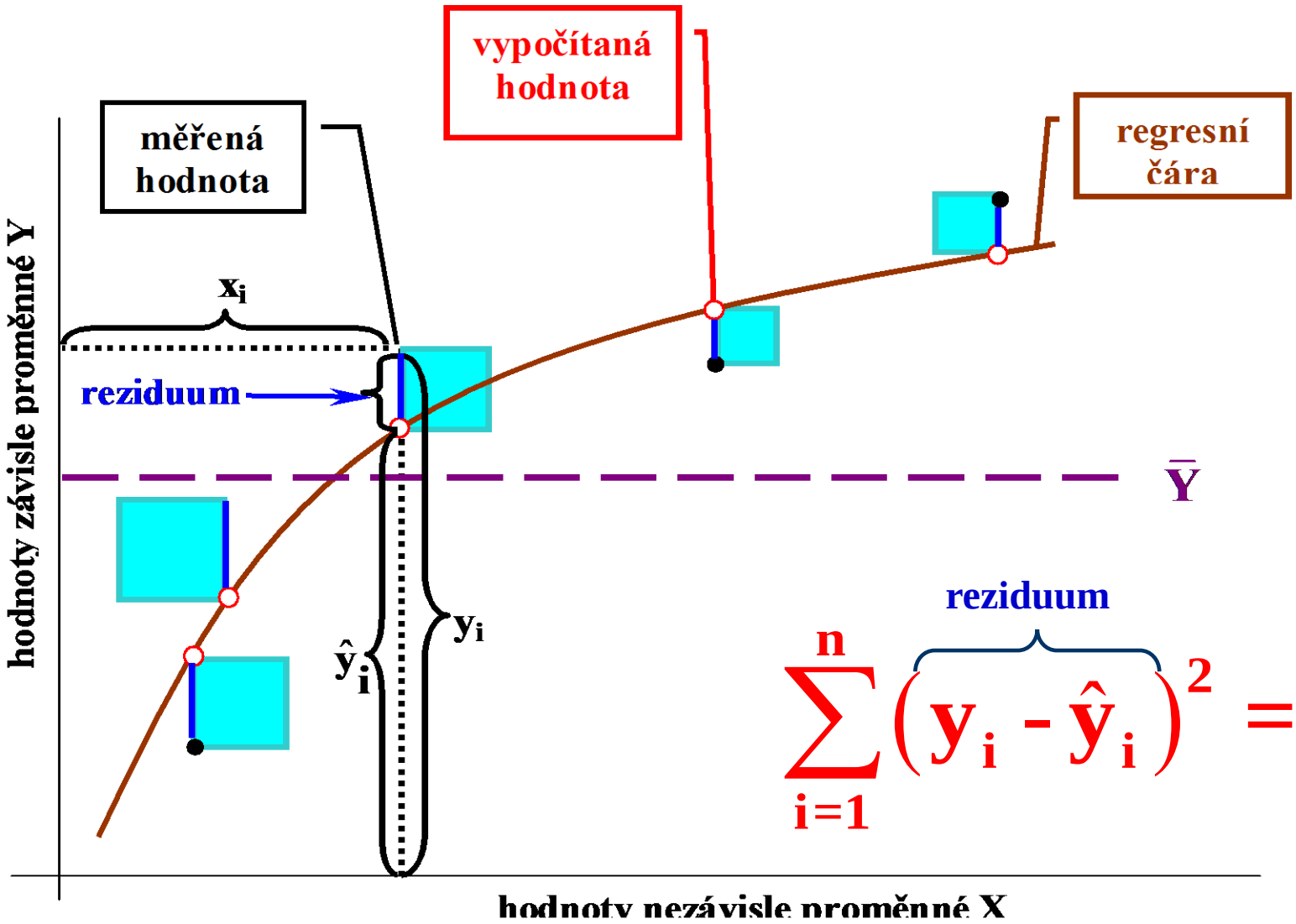
minimalizace vzdálenosti $\underline{\hat{e}}$
metodou NČ

$$\underline{\hat{e}} = \underline{y} - \underline{\hat{y}}_P = \underline{y} - X \cdot \underline{b}$$

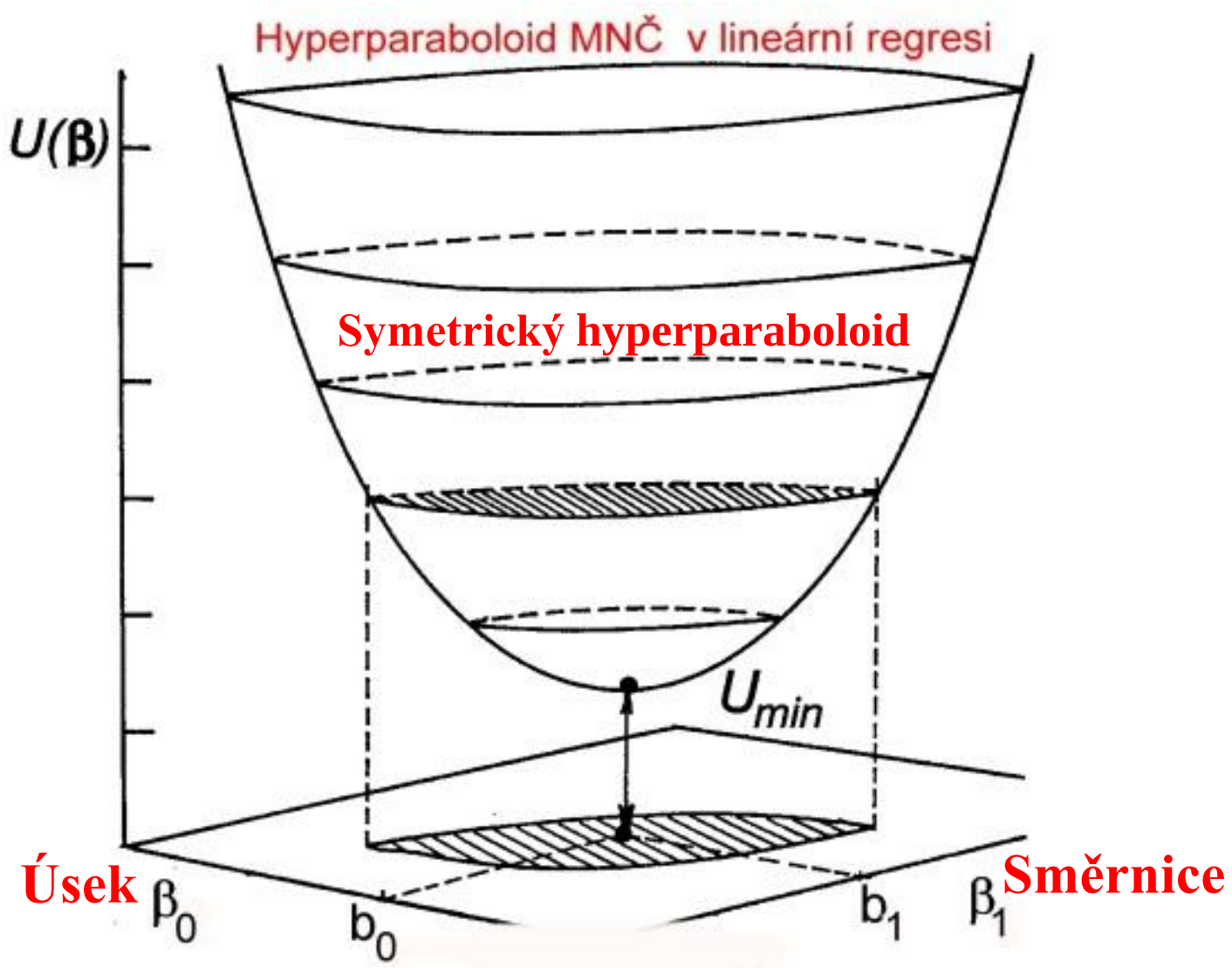
SČÍTÁNÍ A ODČÍTÁNÍ VEKTORŮ:



Vyčíslení odhadů parametrů lineárního regresního modelu metodou nejmenších čtverců (MNČ)



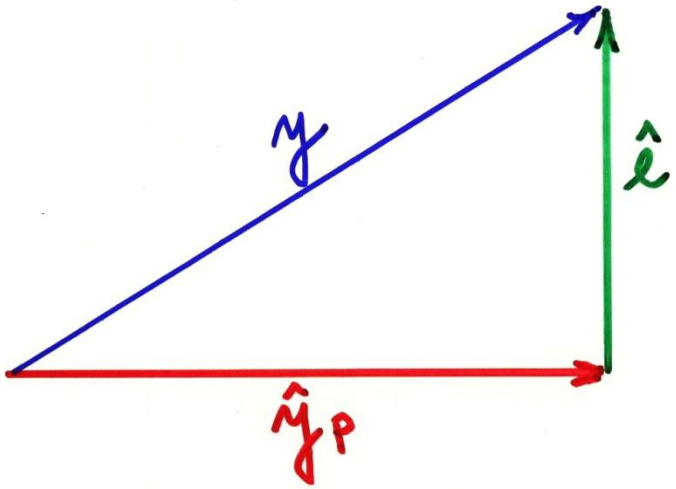
$$\sum_{i=1}^{n} \overbrace{\left(y_i - \hat{y}_i\right)}^{\text{reziduum}}{}^2 = \min.$$



Nejlepší odhady
parametrů úseku a směrnice

Účelová funkce U dosáhne minima pro nejlepší odhady parametrů úseku a směrnice

Metoda projekčních matic



y	vektor proměnné
$\hat{y}_P$	vektor predikce
$\hat{e}$	vektor reziduí

Cílem: minimální délka vektoru $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}_\mathbf{p}$

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

Odhady modelových parametrů **b** minimalizují čtverec délky vektoru

$$D^2 = \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} x_{ij} b_j \right]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_{P,i})^2$$

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

$$\langle \mathbf{x}_j, \hat{\mathbf{e}} \rangle = \sum_{i=1}^{n} x_{ij}\, \hat{e}_i = 0, \qquad j = 1, \ldots, m$$

a maticově $\mathbf{X}^T \hat{\mathbf{e}} = 0$

a dosazením $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\mathbf{b}$ bude

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X}\, \mathbf{b}$$

Odhad **b**, minimalizující vzdálenost D, má tvar

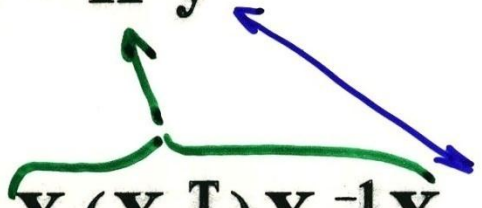
$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

projekční matice **H**, pro kterou platí

$$\hat{\mathbf{y}}_P = \mathbf{H}\ \mathbf{y}$$

a pomocí vektoru **b**

$$\hat{\mathbf{y}}_P = \mathbf{X}\ \ \mathbf{b} = \mathbf{X}(\mathbf{X}^T)\mathbf{X}^{-1}\mathbf{X}\ \ \mathbf{y}$$



Vlastnosti projekční matice H:

1. Projekční matice $\mathbf{H} = \mathbf{X}\ (\mathbf{X}^T\ \mathbf{X})^{-1}\ \mathbf{X}^T$ promítne libovolný vektor **V** do roviny L.

2. Pokud vektor **V** již leží v rovině L, platí $\mathbf{H}\ \mathbf{V} = \mathbf{V}$.

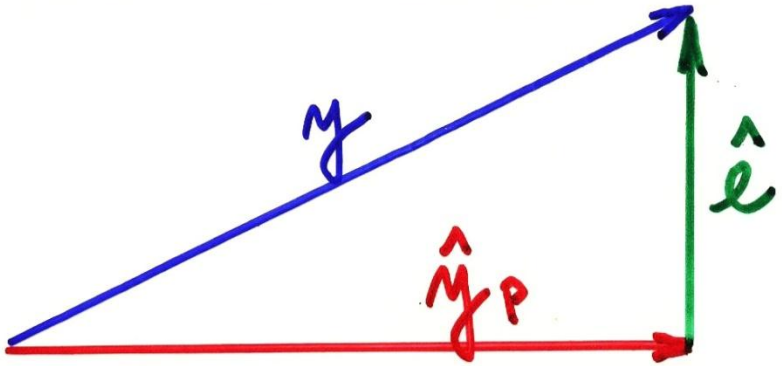
3. Pokud je však vektor **V** kolmý na rovinu L, platí $\mathbf{H}\ \mathbf{V} = \mathbf{0}$, kde **0** je nulový vektor.

Vlastnosti projekční matice P:

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

$$\mathbf{P} = \mathbf{E} - \mathbf{H}$$

kde **E** je jednotková matice.



Závěr: Rozklad vektoru **y** do dvou složek

$$\mathbf{y} = \mathbf{H}\,\mathbf{y} + \mathbf{P}\,\mathbf{y} = \hat{\mathbf{y}}_{\mathbf{P}} + \hat{\mathbf{e}}$$

Geometricky: vektror **y** byl rozložen na dva kolmé vektory

Nejlepší nevychýlené lineární odhady **b** regresních parametrů $\boldsymbol{\beta}$:

B **Nejlepší odhady b:** jejich libovolná lineární kombinace má **nejmenší** rozptyl ze všech lineárních nevychýlených odhadů. (Gaussova-Markova věta).

U **Nevychýlené odhady b:** $E(\boldsymbol{\beta} - \mathbf{b}) = 0$, střední hodnota vektoru odhadů $E(\mathbf{b})$ je rovna vektoru regresních parametrů $\boldsymbol{\beta}$.

L **Lineární odhady b:** lze je zapsat jako lineární kombinaci měření **y** s váhami Q_{ij}, které závisí pouze na polohách proměnných x_j, $j = 1, ..., m$.

E Je-li $\mathbf{Q} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, můžeme psát

$$b_j = \sum_{i=1}^{n} Q_{ij}\, y_i$$

Předpoklady metody nejmenších čtverců

1. Parametry $\boldsymbol{\beta}$ mohou nabývat libovolných hodnot. Omezení jsou pouze fyzikálního smyslu.
2. Model je lineární v parametrech $\boldsymbol{\beta}$; platí přitom additivní model měření $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
3. Matice $\mathbf{X}$ je nenáhodná, nastavitelných hodnot nezávisle proměnných. Má hodnost m:
   (a) Žádné dva sloupce x_j a x_k nejsou kolineární (čili paralení).
   (b) $\mathbf{X}^T\mathbf{X}$ je pak symetrická regulární matice.
   (c) Rovina L je m rozměrná a vektory $\mathbf{X}\,\mathbf{b}$ jsou jednoznačné.

4. Náhodné chyby ϵ_i mají nulovou střední hodnotu $E(\epsilon_i) = 0$.

   Je-li $E(\epsilon_i) = K$, je nutno zavést absolutní člen a pak bude

   $E(\epsilon_i) = 0$.

