Přednáška 10 – Testy v regresi. Validace regrese

- Testy na vhodnost k regresní analýze
- Testy hypotéz pro validaci regrese

Testy na vhodnost k regresní analýze = Testy nezávislosti pro spojitá data

Parametrické testy s předpokladem normality

Pearsonův test

H_0: veličiny jsou lineárně nezávislé

V případě zamítnutí H0:

data jsou vhodná k lineární regresi

$$y = b_0 + b_1 x$$

Neparametrické testy bez předpokladu normality

Spearmanův test

H_0: veličiny jsou nezávislé

V případě zamítnutí H0:

data jsou vhodná k nelineární regresi

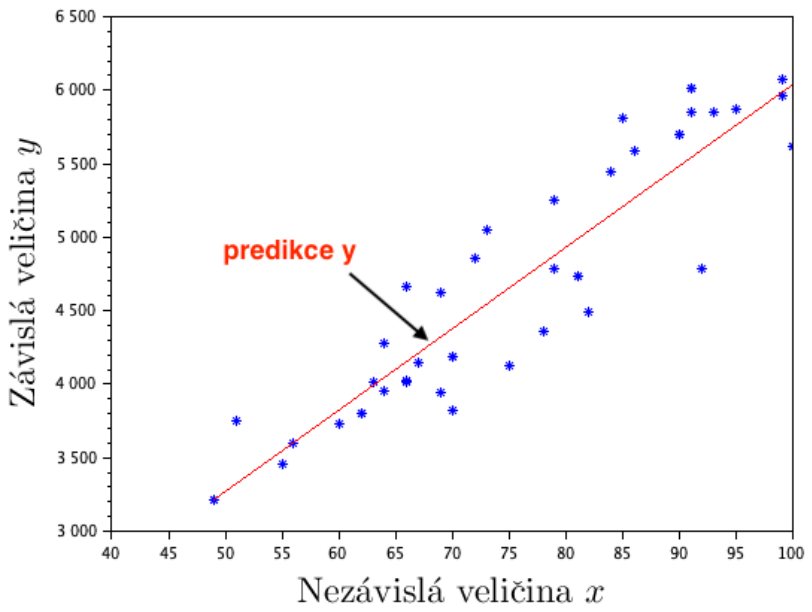
$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n$$

$$y = b_0 e^{b_1 x}$$

http://staff.utia.cas.cz/uglickich/pdfka/volbaTH.pdf

Validace regrese

- data vhodná k regresi = můžeme použít regresi ≠ úspěšná regrese
- po provedení regrese – ověření výsledků regrese
  (zda vybraná regresní metoda vyhovovala naměřeným datům)
- proces ověření výsledků regrese – validace regrese



Základ validace:

- porovnání dat y a predikce $\hat{y}$

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

- x_i – data
- $\hat{b}_0, \hat{b}_1$ – odhad
- $\hat{y}_i$ – predikce y

- testujeme, zda predikce ukazuje správně trend vývoje dat

- testy hypotéz validace regrese rozlišujeme podle typu regrese (lineární, nelineární)

F-test podílu vysvětleného a nevysvětleného rozptylu

Odchylka dat od průměru

testování shody dat y a predikce $\hat{y}$

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} =$$

$$= \underbrace{y_i - \hat{y}_i}_{\text{odchylka dat od predikce}} + \underbrace{\hat{y}_i - \bar{y}}_{\text{odchylka predikce od průměru}},$$

$$\underbrace{\underbrace{\text{reziduum } e_i = y_i - \hat{y}_i}_{\text{nevysvětlená odchylka}}}_{\text{vysvětlená odchylka}}$$

$$F = \frac{(n-2)\ \text{vysvětlená odchylka}}{\text{nevysvětlená odchylka}} = \frac{(n-2)\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \sim \text{Fisherovo}$$

- pouze pravostranný

Nulová hypotéza:

H_0 : zvolená regrese je **nevhodná**

Alternativní hypotéza:

H_A : regrese je vhodná

- je výhodněji zamítnout
- pokud nezamítneme – jiná regresní metoda

Příklad: Sledujeme vývoj ceny 100g kakaa a mléčné čokolády v Kč ročně v období 2004-2018. Jsou data vhodná k regresi? Pokud ano, potřebujeme ověřit výsledky regrese.

Jsou data vhodná k regresi?

