Inovace bakalářského studijního oboru Aplikovaná chemie

http://aplchem.upol.cz

CZ.1.07/2.2.00/15.0247

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Regrese

Závislost proměnných

- funkční $y = f(x)$
- regresní $y = f(x) + \varepsilon$
  - lze odlišit nezávisle a závisle proměnnou, jasně definovaný příčinný vztah (není ale podmínkou)
  - i nelineární vztahy
- korelační $y \sim x$ resp. $x \sim y$
  - nelze odlišit nezávisle a závisle proměnnou, jedná se o dvě náhodné proměnné se souvislostí
  - dokážeme pouze existenci lineárního vztahu

Lineární regrese

$$y = kx + q$$

příklady:

$pV = nRT$ $\quad s = vt$ $\quad A = \varepsilon cd$

$U = IR$ $\quad o = 2\pi r$ $\quad F = ma$

Lineární regrese

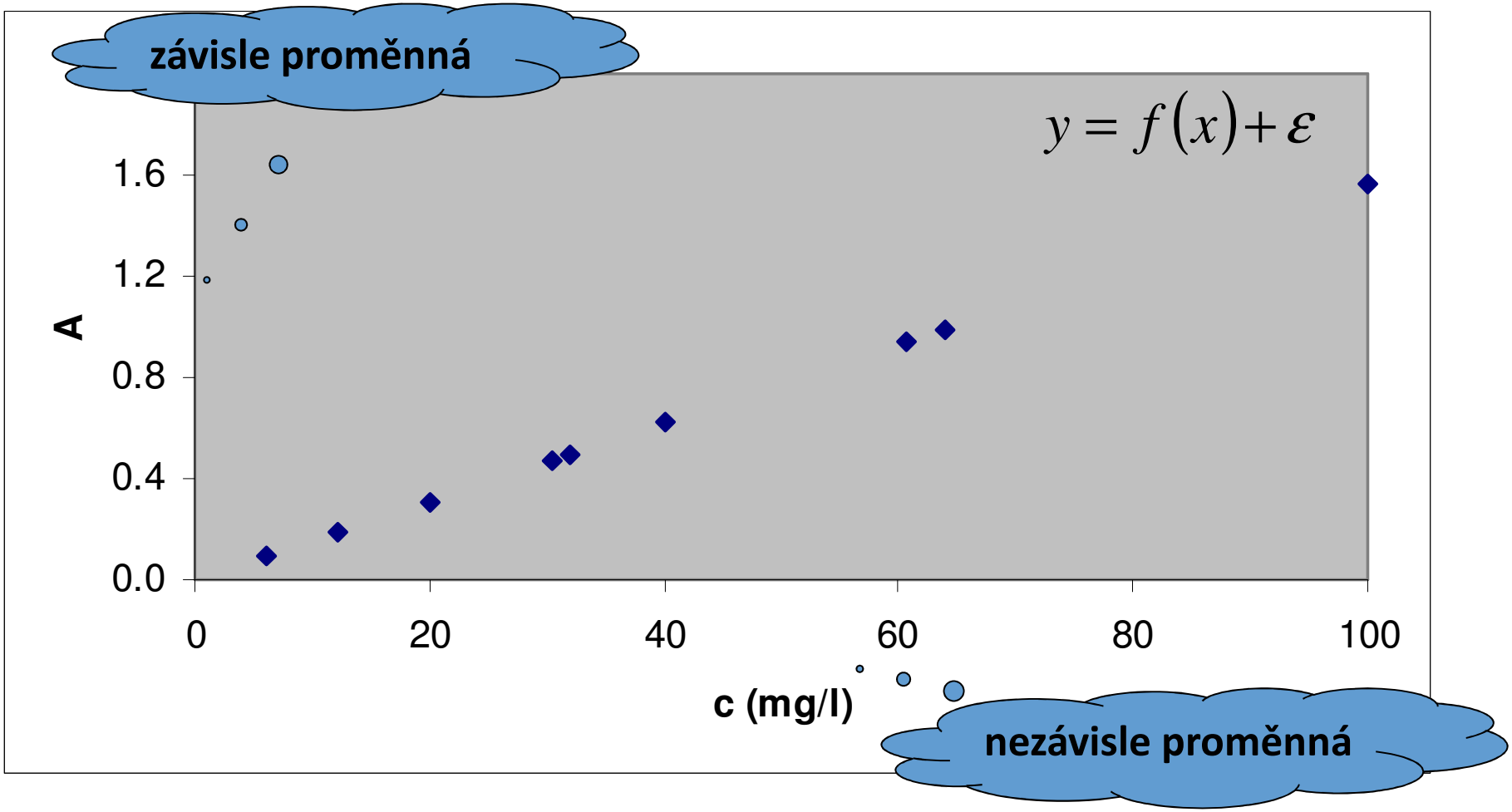
- lineární vzhledem k parametrům!