5. Náhodné chyby ϵ_i mají konstantní rozptyl, homoskedasticita,

   $E(\epsilon_i^2) = \sigma^2$.

6. Náhodné chyby ϵ_i jsou vzájemně nekorelované, $\text{cov}(\epsilon_i \, . \, \epsilon_j) =$

   $E(\epsilon_i \, . \, \epsilon_j) = 0$.

7. Náhodné chyby mají normální rozdělení $\epsilon \approx N(0, \sigma^2)$.

Typy regresního modelu

Regresní model předpokládá, že

nezávislá proměnná (proměnné) je **nenáhodná**
(tj. pevně určena experimentátorem) a

závislá proměnná je **náhodná** (měřená).

Tento předpoklad nebývá striktně splněn (v mnoha případech jsou obě nebo všechny veličiny měřené (to znamená náhodné zatížené náhodným šumem) a potom mluvíme o tzv. **korelačním modelu**.

Rozeznáváme:

- **Regresní modely lineární** – mají lineární postavení parametrů
- **Regresní modely nelineární** –mají nelineární postavení parametrů

Typy regresního modelu

Příklady **lineárních** regresních modelů:

$y = \mathbf{a} + \mathbf{b}x$ - přímka

$y = \mathbf{a} + \mathbf{b}x + \mathbf{c}x^2$ - parabola

$y = \mathbf{a} + (\mathbf{b}/x)$ - hyperbola

Lineární modely mohou být i modely, jejichž grafickým vyjádřením je křivka!!

Příklady **nelineárních** regresních modelů:

$y = \mathbf{a} \cdot x^{\mathbf{b}}$

$y = \mathbf{a} \cdot e^{\mathbf{b}x}$

$y = \mathbf{a} \cdot e^{\frac{\mathbf{k}}{x}}$

Výhody nelineární modelů: jsou schopny modelovat složité reálné děje, např. růst, včetně reálné predikce.

Nevýhody nelineárních modelů: daleko složitější výpočet

Obecný postup regresní analýzy

1. Navrhnout vhodný tvar regresního modelu čili postavit příslušnou rovnici či vzorec, který bude popisovat závislost y na x.

2. Určit parametry modelu β vyčíslením jejich konkrétních odhadů b .

3. Určit statistickou významnost modelu, tj. testovat zda navržený model významným způsobem přispěje ke zpřesnění odhadu závisle proměnné oproti použití pouhého průměru všech hodnot y.

4. Predikované hodnoty regresním modelem vysvětlit z hlediska zadání.

Intervaly spolehlivosti v korelační a regresní analýze

- IS korelačního koeficientu (koeficientu determinace)
- IS regresních parametrů
- IS modelových hodnot (modelu)
- IS predikovaných hodnot (pás spolehlivosti)

Statistické vlastnosti metody MNČ

Matice **H** je nenáhodná a pro **kovarianční matici predikce** platí

$$D(\hat{\mathbf{y}}_p) = \sigma^2 \mathbf{H}$$

a analogicky platí pro **kovarianční matici reziduí**

$$D(\hat{\mathbf{e}}) = \sigma^2 \mathbf{P} = \sigma^2 (\mathbf{E} - \mathbf{H})$$

Součet čtverců reziduí RSC lze napsat

$$RSC = \hat{\mathbf{e}}^T \hat{\mathbf{e}} = \mathbf{y}^T (\mathbf{E} - \mathbf{H}) \mathbf{y} = \mathbf{y}^T \mathbf{P} \mathbf{y}$$

a jeho střední hodnotu

$$E(RSC) = \sigma^2 \, tr(\mathbf{P}) = \sigma^2 (n - m)$$

kde tr(**P**) je stopa matice **P** a je rovna její hodnosti.

Pro nestranný odhad rozptylu chyb σ^2 lze využít

$$\hat{\sigma}^2 = \frac{RSC}{n - m} = \frac{\hat{\mathbf{e}}^T \hat{\mathbf{e}}}{n - m}$$

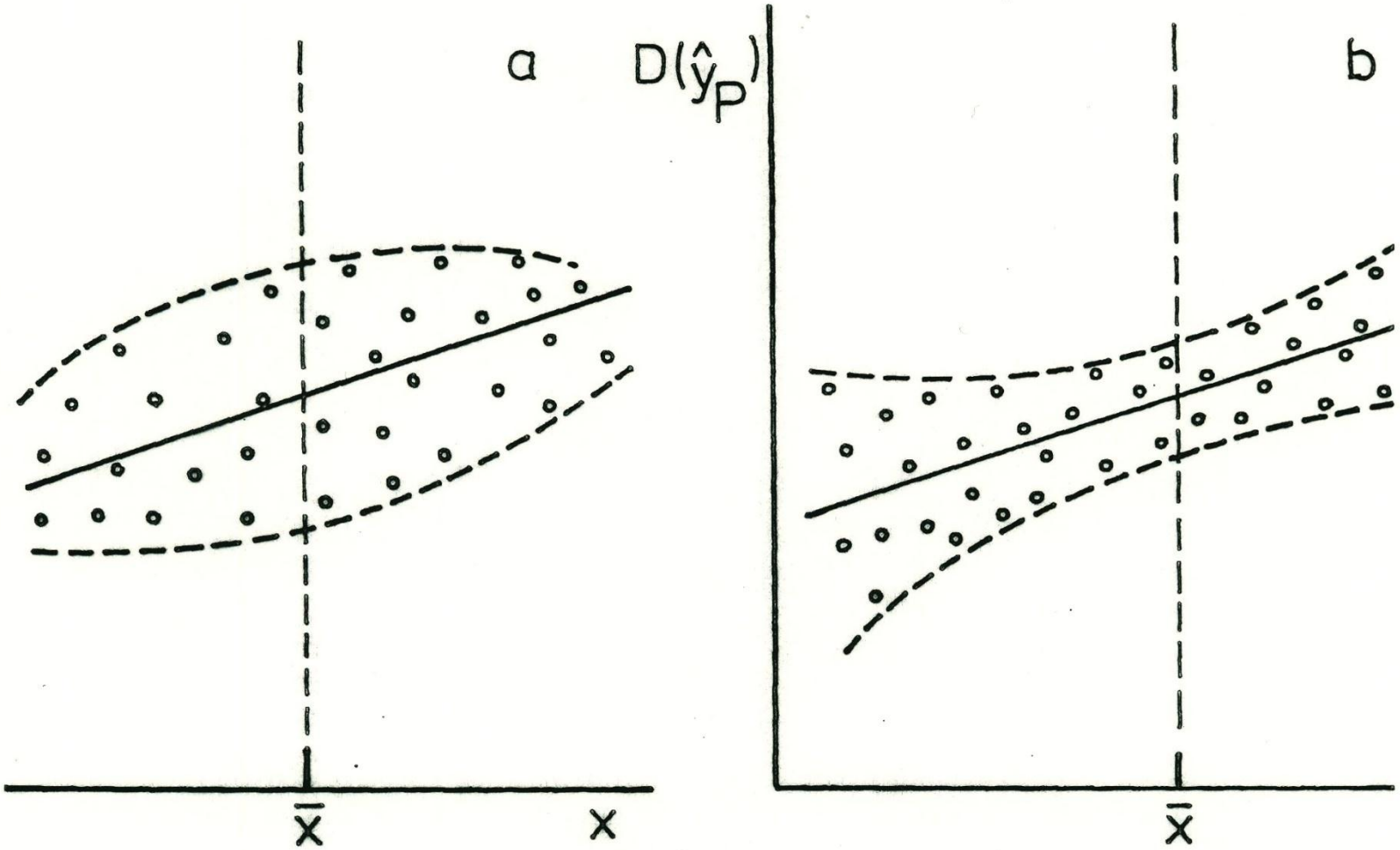
Příklad 6.4 *Odhady rozptylu predikce a reziduí pro model přímky*
Odvoďte vztahy pro výpočet odhadu rozptylu predikce $D(\hat{y}_{P,i})$ a rozptylu reziduí $D(\hat{e}_i)$ regresní přímky.

Řešení: Pro odhad rozptylu predikce platí vztah

$$D(\hat{y}_{P,i}) = \sigma^2 \left[\frac{1}{n} + \frac{n(x_i - \bar{x})^2}{D} \right]$$

pro odhad rozptylu reziduí

$$D(\hat{e}_i) = \sigma^2 \left[\frac{n-1}{n} - \frac{n(x_i - \bar{x})^2}{D} \right]$$



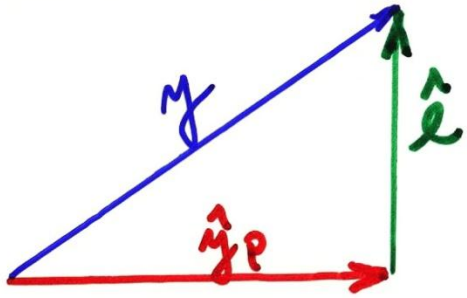
Závěr:

1. V místě $\bar{x}$ je rozptyl predikce minimální, zatímco rozptyl reziduí maximální.
2. Čím více se bod x_i vzdaluje od $\bar{x}$, tím je predikce méně přesná, zatímco odhad rezidua přesnější.

Rozklad celkového součtu čtverců

Čtverec délky vektoru **y** lze rozložit na součet čtverců délek vektoru $\hat{\mathbf{y}}_P$ a $\hat{\mathbf{e}}$,

$$\mathbf{y}^T \mathbf{y} = \mathbf{y}_P^T \mathbf{y}_P + \mathbf{e}^T \mathbf{e}$$



zkráceně

CS	=	TS	+	RSC
Celkový součet čtverců	=	Regresní součet čtverců	+	Reziduální součet čtverců
		Objasněná část		Neobjasněná část

Pro střední hodnotu součtu čtverců platí

$$E(TS) = E(\mathbf{y}^T \mathbf{H}\, \mathbf{y}) = \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\, \boldsymbol{\beta} + m\sigma^2$$

Místo veličin RSC a TS se lépe užívají průměrné hodnoty součtu čtverců:

1. Průměrná hodnota regresního součtu čtverců je

$$MTS = \frac{TS}{m}$$

a její **střední hodnota** je

$$E(MTS) = \sigma^2 + \frac{1}{m} \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\, \boldsymbol{\beta}$$

2. Průměrná hodnota reziduálního součtu čtverců je

$$\text{MRS} = \frac{\text{RSC}}{n - m}$$

a její **střední hodnota**

$$E(\text{MRS}) = \sigma^2$$

Test hypotézy: H_0: $\beta = 0$ *vs.* H_A: $\beta \neq 0$ a platí, že

(a) **TS je nezávislé na RSC,**

(b) $\text{TS} = \sigma^2 \chi^2(m)$,
kde $\chi^2(m)$ je náhodná veličina s rozdělením χ^2 s m stupni volnosti,

(c) $\text{RSC} = \sigma^2 \chi^2(n-m)$,
kde $\chi^2(n-m)$ je náhodná veličina s χ^2 rozdělením s (n-m) stupni volnosti.

Testování: Je-li $F = \frac{\text{MTS}}{\text{MRS}} < F_{1-\alpha}(m, n-m)$ je H_0 přijata a všechny koeficienty regresního modelu jsou nulové.

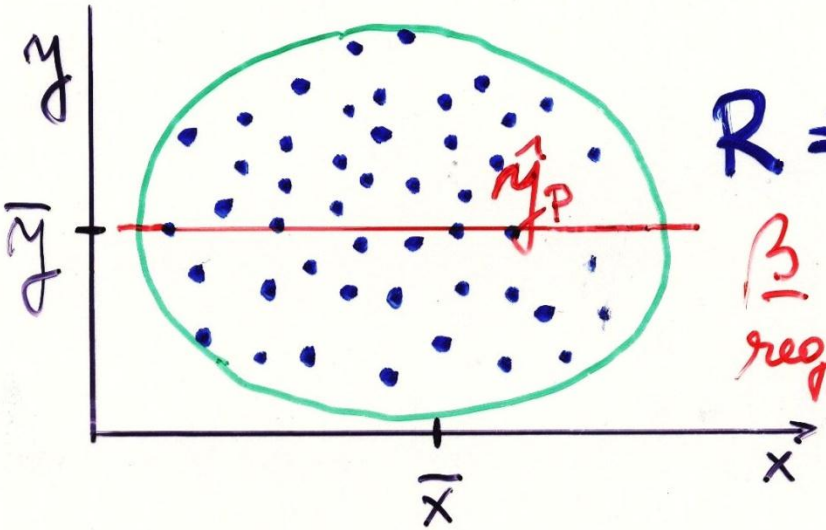
VYUŽITÍ VÍCENÁSOBNÉHO KORELAČNÍHO KOEFICIENTU R

$$R = \sqrt{\frac{TSC}{CSC}} = \sqrt{1 - \frac{RSC}{CSC}} = \cos\alpha$$

$$= \sqrt{1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

DISKUSE: (a) $\hat{R} \to 0$, pak platí $H_0: \hat{y}_p = \underline{1} \cdot \bar{y}$

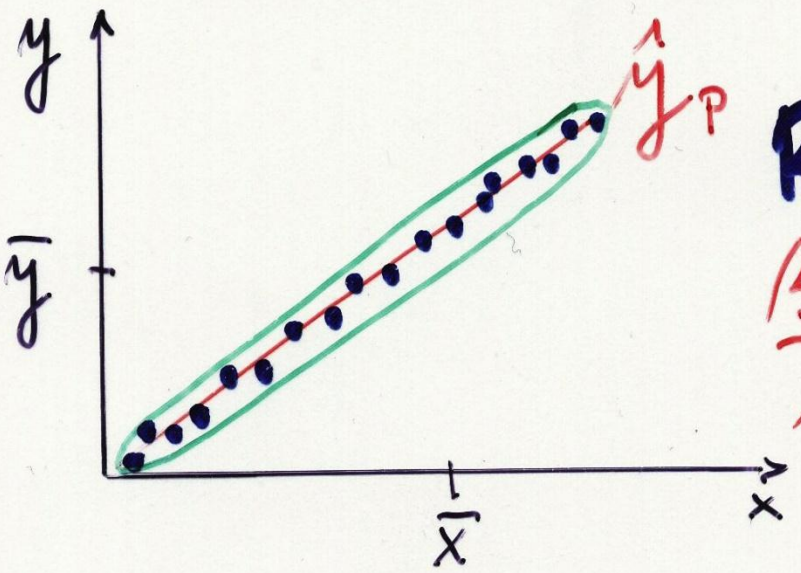
čili $H_0: \underline{\beta} = 0$



$R \doteq 0$

$\underline{\beta} = 0$, všechny regresní koeficienty jsou rovny nule

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



$R \doteq 1$

$\underset{\sim}{\beta} \neq 0$, všechny regresní koeficienty jsou různé od nuly

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

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

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

Diskuse 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:*** parametry ..... **nejsou 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 ($\times 100\%$):
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 6.8 *Omezení klasické analýzy lineárního modelu*