- oba výběry pochází z normálního rozdělení
- Pearsonův test: p-hodnota = 8.7405e-04
- data jsou vhodná k lineární regresi

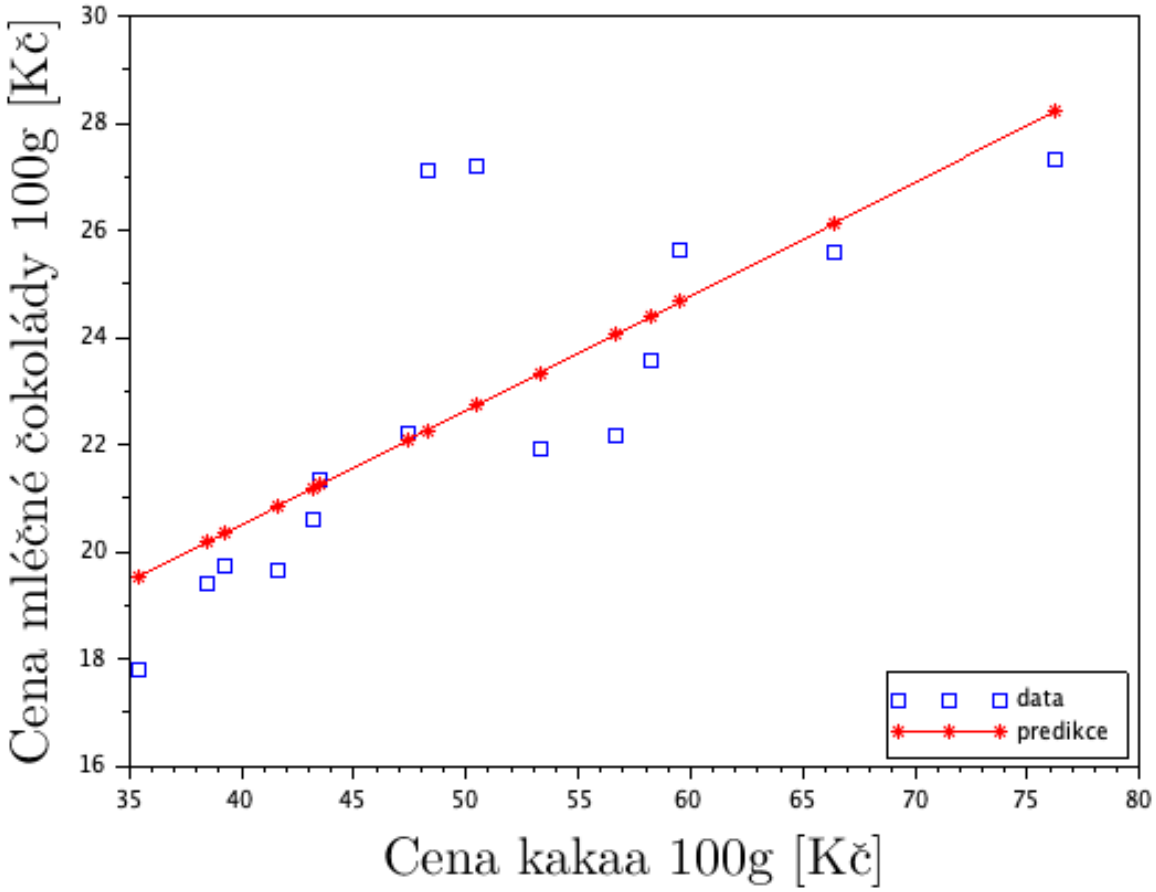
$$y = b_0 + b_1 x = 11.99 + 0.21x$$

- predikce ceny čokolády $\hat{y}$ – hodnoty na přímce

Byla regrese vhodně zvolena?

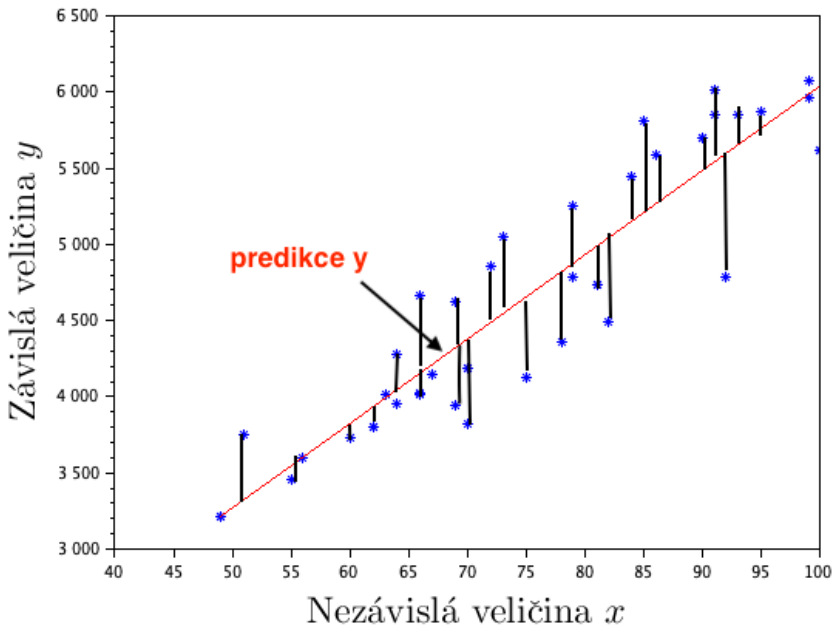
- F-test: p-hodnota = 0.000874
- zvolená regrese je vhodná