$$\frac{df\left(\beta_i, x\right)}{d\beta_i} = konst.$$

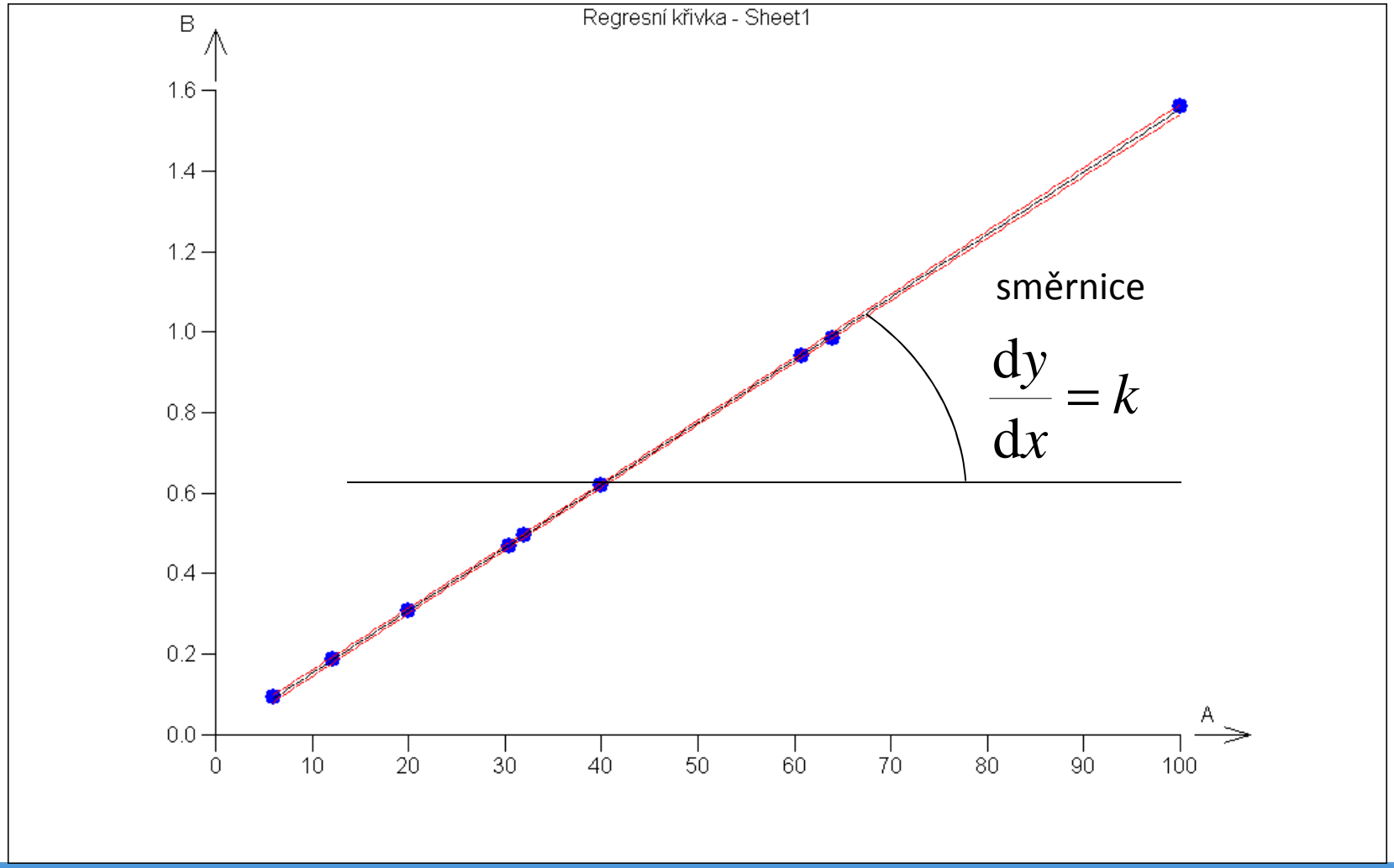
- je lineární vzhledem k parametru *k* a *q*!!!

$$y = k\sin(x) + q$$

Závislost y (f(x)) na x



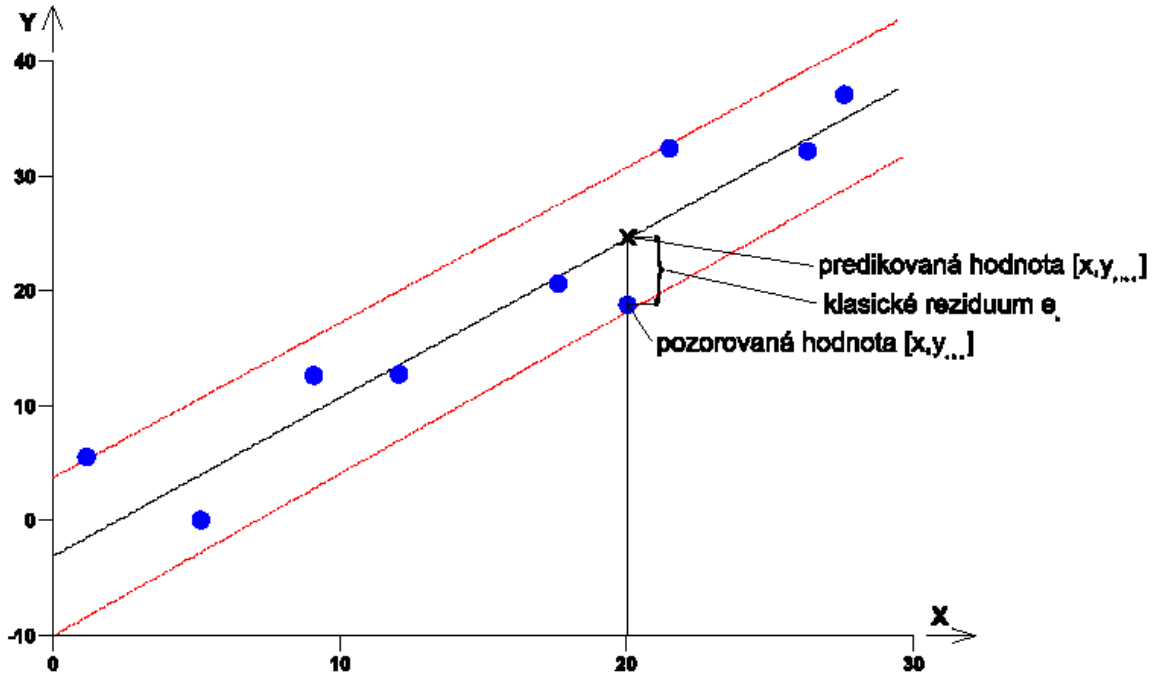
Lineární regrese



Metoda nejmenších čtverců

$$S = \sum_{k}^{N} e_k^2 = \mathbf{e}^T\mathbf{e} = \sum_{k}^{N}(y_k - f(\mathbf{x}_k, \Theta))^2 = (\mathbf{y} - \mathbf{xb})^T(\mathbf{y} - \mathbf{xb})$$