Anscomb[5] 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

$$b_1 = 0.5,\ b_2 = 3.0,\ D(b_1) = 0.0139 \text{ a } D(b_2) = 1.2656.$$

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

Statistika	P608a	P608b	P608c	P608d
Úsek, b_0, s_0				
Směrnice b_1, s_1				
Test významnosti úseku, t_0				
Test významnosti směrnice, t_1				
Test celkové regrese, F_R				
Korelační koeficient, R				
Koeficient determinace, D				
Směrodatná odchylka, $s(y)$				
Trend v reziduích				
Závěr: model je				

Řešení:

Lineární regresní model $E(y/x) = \beta_1 x + \beta_2$ má pro všechny výběry

a) ***Stejné*** **odhady parametrů:**

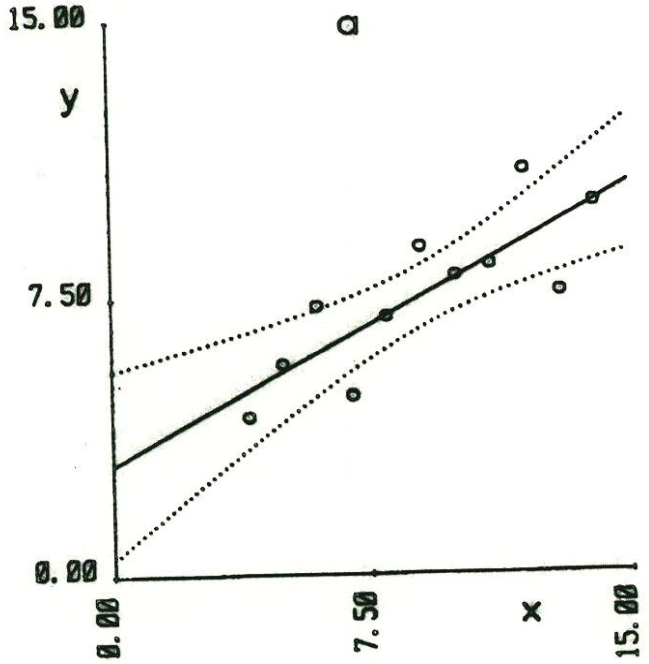
$b_1 = 0.5$, $b_2 = 3.0$, $D(b_1) = 0.0139$, $D(b_2) = 1.2656$.

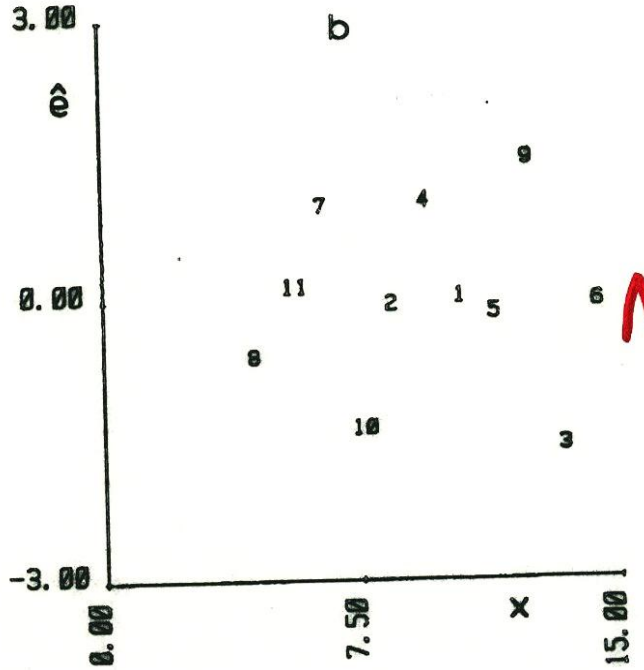
b) ***Stejné*** **testační statistiky významnosti parametrů:**

$T_1 = 2.667$ a $T_2 = 4.241$.

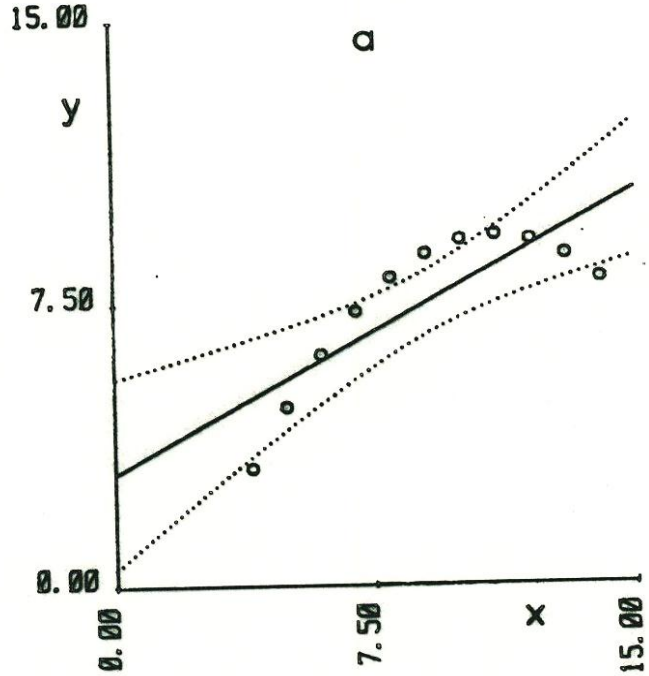
c) ***Stejné*** **testační statistiky:** $F_R = 17.97$, $\hat{R}^2 = 0.66$, $\hat{\sigma} = 1.237$ ukazují,

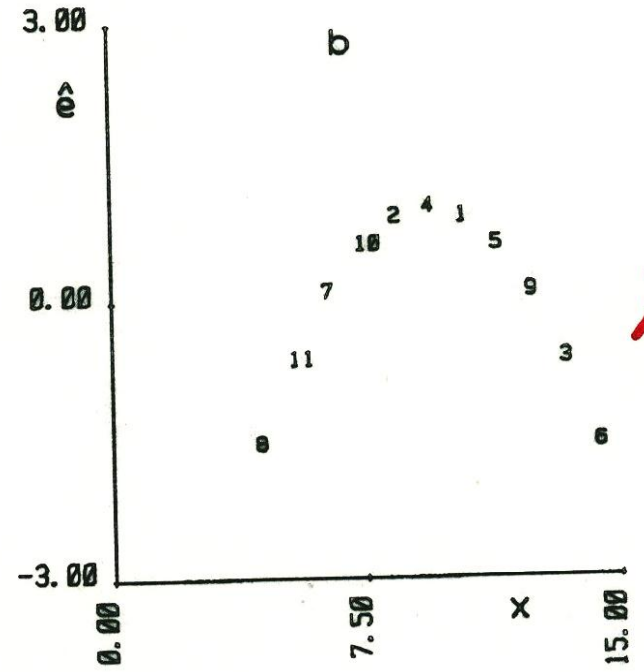
že β_1 a β_2 jsou významně odlišné od nuly.


a
15.00
y
7.50
0.00
0.00
7.50
x
15.00

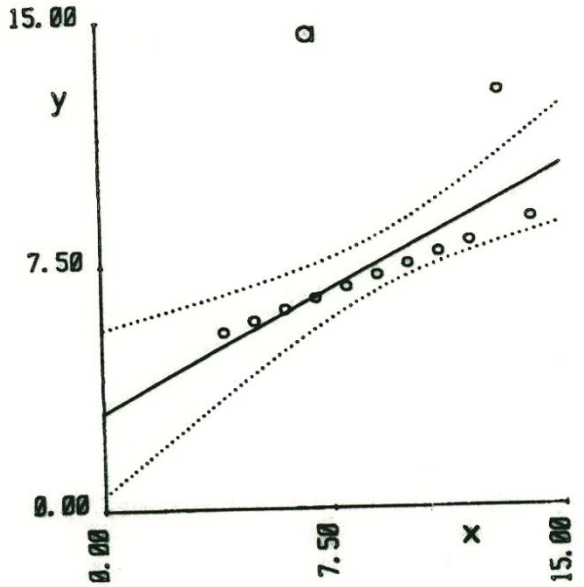

b
3.00
ê
0.00
-3.00
0.00
7.50
x
15.00
9
7
4
11
2
1
5
6
8
10
3

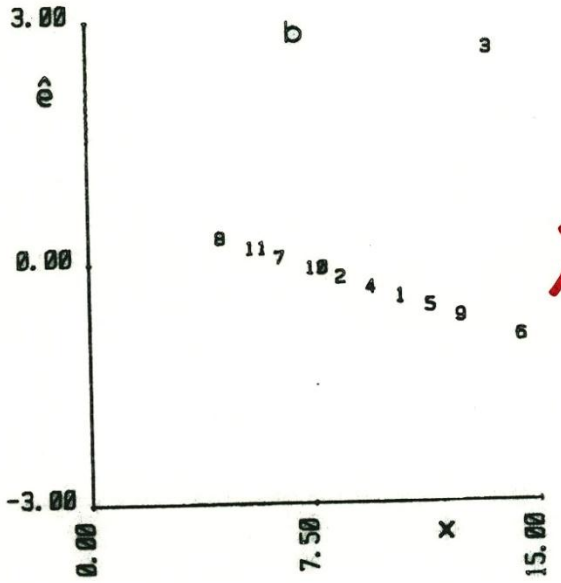
weak


a
15.00
y
7.50
0.00
0.00
7.50
x
15.00

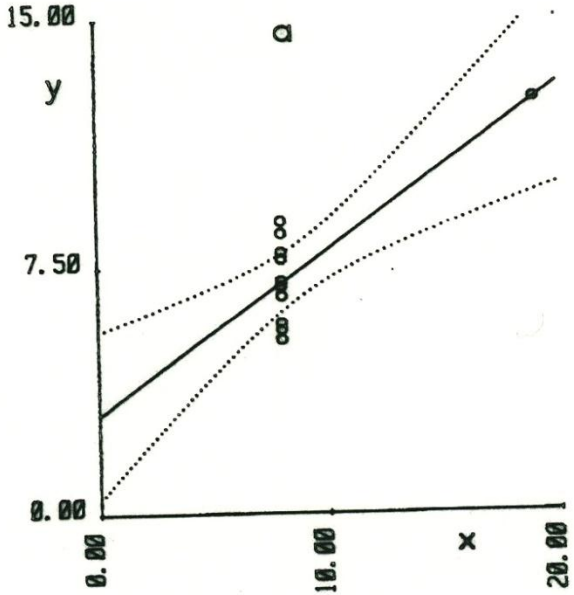

b
3.00
ê
0.00
-3.00
0.00
7.50
x
15.00
2
4
1
10
5
7
9
11
3
8
6

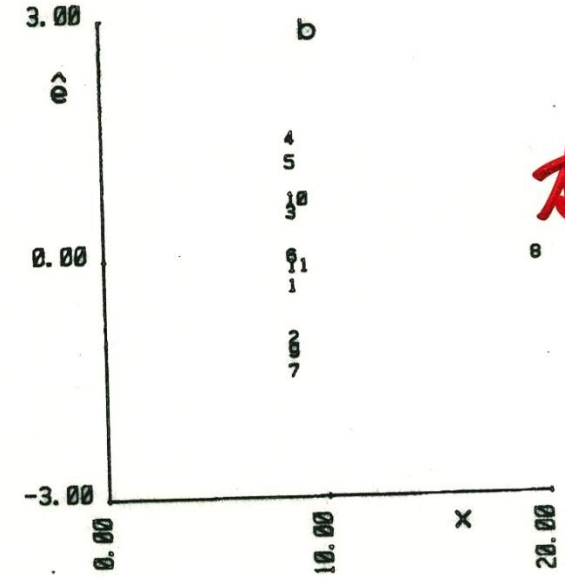
trend





trend





trend

Závěr: **Neshodu modelu s daty indikuje grafická analýza reziduí**

Konstrukce intervalů spolehlivosti

Bodové odhady parametrů β mají pro praxi menší význam.

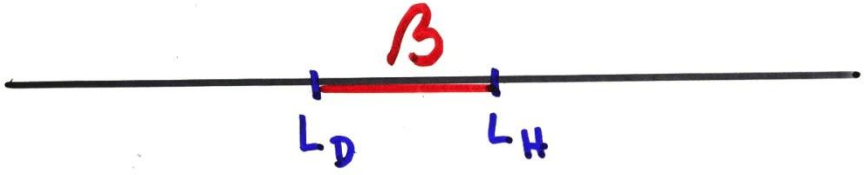
Intervaly spolehlivosti (konfidenční oblasti), ve kterých leží teoretická hodnota β se zvolenou pravděpodobností $(1 - \alpha)$.

1. 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

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

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

kde c_{jj} je j-tý diagonální prvek matice $(\mathbf{X}^T\,\mathbf{X})^{-1}$,
$t_{1-\alpha/2}(n-m)$ je kvantil Studentova rozdělení s $(n-m)$ stupni volnosti.

Pro korelované parametry jsou ***jednoduché i. s.*** příliš úzké.

Extrémní intervaly spolehlivosti jsou extrémy na elipsoidech spolehlivosti dle

$$b_j - p\,\sqrt{c_{jj}} \le \beta_j \le b_j + p\,\sqrt{c_{jj}}$$

Elipsoid spolehlivosti pro q regresních parametrů: jedná se o posledních q složek vektoru $\boldsymbol{\beta}$.

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

$$(\mathbf{b}_2 - \boldsymbol{\beta}_2)^T \mathbf{D}_2^{-1} (\mathbf{b}_2 - \boldsymbol{\beta}_2) = q\, \hat{\sigma}^2\, F_{1-\alpha}(q, n-m)$$

kde

- matice $\mathbf{D}_2$ rozměru $(q \times q)$ vznikla z matice $(\mathbf{X}^T \mathbf{X})^{-1}$ vynecháním prvních $(m - q)$ sloupů a $(m - q)$ řádků,

- symbol $\boldsymbol{\beta}_2$ značí vektor posledních q složek vektoru $\boldsymbol{\beta}$.

IS modelových hodnot přímky

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

Intervalové odhady parametrů

Pro různý počet pozorování se mohou odhadnuté regresní parametry b_0 a b_1 lišit.

Vedle bodových odhadů regresních parametrů lze vyčíslit i jejich *intervalové odhady*:

$$b_i - t_{1-\alpha/2}(n-m)\cdot s(b_i) < \beta_i < b_i + t_{1-\alpha/2}(n-m)\cdot s(b_i)$$

kde b_i je bodový odhad regresního parametru,

$t_{1-\alpha/2}(n-p)$ je kvantil Studentova t rozdělení,

m je počet parametrů modelu,

$s(b_i)$ je směrodatná chyba odhadu parametru.

Příklad: Český hydrometeorologický ústav v Praze měřil ...

95% interval spolehlivosti pro parametr b_0.

