

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.



REGRESNÍ DIAGNOSTIKA

31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Regresní diagnostika

Vyšetřuje **regresní triplet**, což představuje

- **Kritiku dat** (zkoumá kvalitu dat pro navržený model)
- **Kritiku modelu** (zkoumá kvalitu modelu pro daná data)
- **Kritiku metody odhadu** (prověřuje splnění všech předpokladů požadovaných metodou MNČ)



EDA - identifikace

1. Nevhodnost dat (malé rozmezí nebo přítomnost vybočujících bodů),
2. Nesprávnost navrženého modelu (skryté proměnné)
3. Multikolinearita
4. Nenormalita, když jsou proměnné náhodné veličiny



Vyšetření vlivných bodů:

- (1) odlehlé body,
- (2) extrémy.

Metoda:

- grafy vlivných bodů,
- vybrané diagnostiky vlivných bodů.



Výskyt vlivných bodů (VB):

- zdrojem řady problémů,
- zkreslení odhadů a růst rozptylů odhadů parametrů.

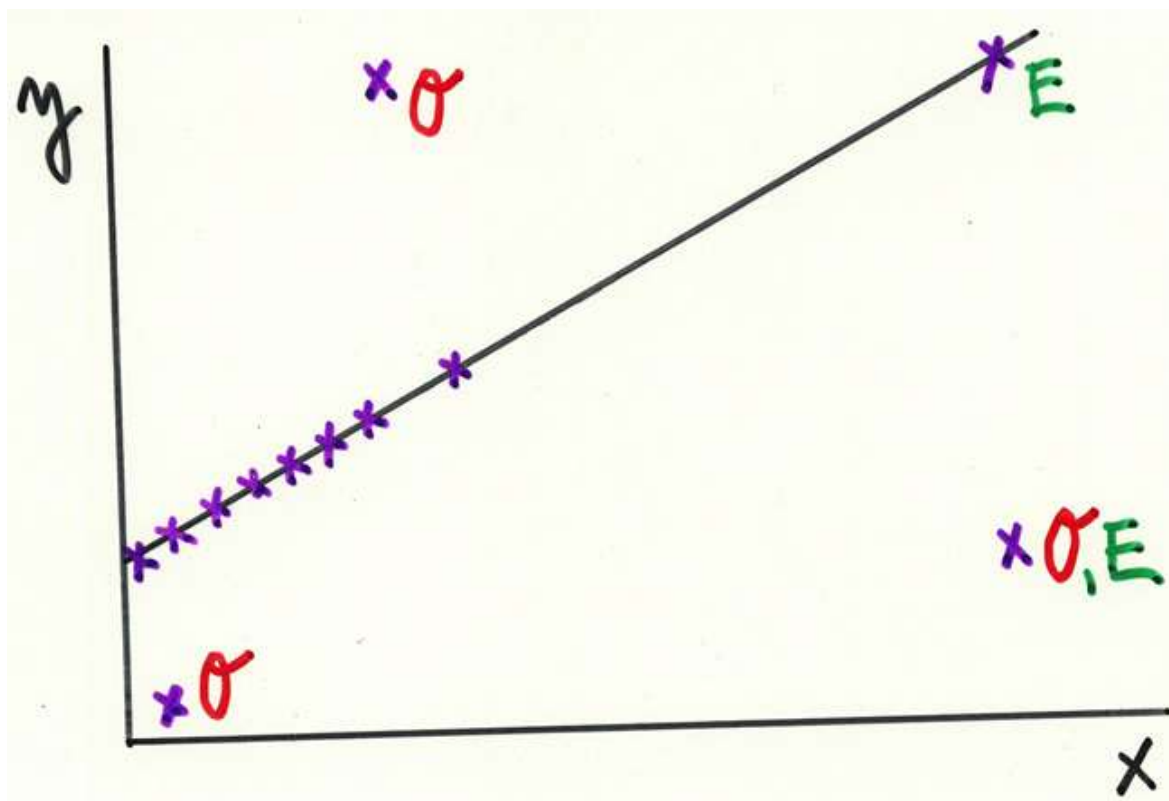
Dělení vlivných bodů dle charakteru:

1. **Hrubé chyby:** vybočující pozorování.
2. **Body s vysokým vlivem** (tzv. golden points): rozšiřují predikční schopnosti modelu.
3. **Zdánlivě vlivné body:** důsledek nesprávného regresního modelu.



Dělení vlivných bodů dle výskytu

1. vybočující pozorování (outliers, O): na y se liší
2. extrémy (high leverage points, E): liší se na ose x





Statistická analýza reziduí

1. Klasická rezidua $\hat{e}_i = y_i - x_i b$,

Nesprávné představy o reziduích:

1. Rozdělení reziduí je stejné jako rozdělení chyb,
2. Vlastnosti reziduí jsou shodné s vlastnostmi chyb,
3. Čím je reziduum $\hat{e}_i$ větší, tím je bod vlivnější, a tím spíše by se měl z dat vyloučit.

$$\hat{e} = Py = P(X\beta + \varepsilon) = P\varepsilon = (E - H)\varepsilon$$



rozepsáním

$$\hat{e}_i = (1 - H_{ii})y_i - \sum_{j \neq i}^n H_{ij}y_j = (1 - H_{ii})\varepsilon_i - \sum_{j \neq i}^n H_{ij}\varepsilon_j$$

slovy: každé $\hat{e}_i$ je lineární kombinací všech chyb ε_i .