$\min S$



Obecná přímka

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

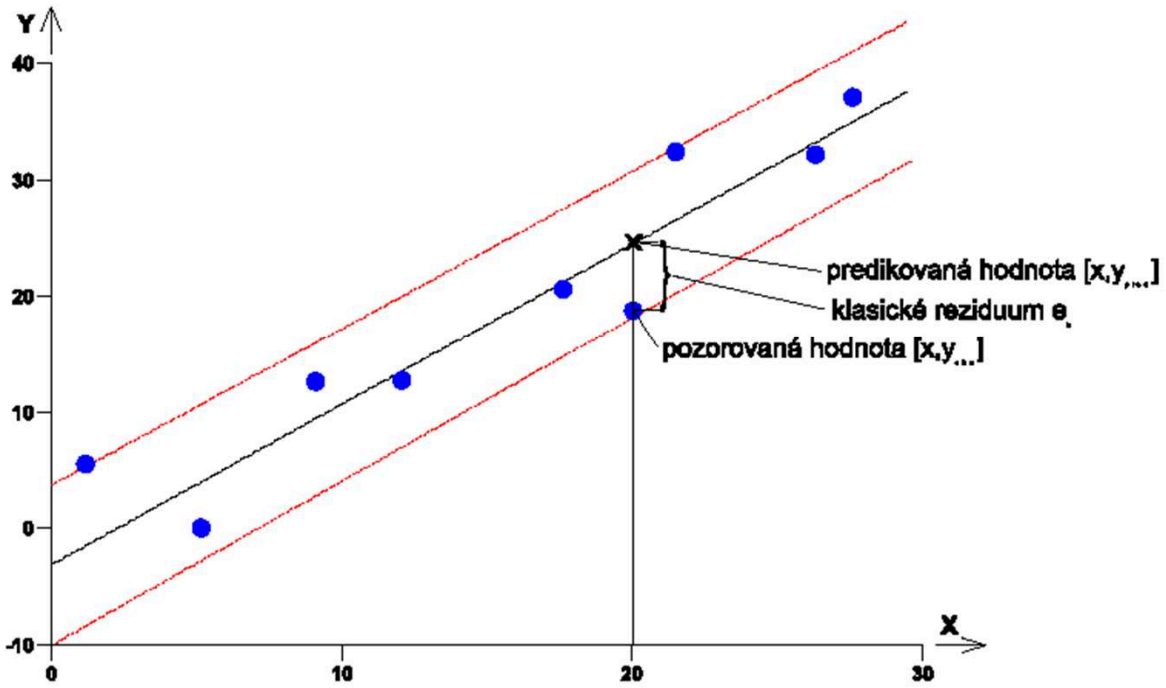
$$b_1 = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2}, \qquad b_0 = \overline{y} - b_1 \overline{x},$$

$$s^2 = \frac{\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i}{n-2}$$

$$b_0 + b_1 x \pm t_{n-2}(\alpha) s \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum x_i^2 - n\overline{x}^2}},$$

pás spolehlivosti

Obecná přímka



$$b_0 + b_1 x \pm t_{n-2}(\alpha) s \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum x_i^2 - n\overline{x}^2}},$$ pás spolehlivosti

Dělení variability

Celková variabilita *y* je

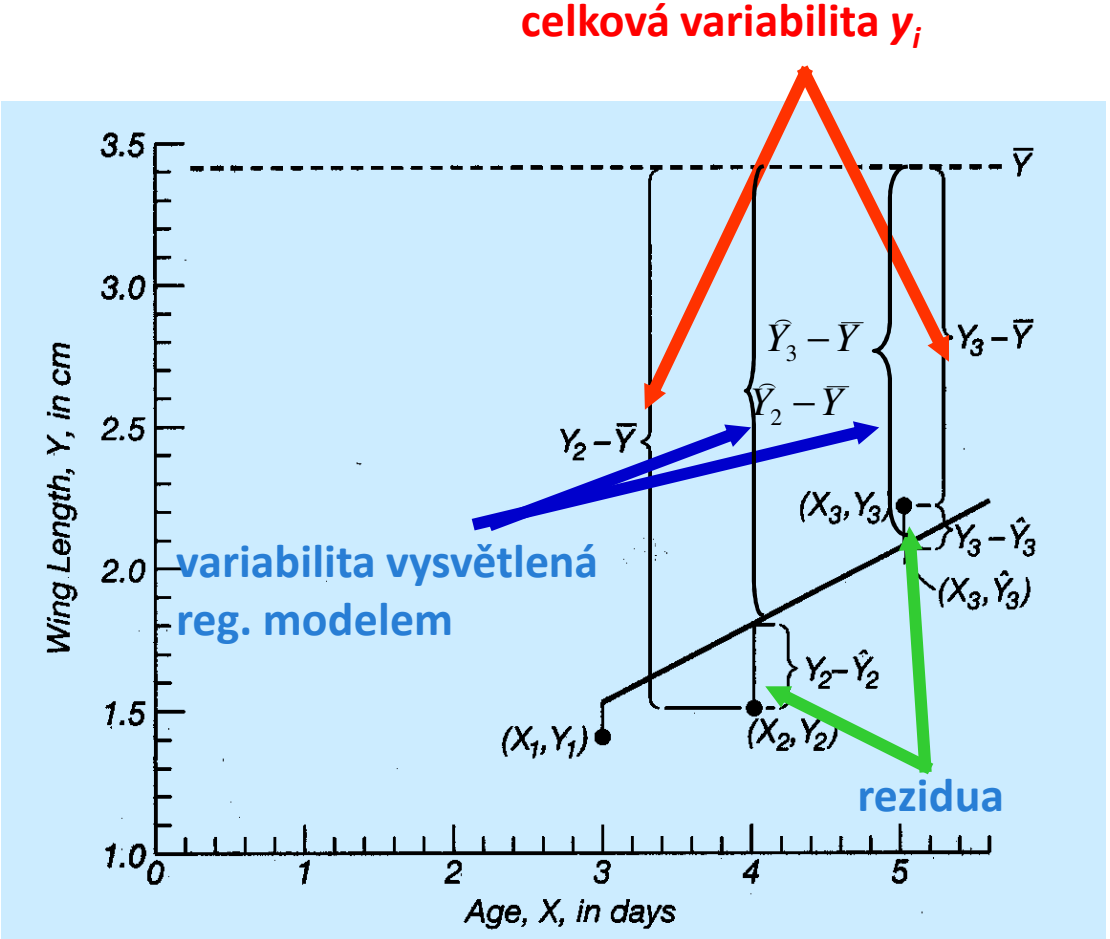
$$SS_{TOTAL} = \sum (y_i - \bar{y})^2$$

lze rozdělit na variabilitu vysvětlenou regresním modelem:

$$SS_{REG} = \sum (\hat{y}_i - \bar{y})^2$$

a zbytkovou, nevysvětlenou variabilitu:

$$SS_{ERROR} = \sum (y_i - \hat{y}_i)^2$$

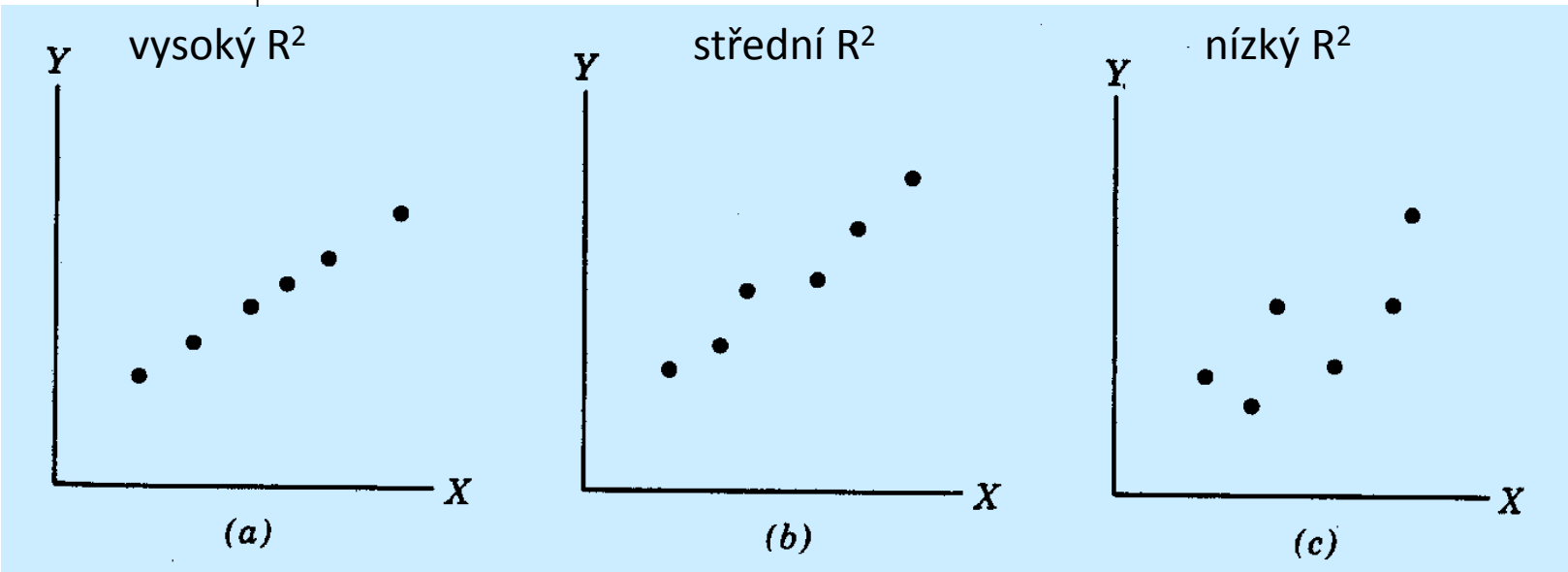


Kvalita modelu

Koeficient determinace (*coefficient of determination*)

udává (*rozsah* $\langle 0,1 \rangle$) část variability závisle proměnné vysvětlenou regresním modelem

$$r^2 = R^2 = \frac{SS_{REG}}{SS_{TOT}}$$



Kvalita modelu

- koeficient determinace
- reziduální směrodatná odchylka
- AIC, MEP (viz dále)
- významnost modelu (F-test) a parametrů (*t*-testy) (viz dále)

příklad

Příklad 7.35. Závislost absorbance $KMnO_4$ při vlnové délce $\lambda = 527$ nm na koncentraci, $d = 1,000$ cm . Byla změřena následující data

c [mg.dm^{-3}]	6,00	12,15	20,00	30,40	32,00	40,00	60,70	64,00	100,00
A	0,094	0,188	0,309	0,470	0,494	0,619	0,940	0,983	1,560

Pro řešení se nejprve vyčíslí potřebné součty tedy

$$\sum c = 365,25 \qquad \sum c^2 = 21912,2725 \qquad \sum cA = 339,8542 \qquad \sum A = 5,657$$