MS EXCEL: *Nástroje – Analýza Dat - Regrese*

	Parametry	***Směrodat. odchylka***	*t exp*	*Hodnota P*	***Dolní 95%***	***Horní 95%***
Hranice	**10,79504**	**0,446866**	24,15722	5,3E-08	**9,738368**	**11,85171**
Nadm. výška	-0,00541	0,00134	-4,03872	0,004941	-0,00858	-0,00224

$$b_i - t_{1-\alpha/2}(n-m)\cdot s(b_i) < \beta_i < b_i + t_{1-\alpha/2}(n-m)\cdot s(b_i)$$

Interval spolehlivosti lze vyčíslit ručně podle vzorce, nebo jej přímo přečíst z výstupu.

Příklad: Český hydrometeorologický ústav v Praze měřil ...

95% interval spolehlivosti pro oba parametry.

MS EXCEL: *Nástroje – Analýza Dat - Regrese*

	Koeficienty	*Chyba stř. hodnoty*	*t stat*	*Hodnota P*	***Dolní 95%***	***Horní 95%***
Hranice	10,79504	0,446866	24,15722	5,3E-08	**9,738368**	**11,85171**
Nadm. výška	-0,00541	0,00134	-4,03872	0,004941	**-0,00858**	**-0,00224**

Výklad úseku: V nadmořské výšce 0 metrů n.m. by se s pravděpodobností 95% měla průměrná teplota půdy nacházet v rozmezí 9,738 ^{0}C až 11,852 ^{0}C.

Výklad směrnice: S každým dalším metrem nadmořské výšky klesá s pravděpodobností 95% průměrná teplota půdy v rozmezí od 0,00858 ^{0}C do 0,00224 ^{0}C.

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

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

$$\mathbf{b} - t_{1-\alpha/2}(n-m)\,\hat{s}_{P,0} \leq \mathbf{x_0}^T \boldsymbol{\beta} \leq \mathbf{x_0}^T \mathbf{b} + t_{1-\alpha/2}(n-m)$$

kde $\hat{s}^2_{P,0}$ je rozptyl predikce

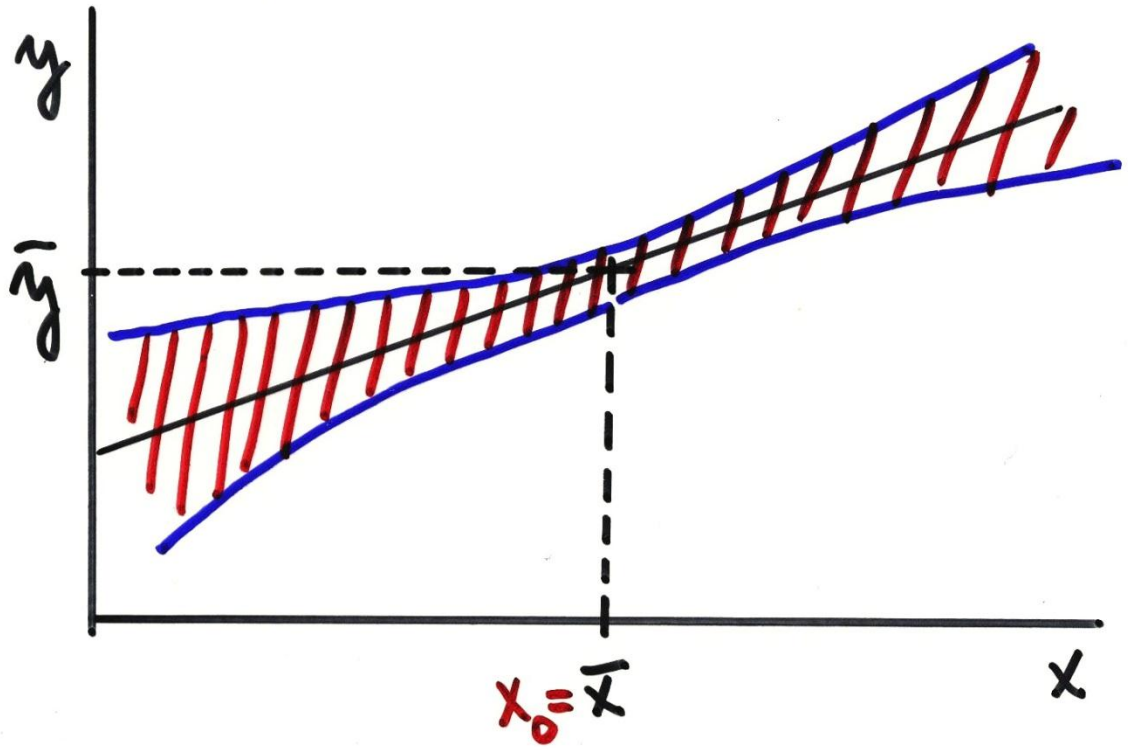
$$D(\hat{y}_{P,0}) = s^2_{P,0} = \hat{\sigma}^2\, \mathbf{x_0}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x_0}$$

100(1 - α)%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 $\mathbf{x}^T \boldsymbol{\beta}$ v oblasti pásů

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

a konfidenční pásy se nazývají ***pásy Workinga-Hottellinga***.



IS *y*-hodnot – Working-Hottelingův pás spolehlivosti

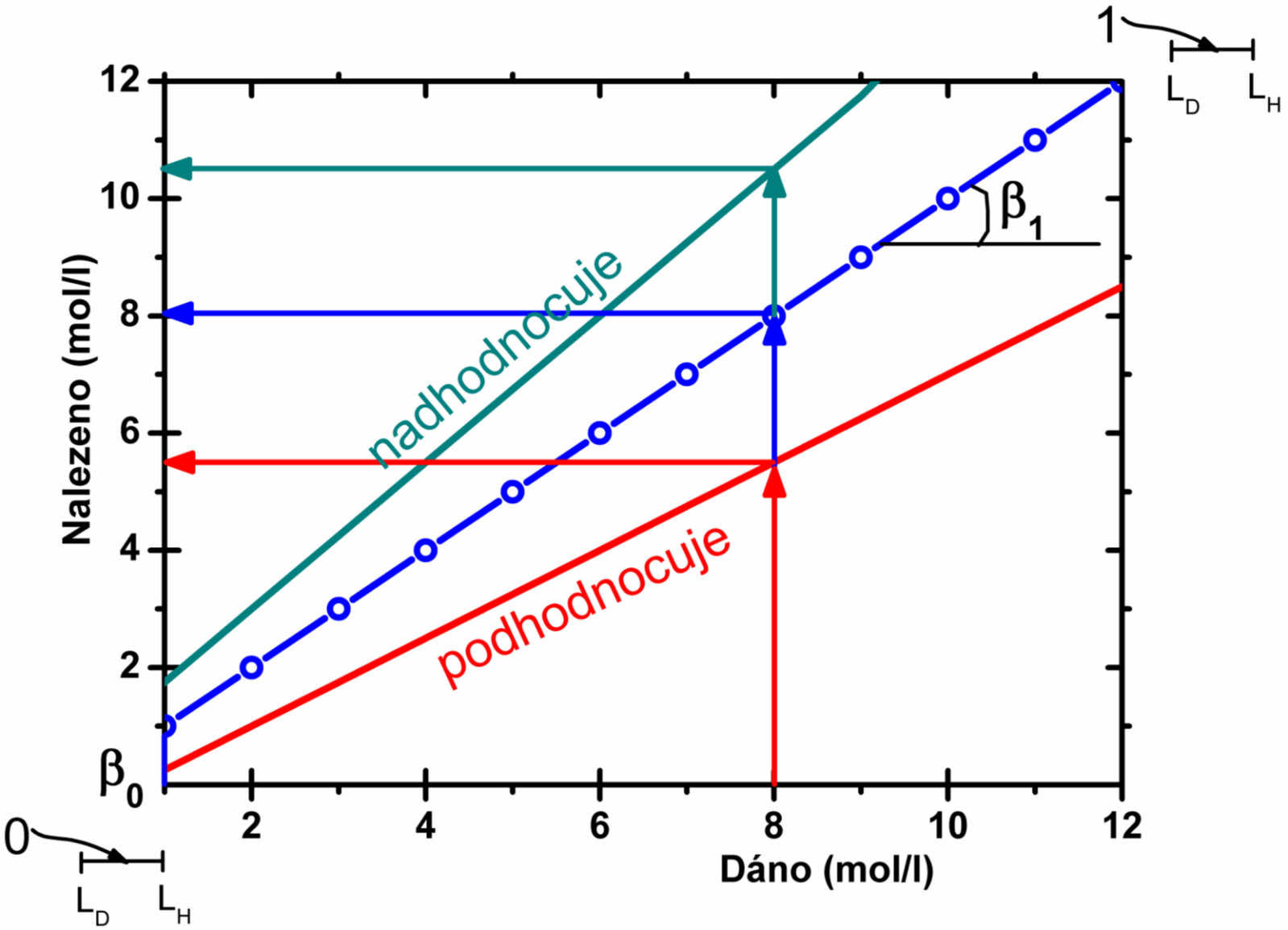
udává rozpětí, ve kterém se budou nacházet hodnoty závisle proměnné se zvolenou pravděpodobností

$1 - \alpha$

$$y_{i(min,max)} = y'_i \pm t_{\frac{\alpha}{2};n-m} \cdot \sigma$$

Validace nové analytické metody

$$y(\text{nalezeno}) = \beta_0 + \beta_1 \cdot x(\text{dáno})$$



Příklad 6.7 *Validizace nové analytické metody*

Proveďte validizaci nové analytické metody porovnáním 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

Obsah látky určený standardní metodou (x) a novou 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		

Řešení:

(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.* $H_A: \beta_2 \neq 0$

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

a dosazením

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

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.* $H_A: \beta_1 \neq 1$

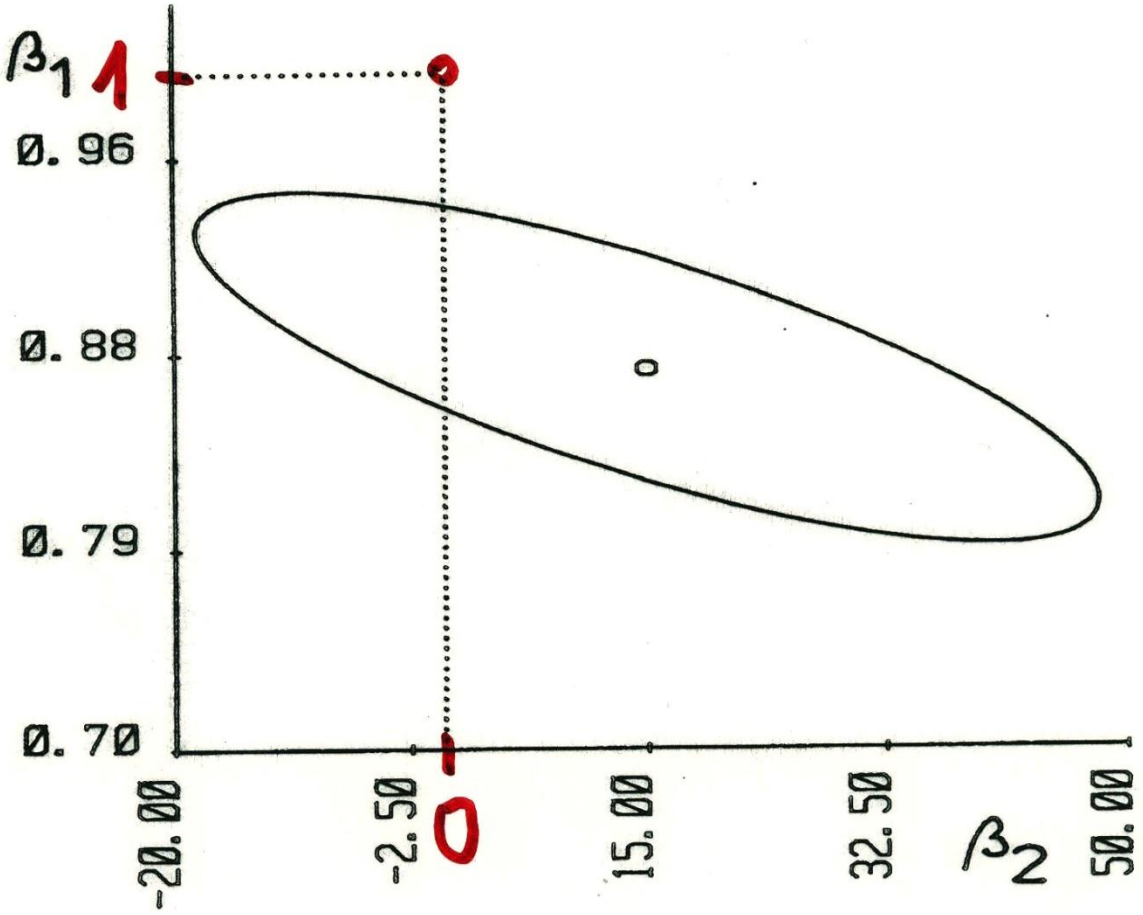
$$0.868 - 2.08 \times 0.0302 \leq \beta_1 \leq 0.868 + 2.08 \times 0.0302$$

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 $\beta_1 \neq 1$



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

(d) Interval spolehlivosti predikce v těžišti:

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

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

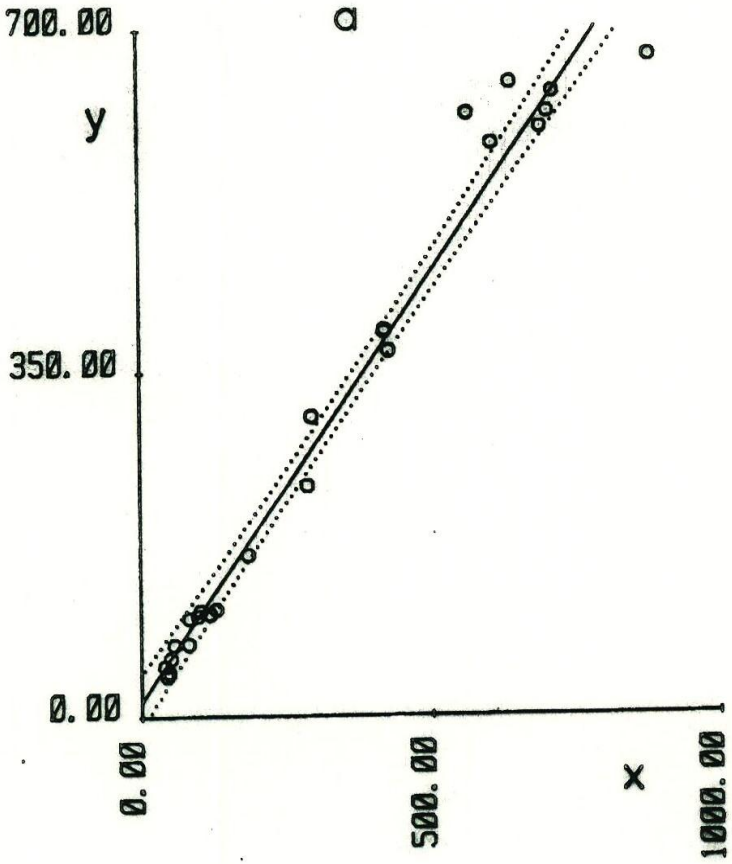
Dosazením

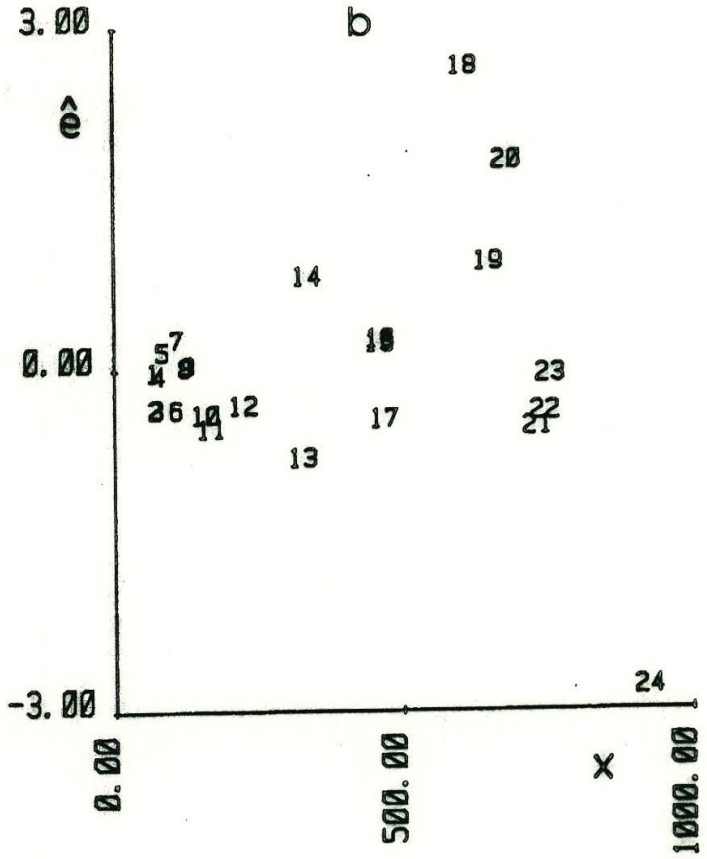
$$14.73 + 0.868 \times 320.7 - \frac{2.08 \times 39.54}{\sqrt{24}} \leq \eta_{\bar{x}} \leq$$

$$\leq 14.73 - 0.868 \times 320.7 + \frac{2.08 \times 39.54}{\sqrt{24}}$$

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

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





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

Nová analytická metoda vede k odlišným výsledkům od standardní.

6.2.1 Úlohy na validaci nové analytické metody

Úloha V6.01 *Validace stanovení molybdenu rentg.-fluoresc. metodou*

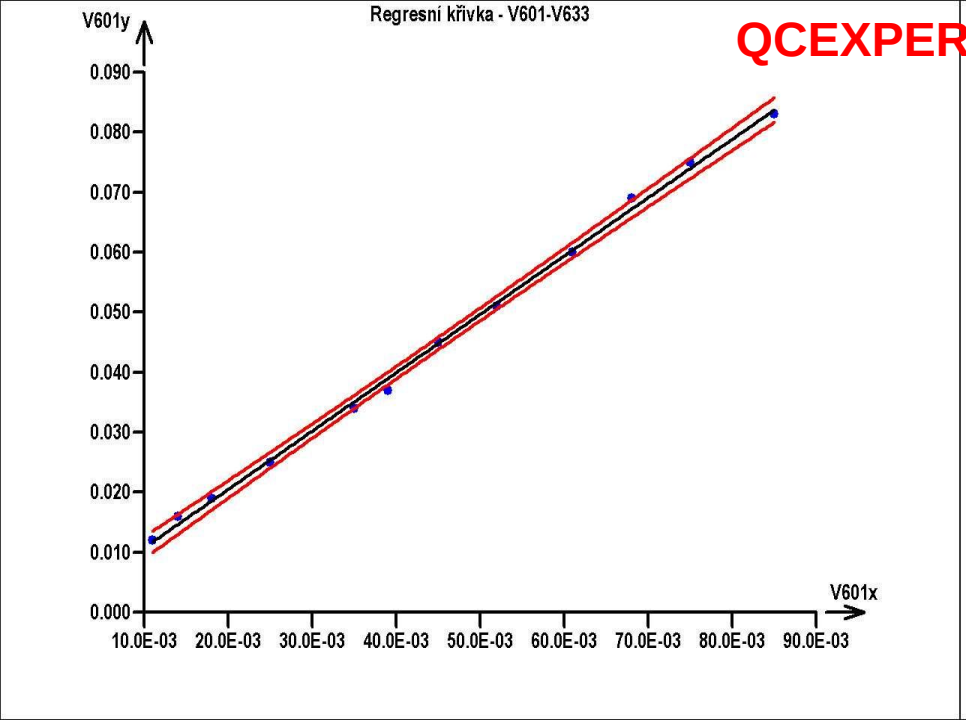
Zadání: U stanovení obsahu molybdenu porovnejte výsledky z rentg.-fluorescenční metody *y* s deklarovaným obsahem standardů ocelí *x*.

Úkoly:

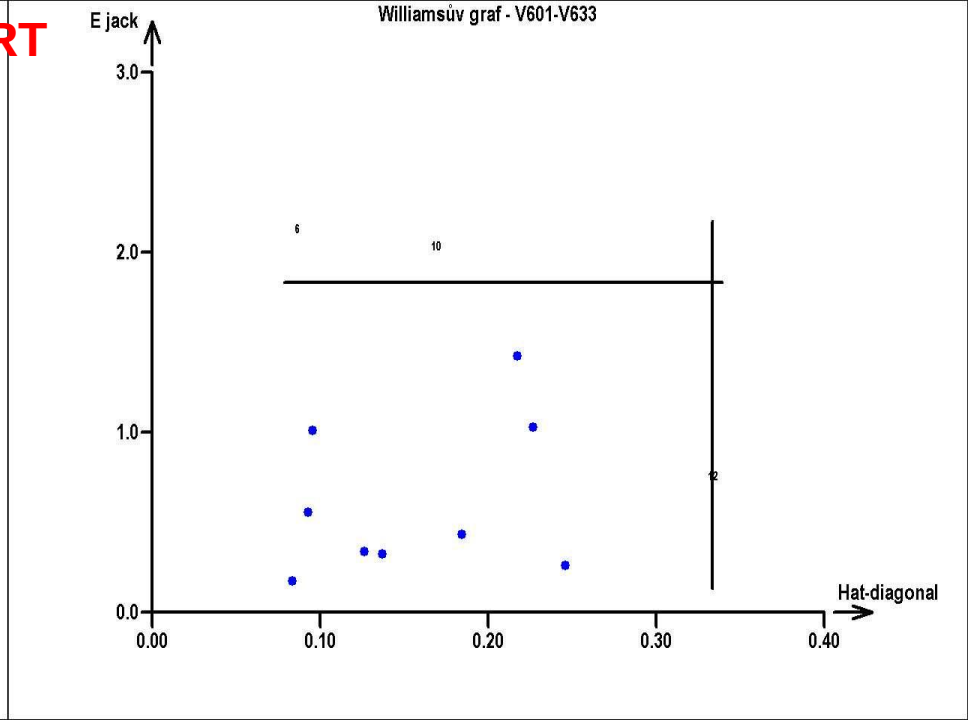
(1) Určete velikost systematické chyby metody (= velikost úseku β_0).
(2) Správnost metody (= směrnice měla být 1).
(3) Pokuste se vyjádřit i přesnost metody.
(4) Jsou v datech vlivné a vybočující body?
(5) Tabulkové indikace vlivných bodů a pět nejdůležitějších grafů identifikace vlivných bodů.

Data: Obsah molybdenu, dáno *x* [%], stanoveno *y* [%]:

Dáno *x*	Stanoveno *y*
0.011	0.012
...	...
0.085	0.083



QCEXPERT



Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	0.001034	0.000686	Nevýznamný	0.163	-0.00049559	0.0025644
V601x	0.972702	0.013748	Významný	7.77E-015	0.9420701358	1.003335592

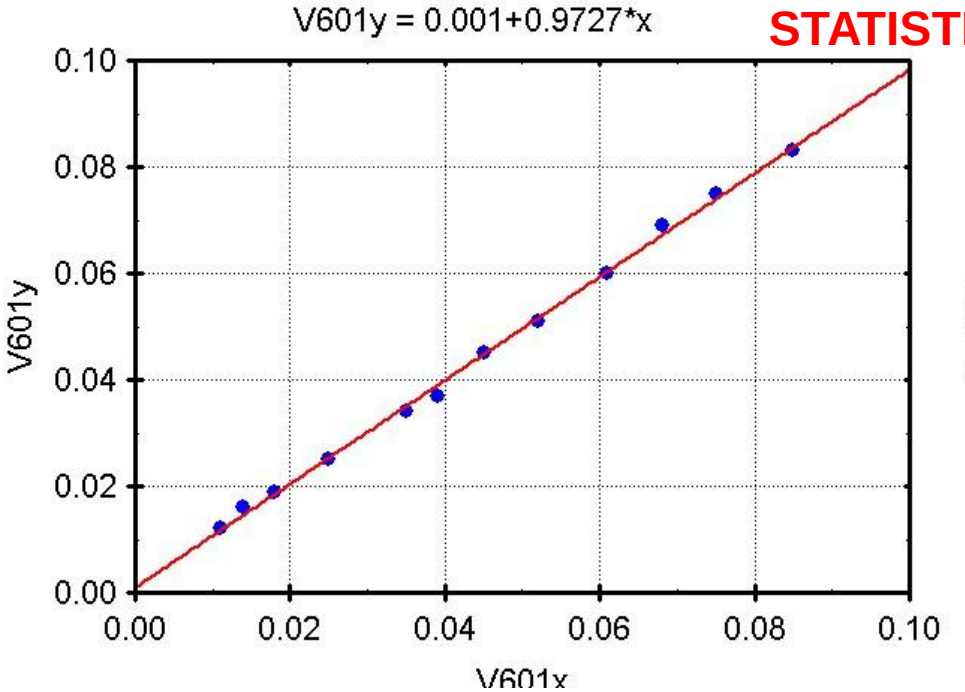
Statistické charakteristiky regrese

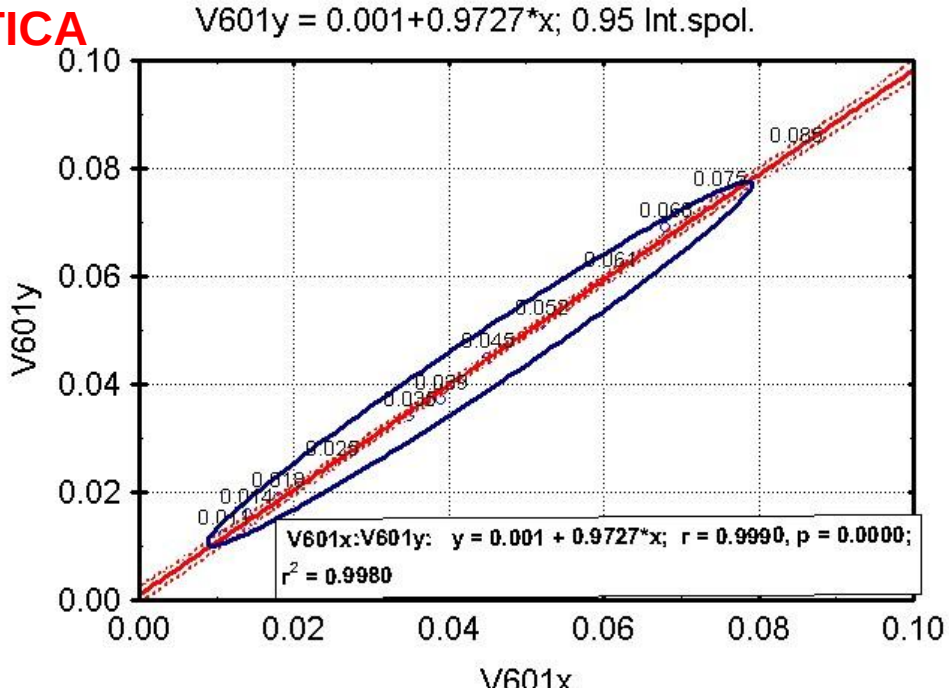
Vícenásobný korelační koeficient R :	0.99900
Koeficient determinace R^2 :	0.99800
Predikovaný korelační koeficient Rp :	0.99434
Střední kvdratická chyba predikce MEP :	1.50063E-006
Akaikeho informační kritérium :	-161.13

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	5005.80
Kvantil F (1-alfa, m-1, n-m) :	4.96460
Pravděpodobnost :	7.75E-015
Závěr :	Model je významný

STATISTICA





Efekt	Odhady parametrů (V6.sta) Sigma-omezená parametrizace							
	V601y Param.	V601y Sm.Ch.	V601y t	V601y p	-95.00% LmtSpol.	+95.00% LmtSpol.	V601y Beta (ß)	V601y Sm.Ch. ß
Abs. člen	0.001034	0.000687	1.50641	0.162882	-0.000496	0.002564		
"V601x"	*0.972703*	*0.013748*	*70.75168*	*0.000000*	*0.942070*	*1.003336*	*0.999003*	*0.014120*

Linear Regression Plot Section

Run Summary Section

Parameter	Value	Parameter		Value
Dependent Variable		V601y	Rows Processed	81
Independent Variable		V601x	Rows Used in Estimation	12
Frequency Variable		None	Rows with X Missing	69
Weight Variable		None	Rows with Freq Missing	0
Intercept	0.0010		Rows Prediction Only	0
Slope	0.9727		Sum of Frequencies	12
R-Squared	0.9980		Sum of Weights	12.0000
Correlation	0.9990		Coefficient of Variation	0.0257
Mean Square Error	1.267129E-06		Square Root of MSE	1.125668E-03

Summary Statement

The equation of the straight line relating V601y and V601x is estimated as: V601y = (0.0010) +(0.9727) V601x using the 12 observations in this dataset. The y-intercept, the estimated value of V601y when V601x is zero, is 0.0010 with a standard error of 0.0007. The slope, the estimated change in V601y per unit change in V601x, is 0.9727 with a standard error of 0.0137. The value of R-Squared, the proportion of the variation in V601y that can be accounted for by variation in V601x, is 0.9980. The correlation between V601y and V601x is 0.9990. A significance test that the slope is zero resulted in a t-value of 70.7517. The signifikance level of this t-test is 0.0000. Since 0.0000 < 0.0500, the hypothesis that the slope is zero is rejected. The estimated slope is 0.9727. The lower limit of the 95% confidence interval for the slope is 0.9421 and the upper limit is 1.0033. The estimated intercept is 0.0010. The lower limit of the 95% confidence interval for the intercept is -0.0005 and the upper limit is 0.0026.