Rozdělení reziduí je závislé:

1. na rozdělení chyb,
2. na prvcích projekční matice H ,
3. na velikosti výběru n .



Efekt supernormality

I když chyby ε nemají normální rozdělení, rozdělení reziduí $\hat{e}$ je bližší normálnímu.

- a) **Malé výběry:** prvky H veliké a převažující roli má $\sum_{j=1}^n H_{ij} \varepsilon_j$. Rozdělení součtu se bude více blížit normalitě než rozdělení původních chyb ε .
- b) **Dostatečně veliké výběry:** $\frac{1}{n} \approx 0$, je $\hat{e}_i \approx \varepsilon_i$, rozdělení reziduí je blízké rozdělení chyb.



Vlastnosti klasických reziduí

1. Rozptyl reziduí

$$D(\hat{e}_i) = (1 - H_{ii})\hat{\sigma}^2$$

je nekonstantní, i když rozptyl chyb $\hat{\sigma}^2$ je konstantní.

2. Rezidua jsou korelovaná: existuje párový korelační koeficient r_{ij} mezi dvěma rezidui e_i a e_j

$$r_{ij} = \frac{-H_{ij}}{\sqrt{(1 - H_{ii})(1 - H_{jj})}}$$

i když chyby ε_i a ε_j jsou nezávislé.

3. Rezidua neindukují silně extrémní hodnoty.



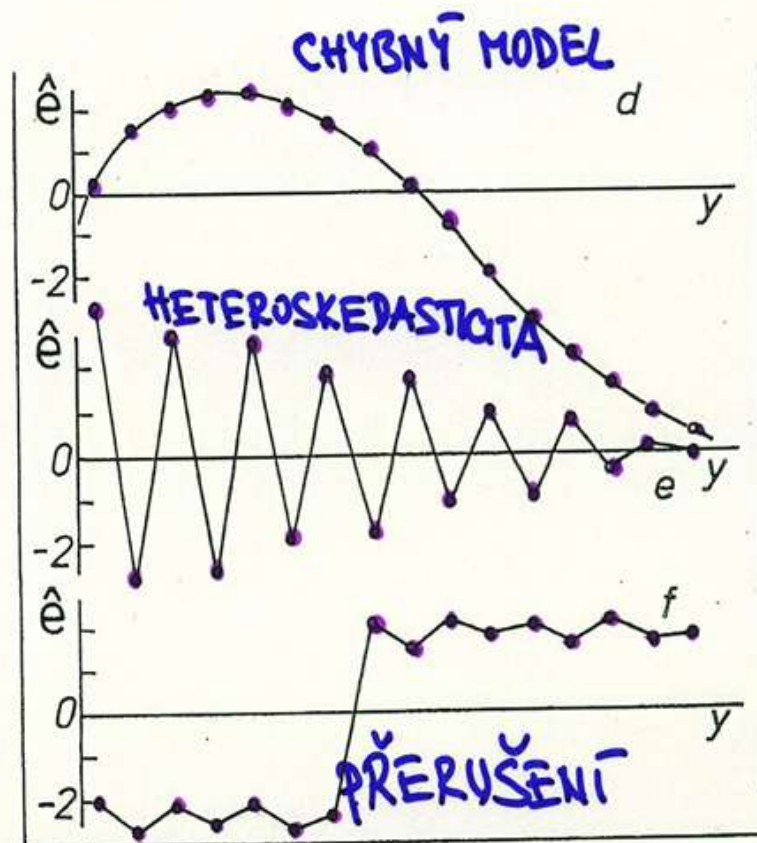
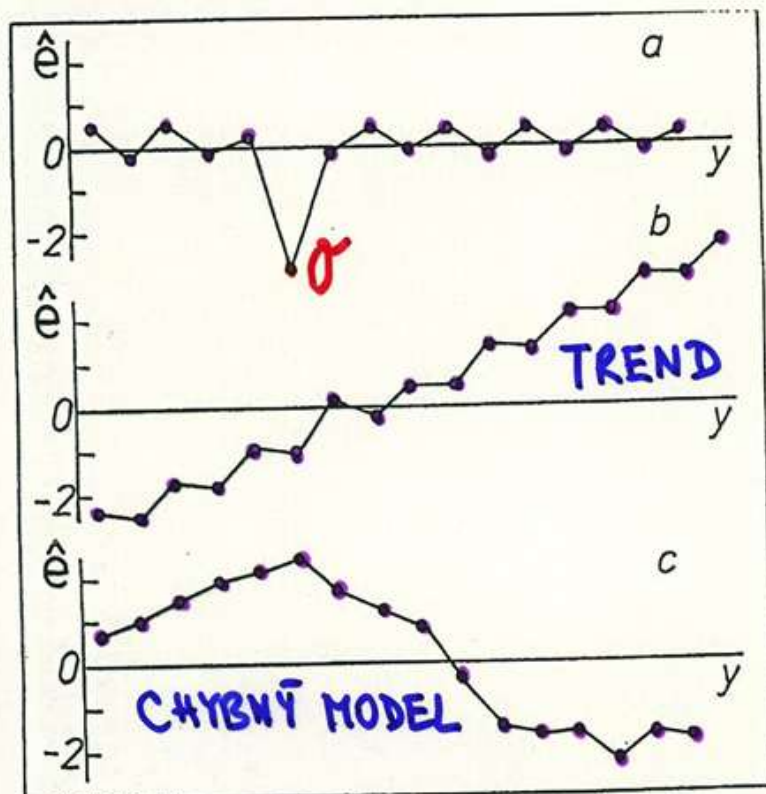
Vlastnosti klasických reziduí