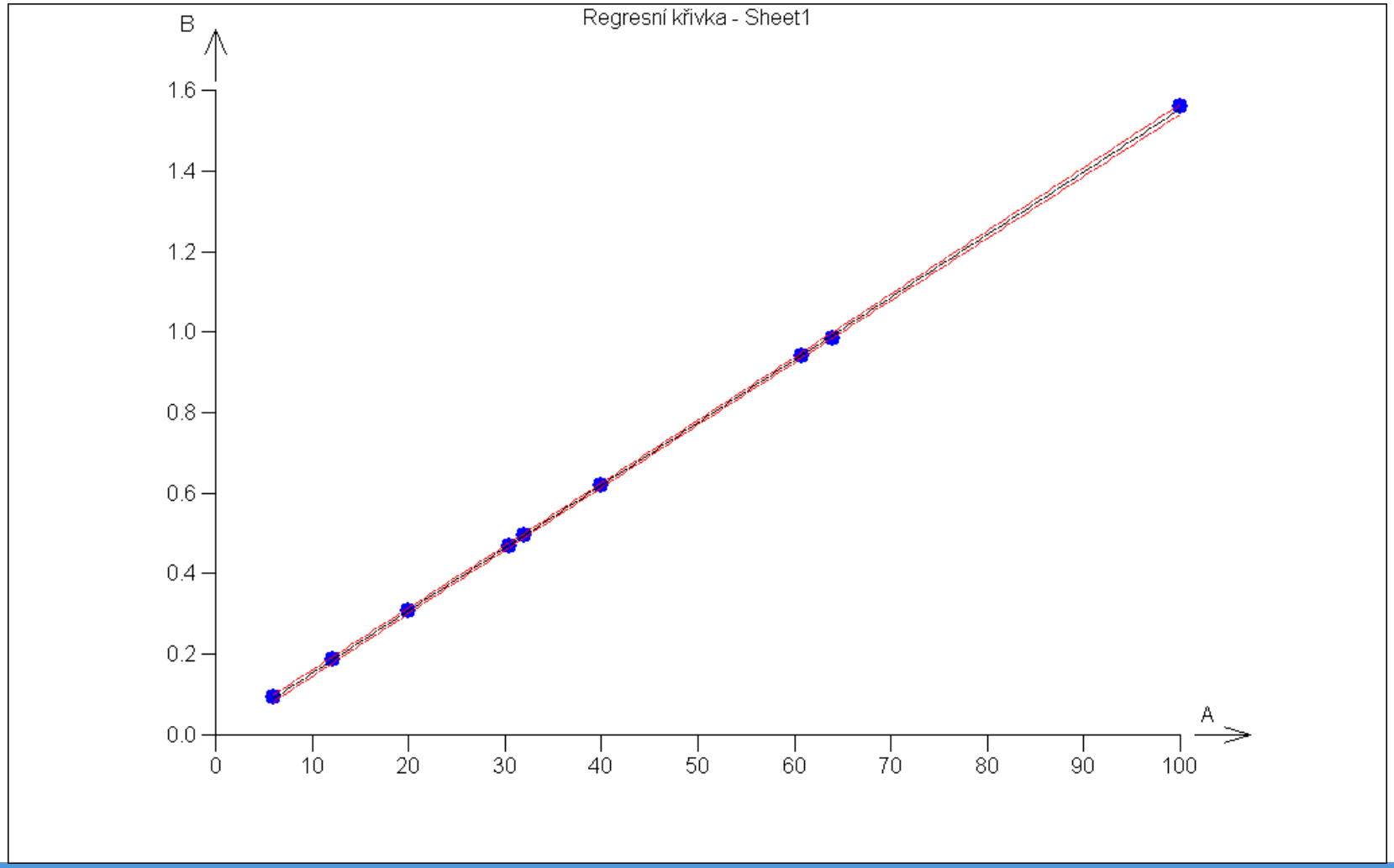
po správném dosazení lze obdržet:

$$b_0 = -2,727.10^{-3} \qquad b_1 = 1,5555.10^{-2} \qquad S_t = 1,71534$$
$$S_e = 1,665.10^{-4} \qquad R^2 = 0,99990 \qquad s^2 = 2,38.10^{-5},$$

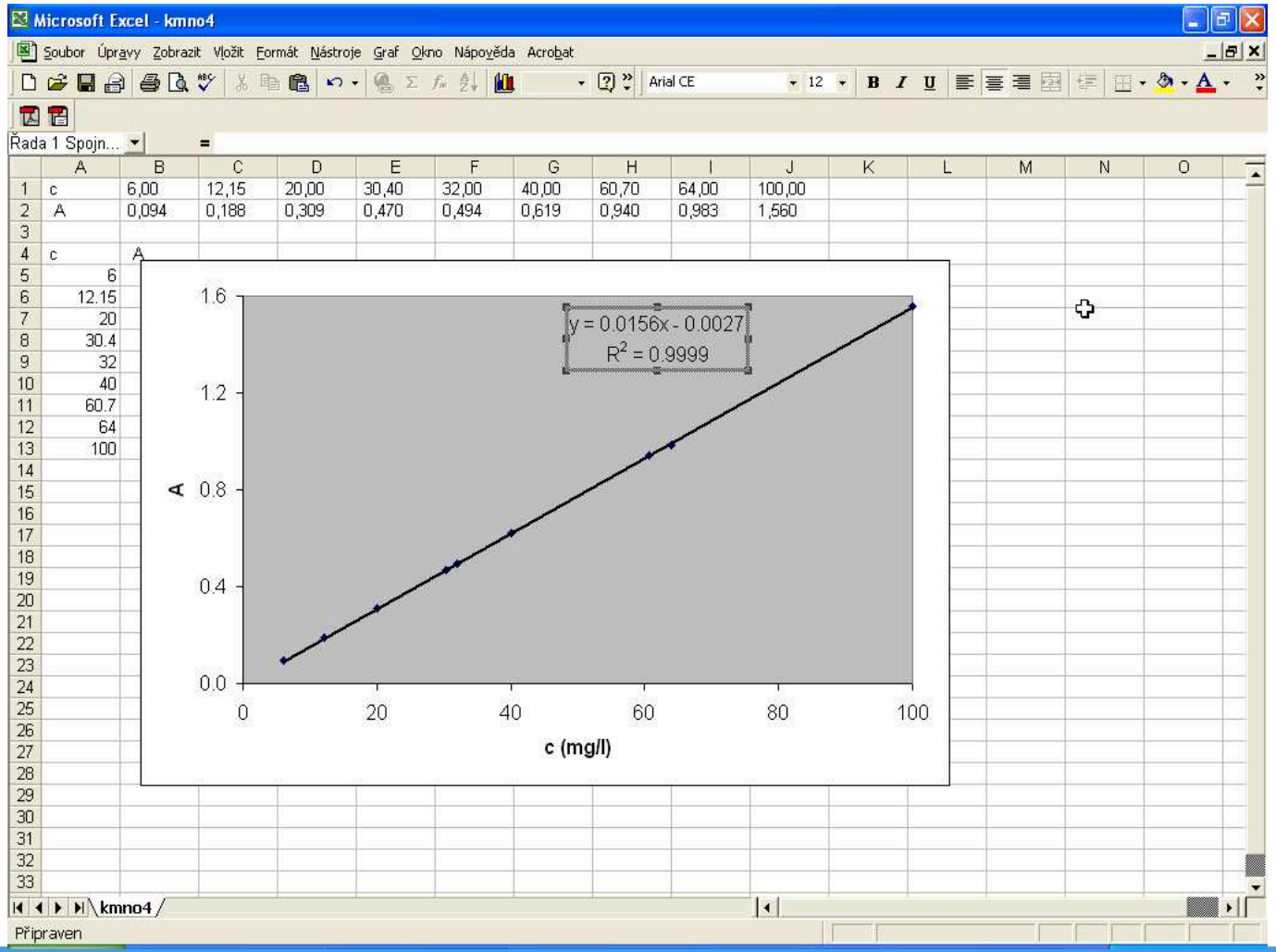
použije-li se model pro přímou úměrnost získají se následující parametry