Descriptive Statistics Section

Parameter	Dependent	Independent
Variable	V601y	V601x
Count	12	12
Mean	0.0438	0.0440
Standard Deviation	0.0240	0.0247
Minimum	0.0120	0.0110
Maximum	0.0830	0.0850

Regression Estimation Section

	Intercept	Slope
Parameter	**B(0)**	**B(1)**
Regression Coefficients	0.0010	0.9727
Lower 95% Confidence Limit	-0.0005	0.9421
Upper 95% Confidence Limit	0.0026	1.0033
Standard Error	0.0007	0.0137
Standardized Coefficient	0.0000	0.9990
T Value	1.5064	70.7517
Prob Level (T Test)	0.1629	0.0000
Reject H0 (Alpha = 0.0500)	No	Yes
Power (Alpha = 0.0500)	0.2759	1.0000
Regression of Y on X	0.0010	0.9727
Inverse Regression from X on Y	0.0009	0.9746
Orthogonal Regression of Y and X	0.0010	0.9736

Notes:

The above report shows the least-squares estimates of the intercept and slope followed by the corresponding standard errors, confidence intervals, and hypothesis tests. Note that these results are based on several assumptions that should be validated before they are used.

Estimated Model: (1.03440731901351E-03) + (.972702863961814) * (V601x)

Correlation and R-Squared Section

Parameter	Pearson Correlation Coefficient	R-Squared	Spearman Rank Correlation Coefficient
Estimated Value	0.9990	0.9980	1.0000
Lower 95% Conf. Limit (r dist'n)	0.9960		
Upper 95% Conf. Limit (r dist'n)	0.9995		
Lower 95% Conf. Limit (Fisher's z)	0.9963		1.0000
Upper 95% Conf. Limit (Fisher's z)	0.9997		1.0000
Adjusted (Rbar)	0.9978		
T-Value for H0: Rho = 0	70.7517	70.7517	
Prob Level for H0: Rho = 0	0.0000	0.0000	0.0000

Notes:

The confidence interval for the Pearson correlation assumes that X and Y follow the bivariate normal distribution. This is a different assumption from linear regression which assumes that X is fixed and Y is normally distributed. Two confidence intervals are given. The first is based on the exact distribution of Pearson's correlation. The second is based on Fisher's z transformation which approximates the exact distribution using the normal distribution. Why are both provided? Because most books only mention Fisher's approximate method, it will often be needed to do homework. However, the exact methods should be used whenever possible. The confidence limits can be used to test hypotheses about the correlation. To test the hypothesis that rho is a specific value, say r0, check to see if r0 is between the confidence limits. If it is, the null hypothesis that rho = r0 is not rejected. If r0 is outside the limits, the null hypothesis is rejected. Spearman's Rank correlation is calculated by replacing the orginal data with their ranks.

This correlation is used when some of the assumptions may be invalid.

Tests of Assumptions Section

Assumption/Test	Test Value	Prob Level	Is the Assumption Reasonable at the 0.2000 Level of Significance?
Residuals follow Normal Distribution?			
Shapiro Wilk	0.9853	0.996849	Yes
Anderson Darling	0.1507	0.962228	Yes
D'Agostino Skewness	0.0094	0.992478	Yes
D'Agostino Kurtosis	0.0319	0.974562	Yes
D'Agostino Omnibus	0.0011	0.999447	Yes
Constant Residual Variance?			
Modified Levene Test	0.1117	0.745133	Yes
Relationship is a Straight Line?			
Lack of Linear Fit F(0, 0) Test	0.0000	0.000000	No

No Serial Correlation?
Evaluate the Serial-Correlation report and the Durbin-Watson test if you have equal-spaced, time series data.

Notes:
A 'Yes' means there is not enough evidence to make this assumption seem unreasonable. This lack of evidence may be because the sample size is too small, the assumptions of the test itself are not met, or the assumption is valid. A 'No' means the that the assumption is not reasonable. However, since these tests are related to sample size, you should assess the role of sample size in the tests by also evaluating the appropriate plots and graphs. A large dataset (say N > 500) will often fail at least one of the normality tests because it is hard to find a large dataset that is perfectly normal.

Normality and Constant Residual Variance:
Possible remedies for the failure of these assumptions include using a transformation of Y such as the log or square root, correcting data-recording errors found by looking into outliers, adding additional independent variables, using robust regression, or using bootstrap methods.

Straight-Line: Possible remedies for the failure of this assumption include using nonlinear regression or polynomial regression.

Úloha V6.02 *Bichromátometrická metoda stanovení železitých iontů*

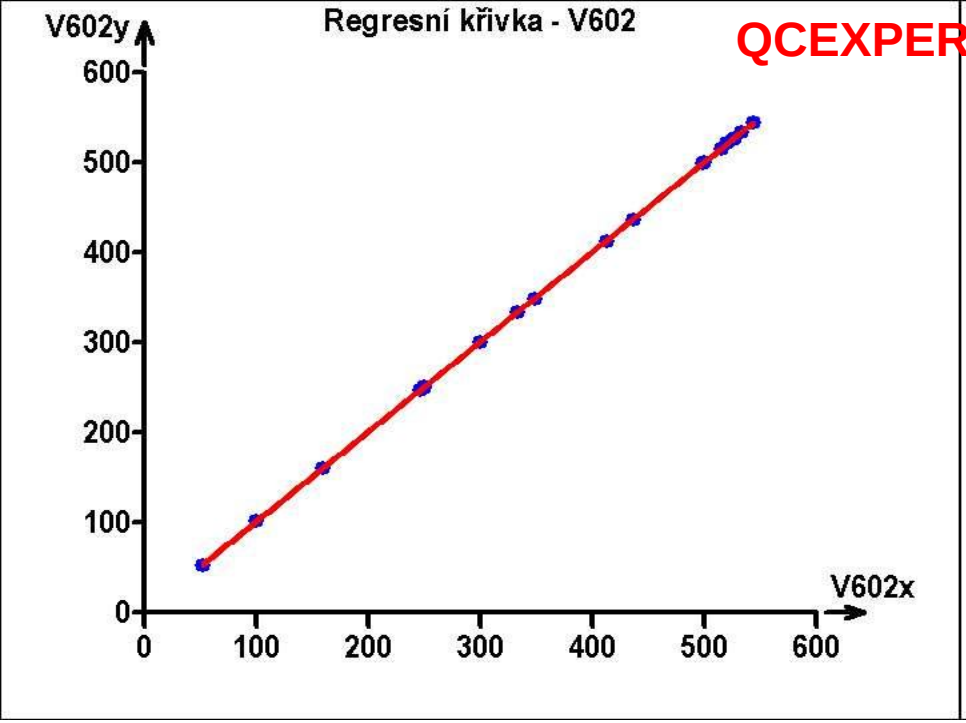
Zadání: Kraft a Dosch[60] navrhli titrační stanovení železa ve vodách. Železité ionty Fe^{3+} v Fe_2O_3 se redukují titanitou solí v přebytku a vzniklé ionty Fe^{2+} se pak stanoví bichromátometricky.

Úkoly:

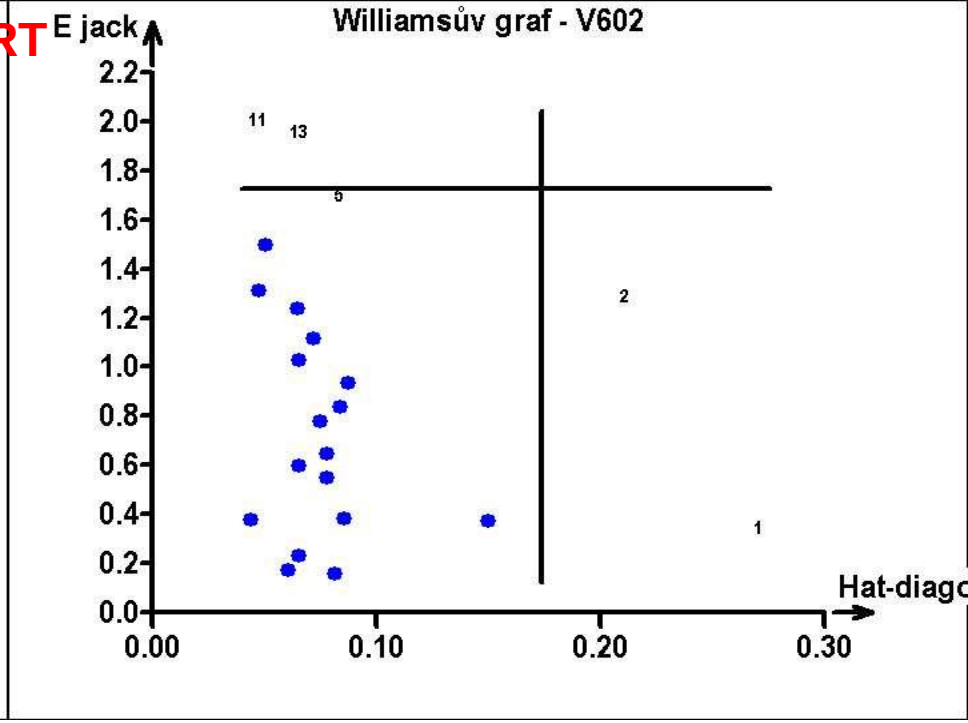
(1) Vede titrační stanovení ke správným výsledkům?
(2) Proveďte Studentův *t*-test významnosti úseku b_0 (má být $\beta_0 = 0$).
(3) Proveďte Studentův *t*-test jednotkové směrnice b_1 (má být $\beta_1 = 1$).
(4) Proveďte kombinovaný test obou parametrů v modelu přímky.
(5) Popište test významnosti absolutního členu.
(6) Popište test vhodnosti lineárního modelu dle Uttsové.

Data: Obsah Fe_2O_3 [mg], dáno *x*, nalezeno *y*:

Dáno *x*	Stanoveno *y*
52.0	52.50
...	...
543.61	543.78



QCEXPERT



Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	0.70845	0.23872	Významný	0.007343	0.21200	1.20490
V602x	0.99834	0.00056	Významný	0	0.99716	0.99951

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0.99999
Koeficient determinace R^2 :	0.99999
Predikovaný korelační koeficient Rp :	0.99998
Střední kvdratická chyba predikce MEP :	0.17699
Akaikeho informační kritérium :	-39.545

Úloha V6.04 *Stanovení kyseliny ftalové tenkovrstvou chromatografií*

Zadání: Obsah kyseliny ftalové byl stanoven tenkovrstvou chromatografií a chromatogram byl vyhodnocován remisním fotometrem.

Úkoly:

(1) Stanovte oba parametry lineárního regresního modelu a vyšetřete, zda je úsek nulový a směrnice jednotková.

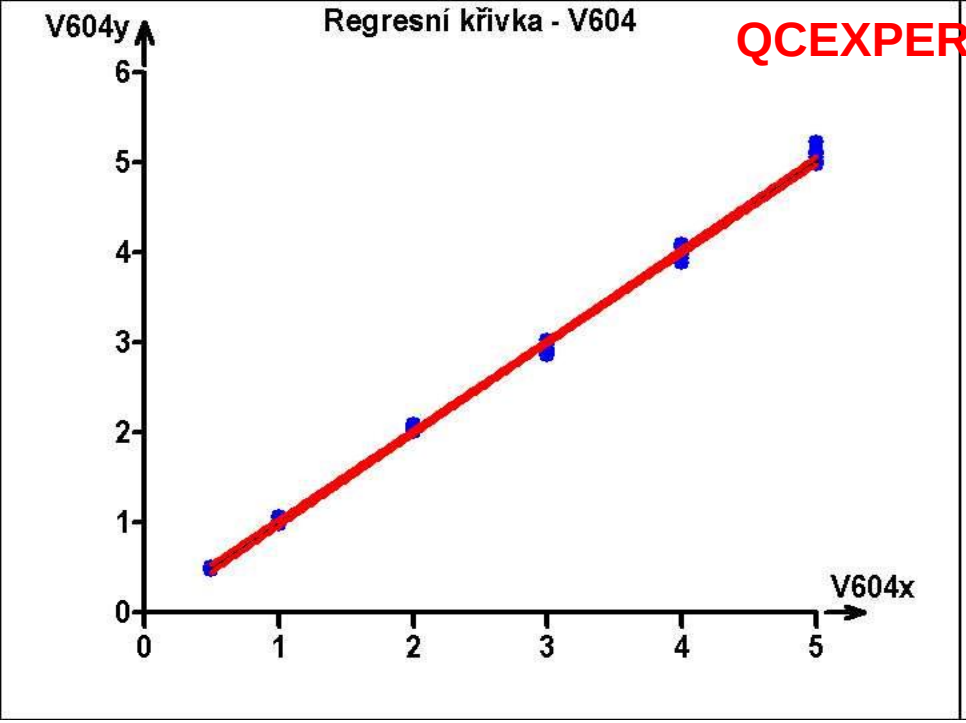
(2) Vyšetřete, zda jsou v datech vybočující hodnoty?

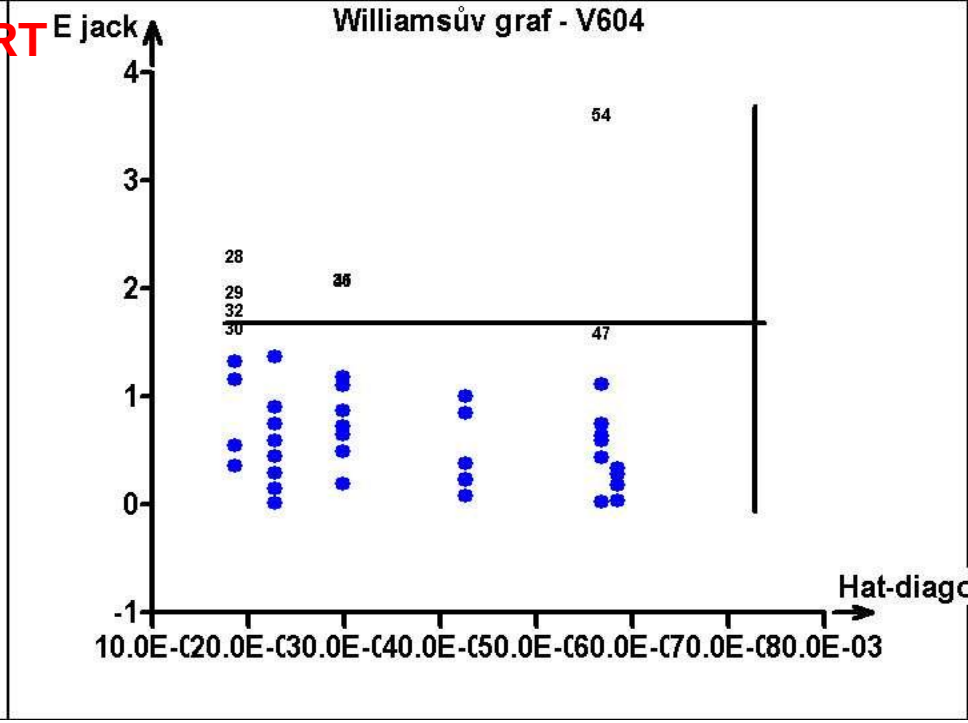
(3) Je stanovení je správné?