kakao	čokoláda	$\hat{y}$
41.61	19.65	20.84
39.29	19.75	20.35
35.38	17.78	19.52
38.48	19.42	20.18
43.21	20.58	21.19
53.3	21.91	23.33
56.67	22.17	24.05
58.24	23.58	24.39
47.46	22.22	22.09
43.46	21.32	21.24
59.53	25.63	24.66
66.45	25.57	26.14
76.24	27.31	28.22
48.27	27.12	22.26
50.52	27.20	22.74


Cena mléčné čokolády 100g [Kč]
Cena kakaa 100g [Kč]
data
predikce
30
28
26
24
22
20
18
16
35
40
45
50
55
60
65
70
75
80

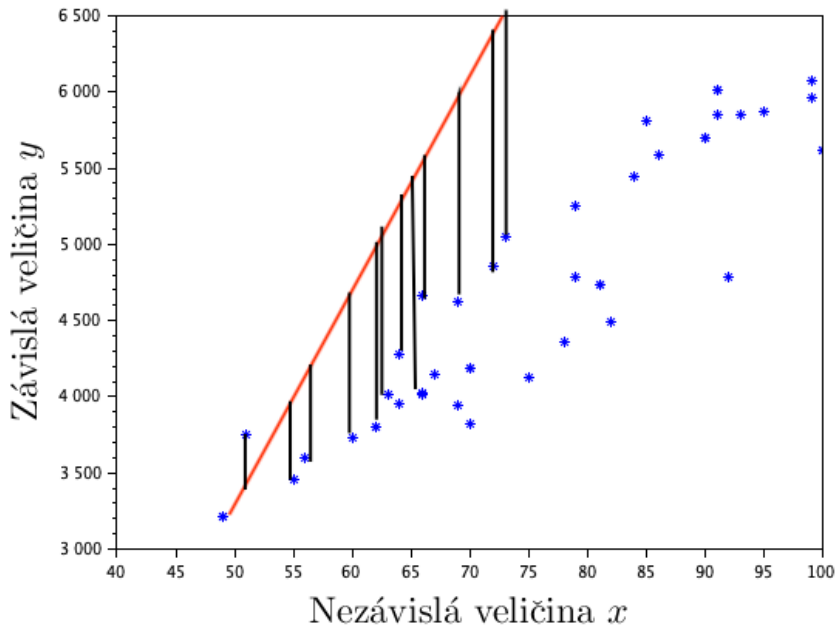
Test nezávislosti (bělosti) reziduí $e_i = y_i - \hat{y}_i$

- regrese je zvolena dobře:
rezidua – kladná, záporná, nezávislá



$$b_i = e_i - \tilde{e}_{0.5}, \quad b = \sum_{i=1}^{n} b_i$$

- regrese je zvolena špatně:
rezidua narůstají



$$T = \frac{2b-(n-2)}{\sqrt{n-1}} \sim N(0,1)$$

Nulová hypotéza:

H_0 : zvolená regrese je **vhodná**

Alternativní hypotéza:

H_A : regrese není vhodná

- je výhodněji nezamítnout
- pokud zamítneme – jiná regresní metoda

Test autokorelace reziduí – Durbinův-Watsonův test dwtest

- Validace nelineární regrese

Nulová hypotéza:

H_0 : rezidua jsou nekorelovaná – zvolená regrese je **vhodná**

Alternativní hypotéza:

H_A : rezidua jsou autokorelovaná – regrese není vhodná

Statistika:

$$T = \frac{\sum_{i=1}^{n-1}(e_{i+1} - e_i)^2}{\sum_{i=1}^{n} e_i^2}$$

je výhodněji nezamítnout

- pokud zamítneme – jiná regresní metoda

Příklad: Sledujeme měsíční spotřebu elektřiny a rozlohu bytů. Jsou tato data vhodná k regresi? Pokud ano, použijeme je a ověříme, zda zvolený typ regrese byl vhodný.