$$b_1 = 1,5510.10^{-2} \qquad S_e = 1,882.10^{-4} \qquad R^2 = 0,99996 \qquad s^2 = 2,35.10^{-5}.$$

Regresní přímka



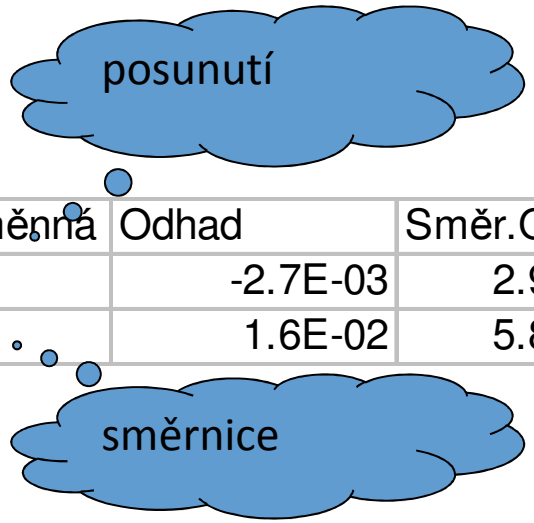
Řešení v MS Excel

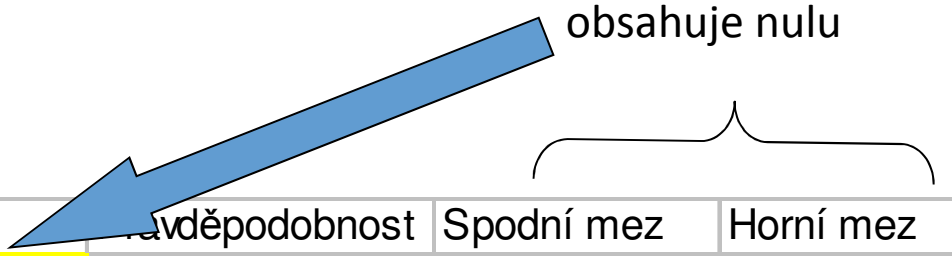


	A	B	C	D	E	F	G	H	I	J
1	c	6,00	12,15	20,00	30,40	32,00	40,00	60,70	64,00	100,00
2	A	0,094	0,188	0,309	0,470	0,494	0,619	0,940	0,983	1,560

	A
4	c
5	6
6	12.15
7	20
8	30.4
9	32
10	40
11	60.7
12	64
13	100

Odhady parametrů





Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-2.7E-03	2.9E-03	Nevýznamný	3.7E-01	-9.5E-03	4.0E-03
A	1.6E-02	5.8E-05	Významný	2.7E-15	1.5E-02	1.6E-02

Statistické charakteristiky regrese	
Vícenásobný korelační koeficient R :	1.000
Koeficient determinace R^2 :	1.000
Predikovaný korelační koeficient Rp :	0.999
Střední kvadratická chyba predikce MEP :	0.000
Akaikeho informační kritérium :	-94.079

Testy hypotéz

roven nule $\boldsymbol{\beta} = \mathbf{0}$. Tento test je také shodný s testem nezávislosti lineárního regresního modelu s $H_0 : R^2 = 0$ oproti $H_1 : R^2 > 0$. Testovací statistika F_e se testuje oproti kritické hodnotě $F_{p-1,n-p}(\alpha)$, kde p je počet regresních parametrů a F_e je definováno jako

$$F = \frac{(n-p)R^2}{(1-R^2)(p-1)}. \tag{7.21}$$

Dále se t testem testují jednotlivé parametry, kde $H_0 : b_i = \beta_i$ a $H_1 : b_i \neq \beta_i$. Často se za parametry β_i opět dosazuje $\beta_i = 0$. Testovací kritérium má tvar

$$t_i = \frac{|b_i - \beta_i|}{\sqrt{s^2(\mathbf{x}^T\mathbf{x})^{-1}}} \tag{7.22}$$

F-test modelu

Fisher-Snedecorův test významnosti modelu	
Hodnota kritéria F :	72108.776
Kvantil F (1-alfa, m-1, n-m) :	5.591
Pravděpodobnost :	2.62E-15
Závěr :	Model je význa

t-testy parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-2.7E-03	2.9E-03	Nevýznamný	3.7E-01	-9.5E-03	4.0E-03
A	1.6E-02	5.8E-05	Významný	2.7E-15	1.5E-02	1.6E-02