(4) Jaký je nutno zvolit postup při porušení předpokladů MNČ?

Data: Obsah kyseliny ftalové [μg], dáno *x*, nalezeno *y* (opakovaně).

Dáno *x*	Stanoveno *y*
0.50	0.48
...	...
5.23	65.02





Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-0.0110	0.0187	Nevýznamný	0.5565	-0.048499	0.026400
V604x	1.00588	0.0059	Významný	0	0.9940574	1.017716

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R : 0.9990902542
Koeficient determinace R^2 : 0.9981813361
Predikovaný korelační koeficient Rp : 0.9961043053
Střední kvdratická chyba predikce MEP : 0.004614241291
Akaikeho informační kritérium : -295.6513242

Úloha V6.06 *Ověření stanovení železa spektrofotometrickou metodou*

Zadání: Ověřte stanovení obsahu železa y v $CoSO_4$ spektrofotometricky SFM *y* porovnáním výsledků standardního stanovení obsahu *x* metodou AAS, u které je předpokládána zanedbatelná náhodná chyba.

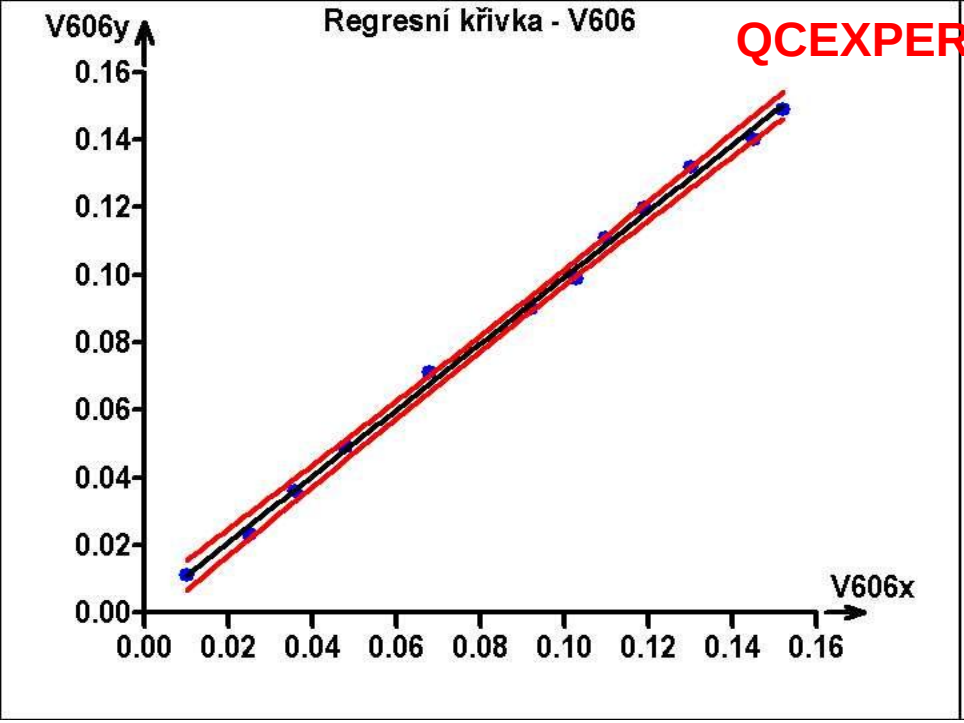
Úkoly:

(1) Vedou obě metody ke shodným výsledkům?

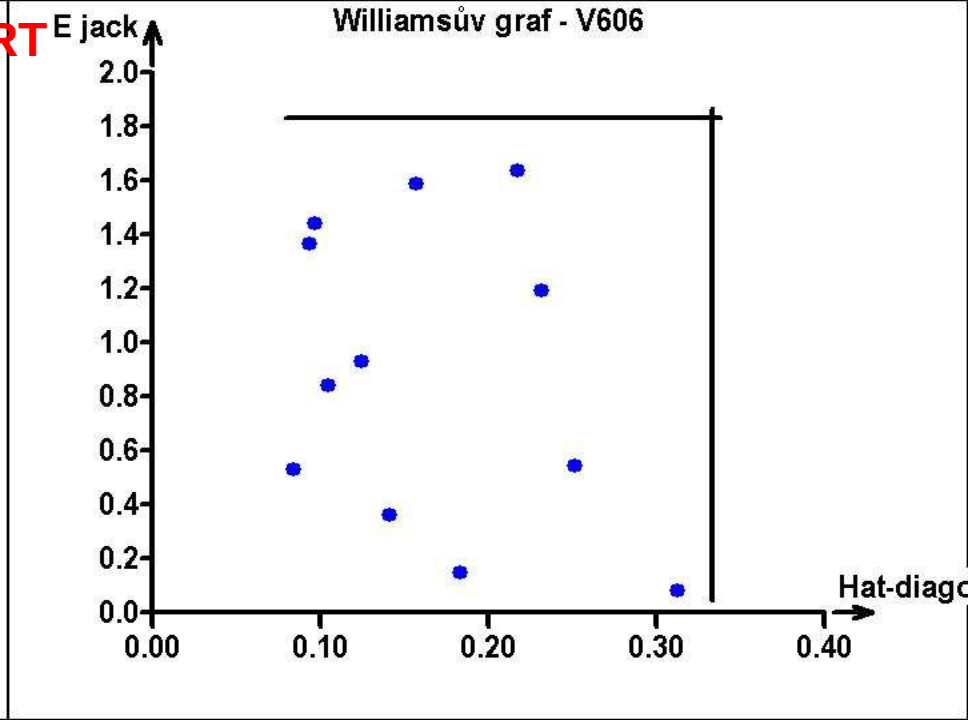
(2) Jsou v datech odlehlé hodnoty? Užijte pět grafů indikace vlivných bodů.

Data: Obsah železa v $CoSO_4$ [%], když je AAS *x* [%], SFM *y* [%]:

Dáno *x*	Stanoveno *y*
0.010	0.011
...	...
0.152	0.149



QCEXPERT



Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	0.001011	0.001531	Nevýznamný	0.5238	-0.0023997	0.0044219
V606x	0.981567	0.015615	Významný	2.5313E-014	0.9467727	1.0163614

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R : 0.9987368986
Koeficient determinace R^2 : 0.9974753927
Predikovaný korelační koeficient Rp : 0.9929318535
Střední kvdratická chyba predikce MEP : 7.273899498E-006
Akaikeho informační kritérium : -142.032264

Úloha V6.07 *Ověření stanovení dusičnanů v pitné a povrchové vodě*

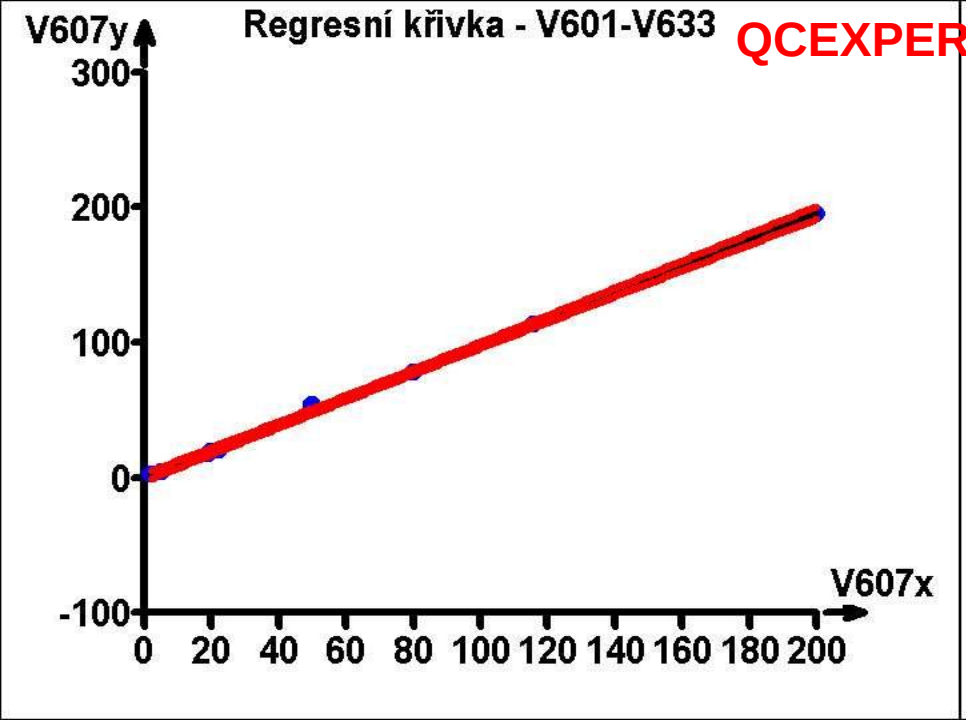
Zadání: V chemických laboratořích geochemické firmy se zavedla nová metoda stanovení obsahu dusičnanů y v pitných ale také povrchových vodách pomocí iontově párové chromatografie.

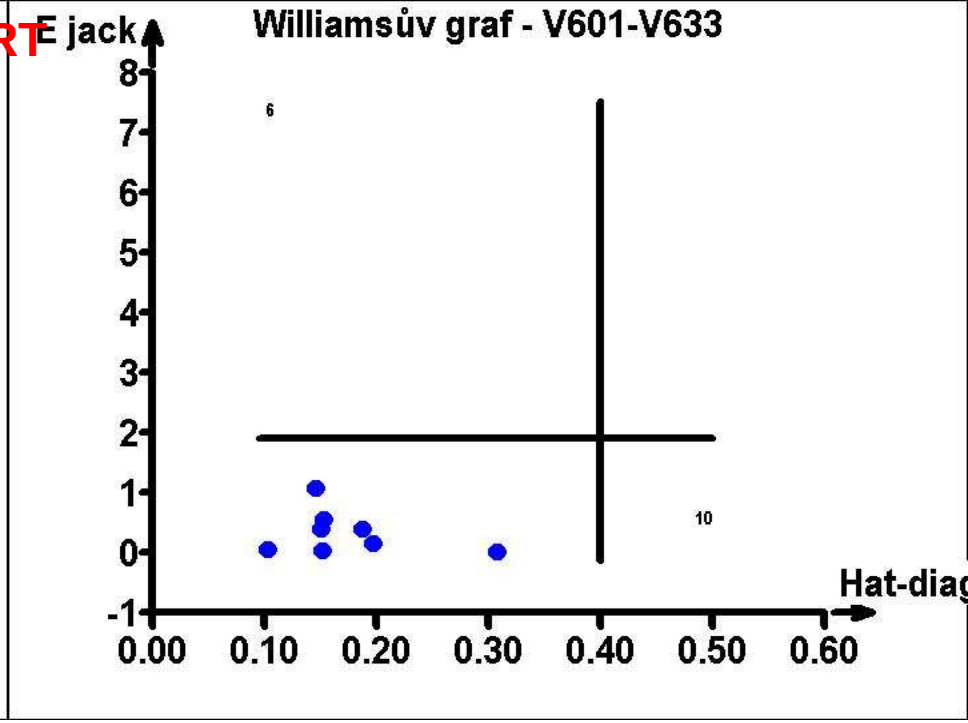
Úkoly:

(1) Validujte novou metodu vůči deklarovaným obsahům NO_3^- [mg/l] *x*.
(2) Odhadněte regresní parametry metodu ortogonální regrese.
(3) Vede nová metoda ke správným výsledkům?
(4) Proveďte simultánní test významnosti úseku a významnosti směrnice, zda je rovna jedné.

Data: Pro obsah dusičnanů NO_3^- [mg/l] je dáno *x*, nalezeno *y*.

Dáno *x*	Stanoveno *y*
2.10	2.20
...	...
200.00	195.00





Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-0.14284	0.9389	Nevýznamný	0.88284	-2.3079	2.0222
V607x	0.9801479567	0.0098	Významný	1.179E-013	0.95739	1.0029

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0.9995948631
Koeficient determinace R^2 :	0.9991898903
Predikovaný korelační koeficient Rp :	0.9978300448
Střední kvdratická chyba predikce MEP :	4.643796689
Akaikeho informační kritérium :	16.4284433

Úloha V6.20 *Validace nové metody stanovení arsenu v odpadní vodě*

Zadání: Je třeba validovat nové jednodušší stanovení arsenu v odpadní vodě. Mezi naměřenou koncentrací arsenu y a známou koncentrací x v μg/ml je předpokládán lineární regresní model $y = \beta_0 + \beta_1 x$.

Úkoly:

(1) Užitím ortogonální regrese ověřte správnost nové metody.

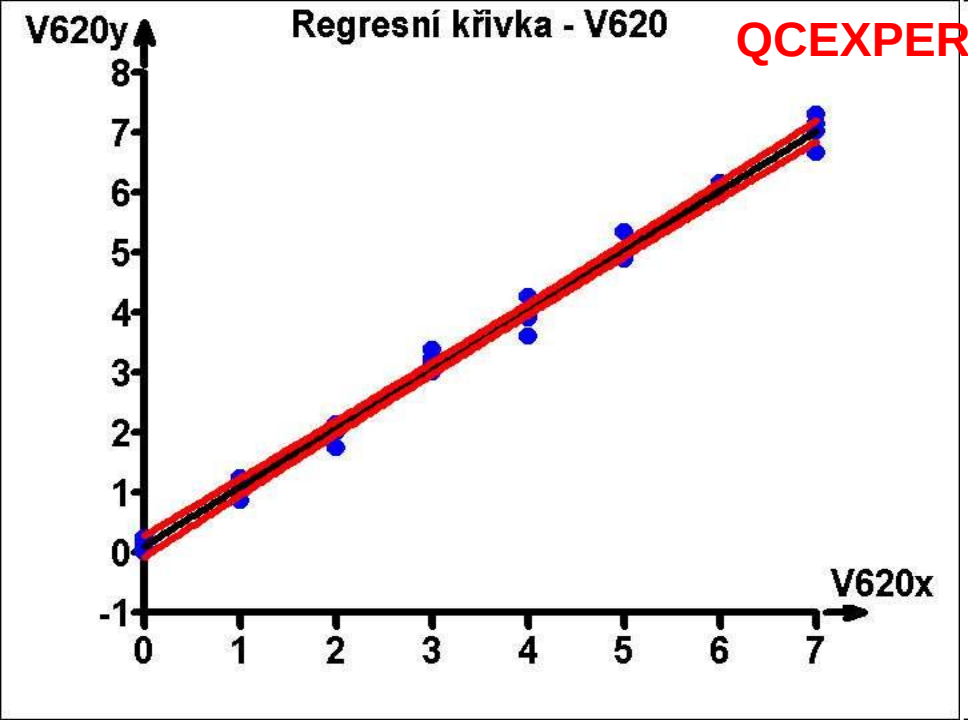
(2) K jakému výsledku dospěje nová metoda, když standard arsen vůbec neobsahuje čili absolutní člen je nulový, $\beta_0 = 0$?