- větší byt – vyšší spotřeba? lineární regrese?

Jsou data vhodná k regresi?

- první z výběrů nemá normalitu
- Spearmanův test: p-hodnota $= 2.841\text{D-}13$
- data jsou vhodná k nelineární regresi

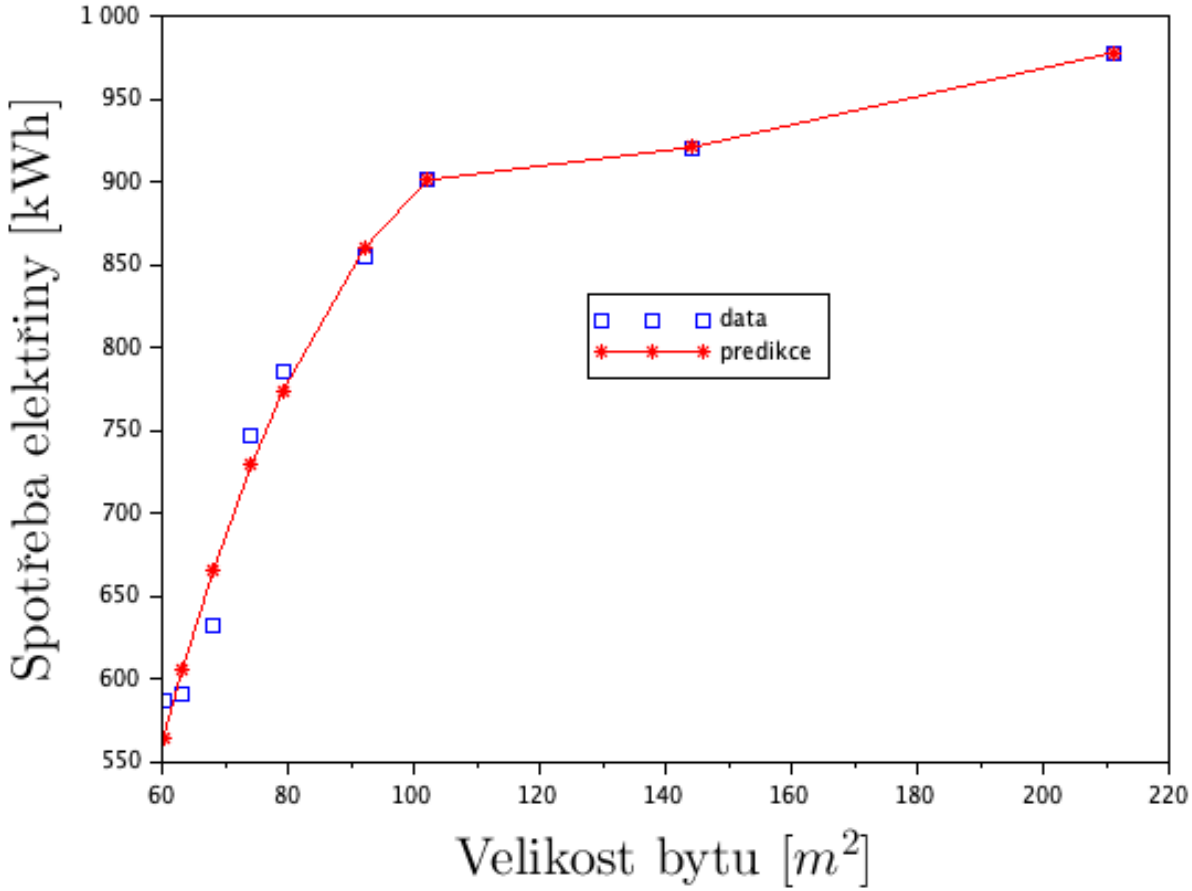
$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$$
$$y = -1011.51 + 40.78x - 0.28x^2 + 0.00062x^3$$

- predikce spotřeby $\hat{y}$ – hodnoty na křivce

Byla regrese vhodně zvolena?

- dwtest: p-hodnota $= 0.9378$
- použili jsme správnou metodu

m^2	kWh	$\hat{y}$
60	591	564.26
63	586	604.68
68	632	665.61
74	747	728.68
79	785	773.4
92	855	859.83
102	902	900.96
144	920	920.68
211	978	977.89
60	591	564.26
63	586	604.68
68	632	665.61
74	747	728.68
79	785	773.40
92	855	859.83
102	902	900.96
144	920	920.69
211	978	977.89



Poznámka:

- na cvičení v Matlabu:
- validace lineární regrese – F-test
- validace nelineární regrese – dwtest