Analýza reziduí

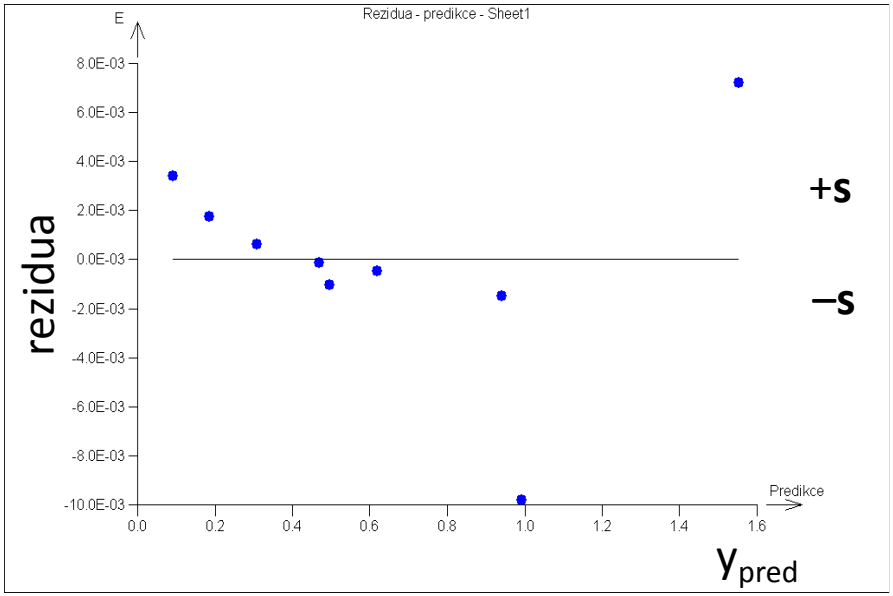
$$\mathbf{e} = \mathbf{y} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} = \mathbf{y} - \mathbf{H}\mathbf{y}$$

projekční matice

konstantní rozptyl

Cook-Weisbergův test heteroskedasticity	
Hodnota kritéria CW :	2.966
Kvantil Chi^2(1-alfa,1) :	3.841
Pravděpodobnost :	0.085
Závěr :	Rezidua vykazují homoskedasticitu.
Jarque-Berrův test normality	
Hodnota kritéria JB :	1.140
Kvantil Chi^2(1-alfa,2) :	5.991
Pravděpodobnost :	0.566
Závěr :	Rezidua mají normální rozdělení.

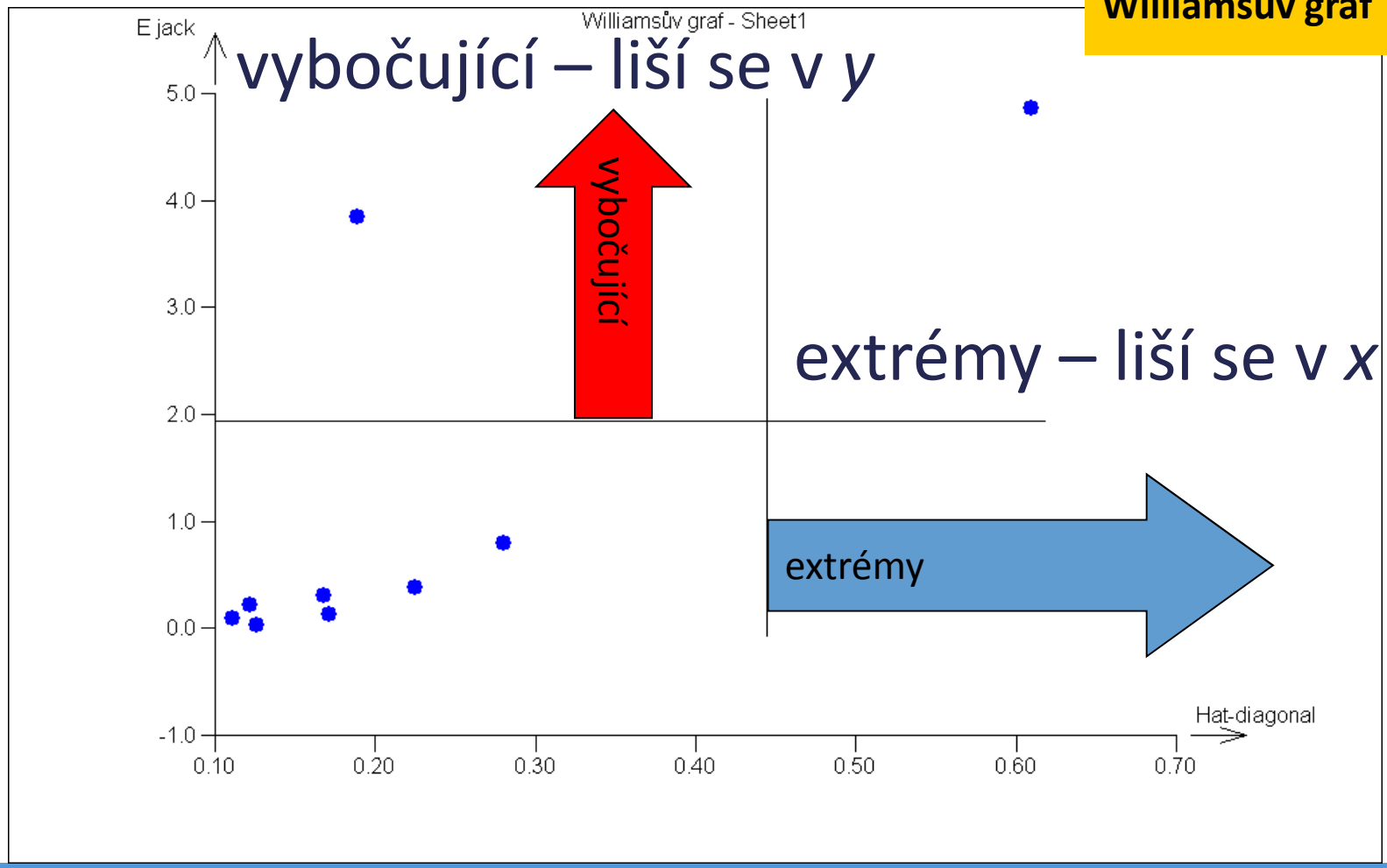
Waldův test autokorelace	
Hodnota kritéria WA :	0.610
Kvantil Chi^2(1-alfa,1) :	3.841
Pravděpodobnost :	0.435
Závěr :	Autokorelace je nevýznamná



evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
OKRESNÍ HOSPODÁŘSKÁ KOMORA OLOMOUC
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Vlivné body

Williamsův graf



evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
OKRESNÍ HOSPODÁŘSKÁ KOMORA OLOMOUC
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Mnohonásobná regrese

$$y = b_1 x_1 + \ldots + b_n x_n + b_0$$

Několik lineárních regresních modelů lze posuzovat, vedle parametrů s a R^2, i na základě tzv. **Akaikova informačního kritéria** definovaného vztahem

$$AIC = n \ln \frac{S_e}{n} + 2p \tag{7.31}$$

nebo na základě **střední kvadratické chyby predikce** (MEP), definované jako

$$MEP = \frac{1}{n} \sum_{i=1}^{n} \frac{e_i^2}{(1 - H_{ii})^2}, \tag{7.32}$$

která pro velké soubory dat n, kdy $H_{ii} \sim 0$ nabývá hodnoty $MEP = S_e/n$. Jako nejlepší se volí model, který má minimální hodnotu AIC a MEP.