(3) Vyšetřete, zda nová metoda nadhodnocuje či podhodnocuje?

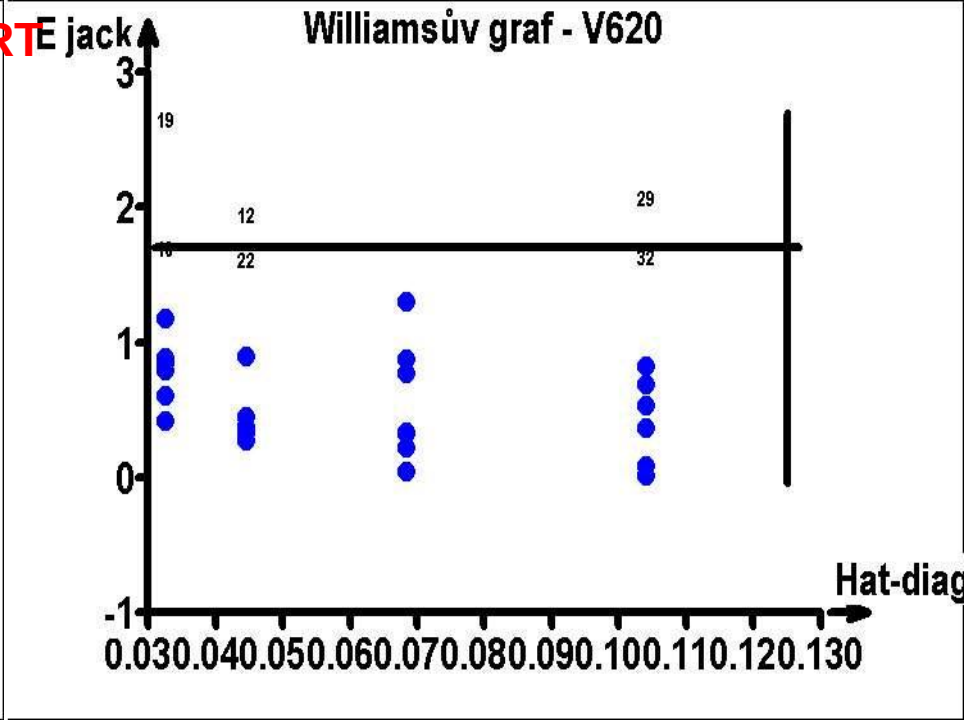
(4) Jakou modifikaci MNČ je třeba použít, když jsou všechny proměnné zatíženy náhodnými chybami?

Data: Koncentrace arsenu daná x [μg. cm^{-3}], nalezená y [μg. cm^{-3}].

Dáno x	Stanoveno y
0	0.17
...	...
7.0	7.30



QCEXPERT



Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	0.10458	0.06051	Nevýznamný	0.0942	-0.01899	0.228167
V620x	0.98770	0.01446	Významný	0	0.958168	1.017248

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0.99679
Koeficient determinace R^2 :	0.99360
Predikovaný korelační koeficient Rp :	0.98563
Střední kvdratická chyba predikce MEP :	0.03715
Akaikeho informační kritérium :	-105.20

Úloha V6.22 *Validace navržené titrační metody ke stanovení modré báze MB H-3R*

Zadání: Při výrobě modré báze MB H-3R byl stanovován její obsah v pastě z kalolisu titračně dusitanem v kyselém prostředí *y* a standardně spektrofotometricky *x*. Za základ byla vzata titrační metoda. Rozptyl této metody se považuje za zanedbatelný vůči rozptylu spektrofotometrické metody.

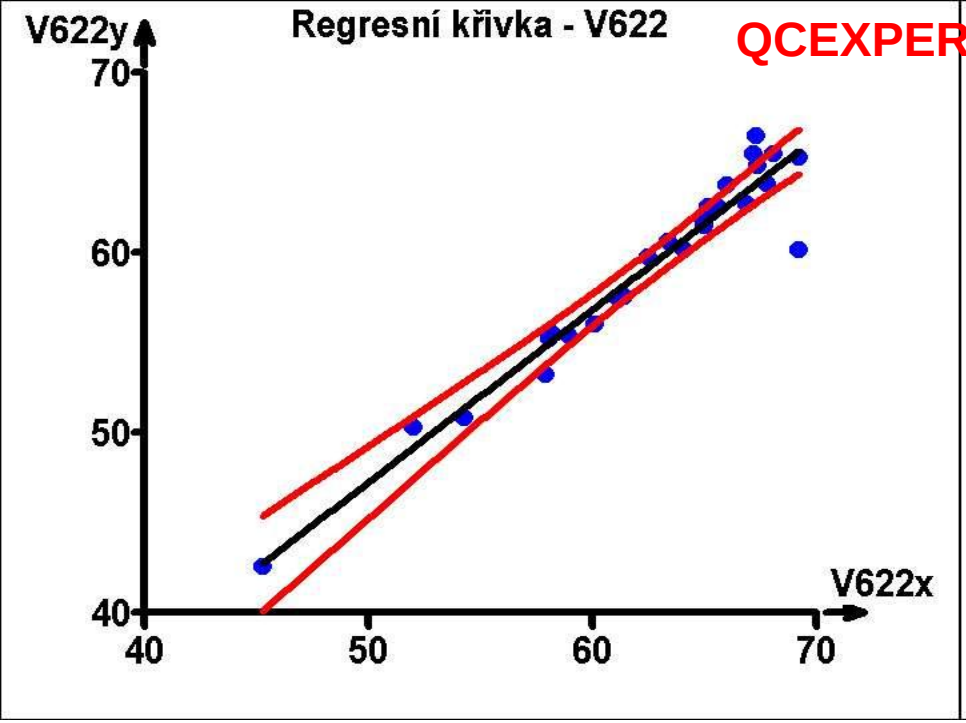
Úkoly:

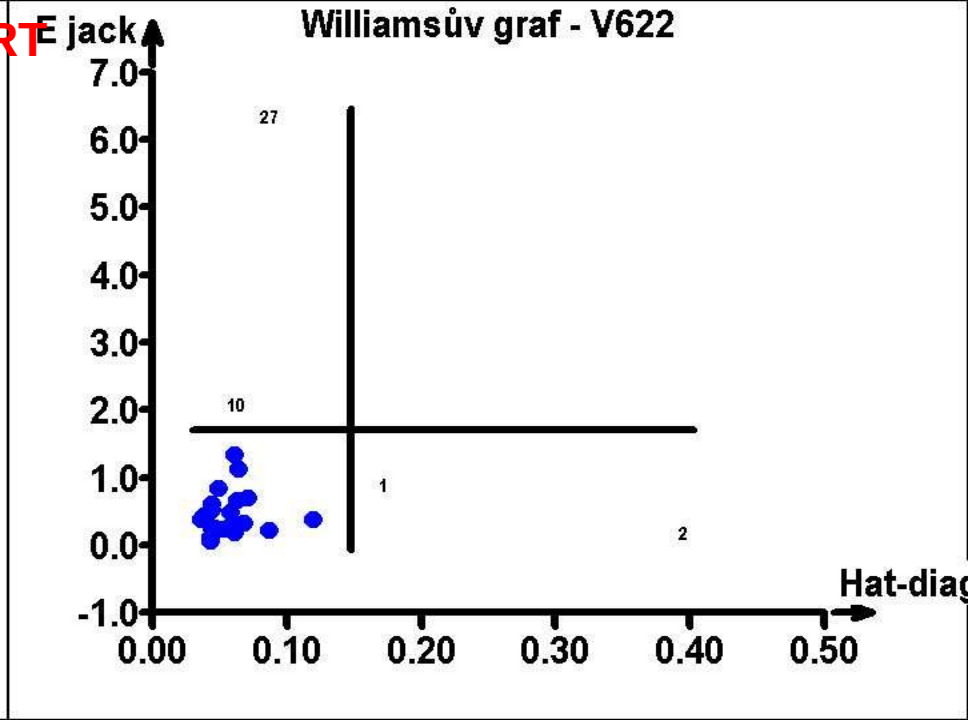
(1) Popište test významnosti absolutního členu.

(2) Vysvětlete test shodnosti odhadu parametru β s předepsanou β_0.

Data: Koncentrace modré báze spektrofotometrickou metodou *x* a titrační metodou *y*.

Dáno *x*	Stanoveno *y*
52.0	50.3
...	...
69.2	60.1





Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-0.65650	3.14318	Nevýznamný	0.83624	-7.1300	5.8170
V622x	0.957240	0.04994	Významný	2.220E-016	0.85438	1.0600

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R : 0.967619281
Koeficient determinace R^2 : 0.936287073
Predikovaný korelační koeficient Rp : 0.8551935541
Střední kvdratická chyba predikce MEP : 2.300430311
Akaikeho informační kritérium : 22.00604283

Úloha V6.31 *Validace stanovení chromu metodou AAS a ICP-AES*

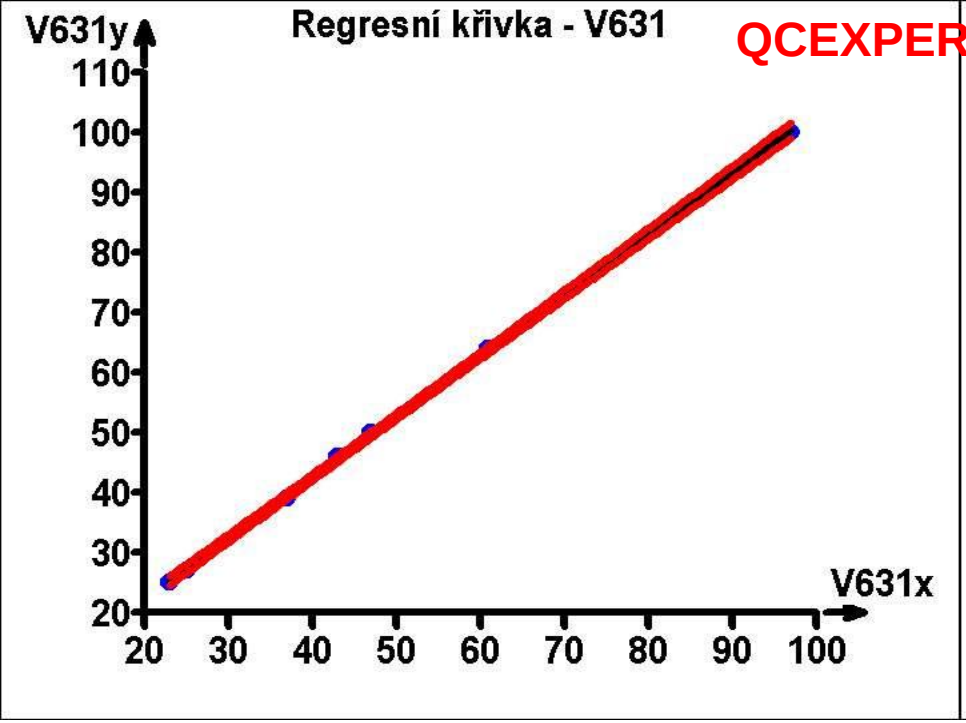
Zadání: Ve vzorcích půdy byl stanoven metodami AAS a ICP-AES obsah chromu.

Úkoly:

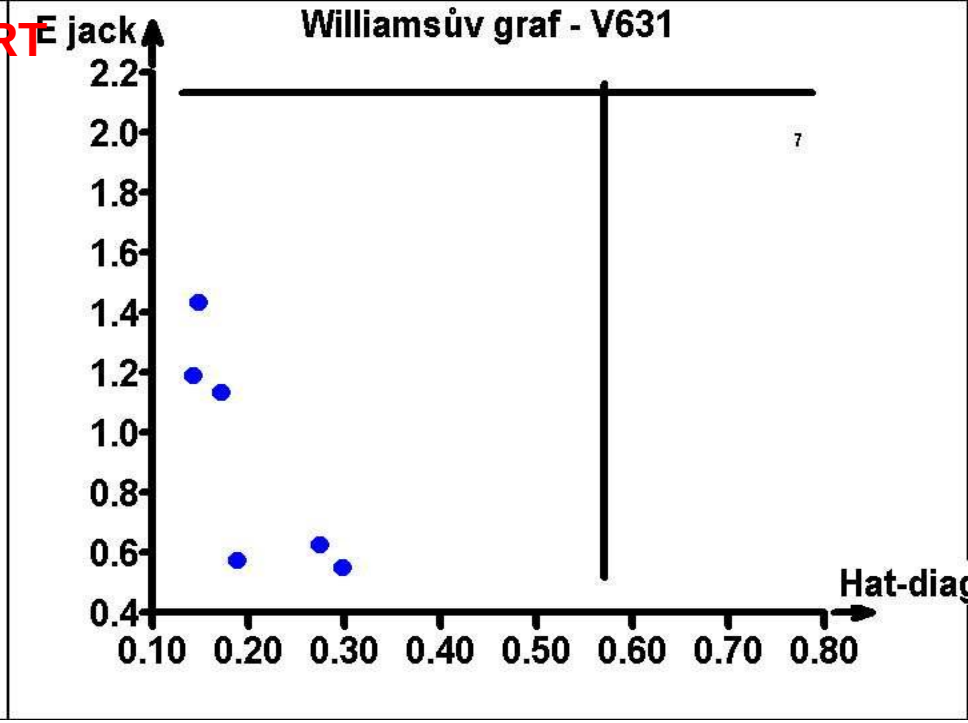
(1) Porovnejte shodnost výsledků stanovení oběma metodami.
(2) Vysvětlete 7 předpokladů MNČ a řešení regresního tripletu.
(3) Ukažte postup validace nové analytické metody testování nulovosti úseku a jednotkovosti směrnice.
(4) Jak se bude řešit tato úloha v případě porušení předpokladů MNČ?

Data: *x* značí AAS [mg/kg], *y* značí ICP-AES [mg/kg]:

Dáno *x*	Stanoveno *y*
25	27
...	...
97	100



QCEXPERT



Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	1.86193	0.35244	Významný	0.00323	0.95595	2.76791
V631x	1.01491	0.00664	Významný	2.27699E-10	0.99784	1.03198

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0.9998929533
Koeficient determinace R^2 :	0.999785918
Predikovaný korelační koeficient Rp :	0.9985017625
Střední kvdratická chyba predikce MEP :	0.4268212711
Akaikeho informační kritérium :	-10.73012051