Validace

- stará/nová metoda
- předpoklady
  - jednotková směrnice
  - nulové posunutí

Kalibrace

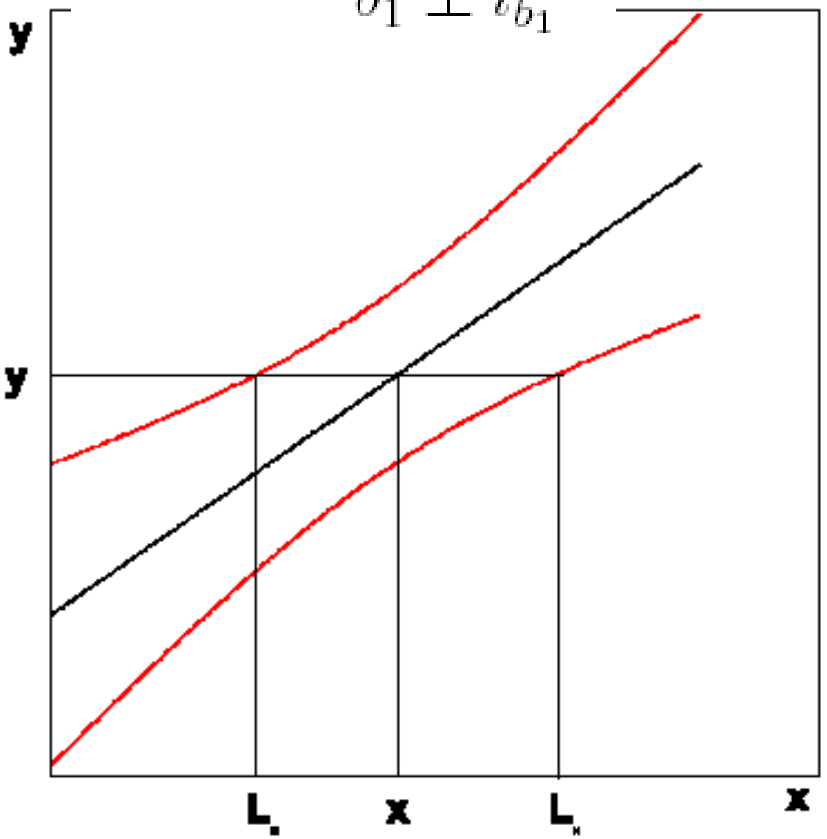
- citlivost – směrnice
- konstrukce bez vlivných bodů
- zahrnutí slepého pokusu
- požadujeme zpětný odhad
  - ze změřeného *y* odhadnout *x*

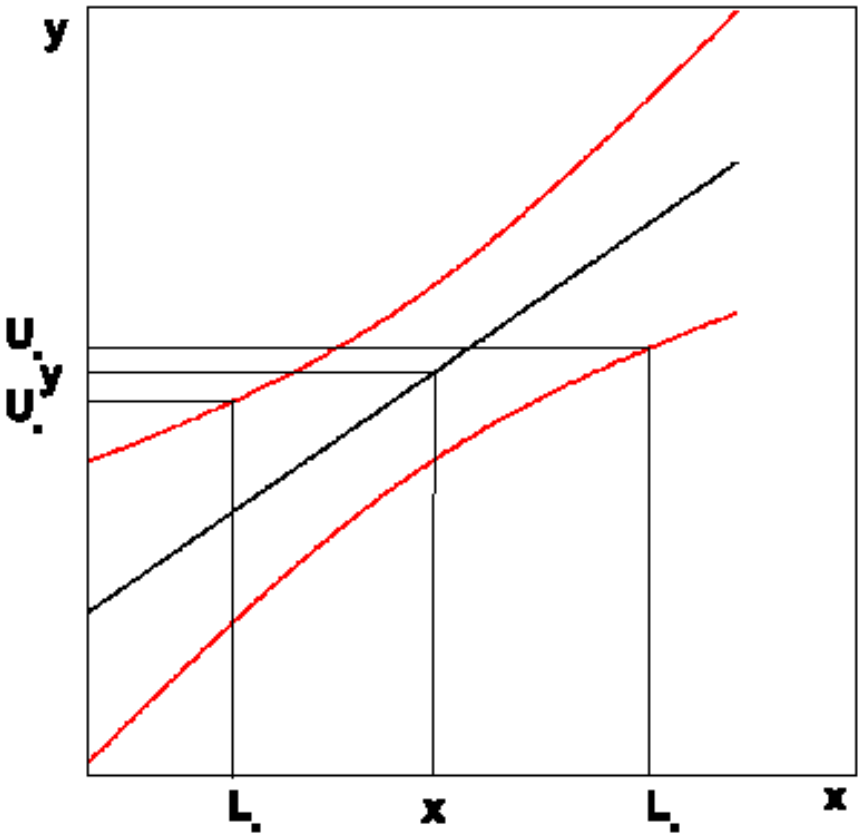
$$x = \overline{x} + \frac{y - \overline{y}}{b_1}, \quad x = \frac{y - b_0}{b_1},$$

Intervaly spolehlivosti

$$L_{D,H} = \frac{y - b_0}{b_1 \pm i_{b_1}},$$

hrubý odhad (správně viz skripta)





Meze

Mez detekce signálu y_d je taková hodnota signálu, nad kterou s pravděpodobností $1 - \alpha$ je signál projevem přítomnosti x ve vzorku.