- 4. Rezidua jsou normálnější než chyby (efekt supernormality)
- 5. U malých výběrů nemusí správně indikovat model.



Statistická analýza klasických reziduí

1. Grafická analýza



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





2. Numerická analýza

- (a) Střední hodnota reziduí $E(\hat{e}) = 0$
- (b) Průměrné reziduum $|\bar{e}| = \frac{1}{n} \sum |\hat{e}_i| \approx \epsilon$
- (c) Směrodatná odchylka střední hodnoty reziduí $s(\hat{e}) \approx \epsilon$
- (d) Koeficient šikmosti souboru reziduí $g_1(\hat{e}) \approx 0$
- (e) Koeficient špičatosti souboru reziduí $g_2(\hat{e}) \approx 3$
- (f) Pearsonův χ^2 – test dobré shody χ_{exp}^2 vs. χ_{crit}^2
- (g) Hamiltonův R-faktor relativní četnosti $R \approx 0.5\%$



2. Normovaná rezidua $\hat{\sigma}_N$

$$\hat{e}_{Ni} = \hat{e}_i / \hat{\sigma}$$

normálně rozdělené veličiny $\hat{e}_{Ni} \approx N(0,1)$.

Diagnostika: pravidlo 3σ , tj. rezidua větší než $\pm 3\hat{\sigma}$ indikují vybočující.

3. Standardizovaná rezidua $\hat{\sigma}_S$

$$\hat{e}_{Si} = \frac{\hat{e}_i}{\hat{\sigma} \sqrt{1 - H_{ii}}}$$

Mají konstantní rozptyl. Maximální hodnota $\hat{e}_{Si}$ je ohraničena velikostí $\sqrt{n - m}$.

Diagnostika: k indikaci heteroskedasticity



4. Jackknife rezidua $\hat{\sigma}_J$ („plně studentizovaná“)

užijeme místo $\hat{\sigma}$ odhadu směrodatné odchylky $\hat{\sigma}_{(-i)}$,

$$\hat{e}_{Ji} = \hat{e}_{Si} \sqrt{\frac{n - m - 1}{n - m - \hat{e}_{Si}^2}} = \sqrt{n - m} \cotg \Theta_i$$

mají Student. rozdělení s $(n - m - 1)$ stupni volnosti.

Diagnostika: k identifikaci vybočujících bodů (outliers).



5. Predikovaná rezidua $\hat{\sigma}_P$

$$\hat{e}_{Pi} = y_i - x_i b_{(i)} = \frac{\hat{e}_i}{1 - H_{ii}}$$

kde $b_{(i)}$ jsou MNČ odhady ze všech bodů, kromě i-tého.

Diagnostika: indikace vybočujících hodnot (outliers).



6. Rekurzivní rezidua $\hat{\sigma}_R$

Dopředná rekurzivní rezidua jsou definována vztahy

$$\hat{e}_{Ri} = 0, \quad i = 1, \dots, m$$

$$\hat{e}_{Ri} = \frac{y_i - x_i b_{i-1}}{\sqrt{1 + x_i (X_{i-1}^T X_{i-1})^{-1} x_i^T}}, \quad i = m + 1, \dots, n$$

kde b_{i-1} jsou odhady získané z prvních $(i - 1)$ bodů.

Diagnostika: umožňují identifikovat autokorelaci, čili nestabilitu modelu, např. v čase.



Rezidua k diagnostice

1. K detekci heteroskedasticity: standard. rezidua $\hat{e}_{Si}$
2. K detekci vybočujících bodů: Jackknife rezidua $\hat{e}_{Ji}$ nebo predikovaná rezidua $\hat{e}_{Pi}$
3. K detekci autokorelace: rekurzivní rezidua $\hat{e}_{Ri}$

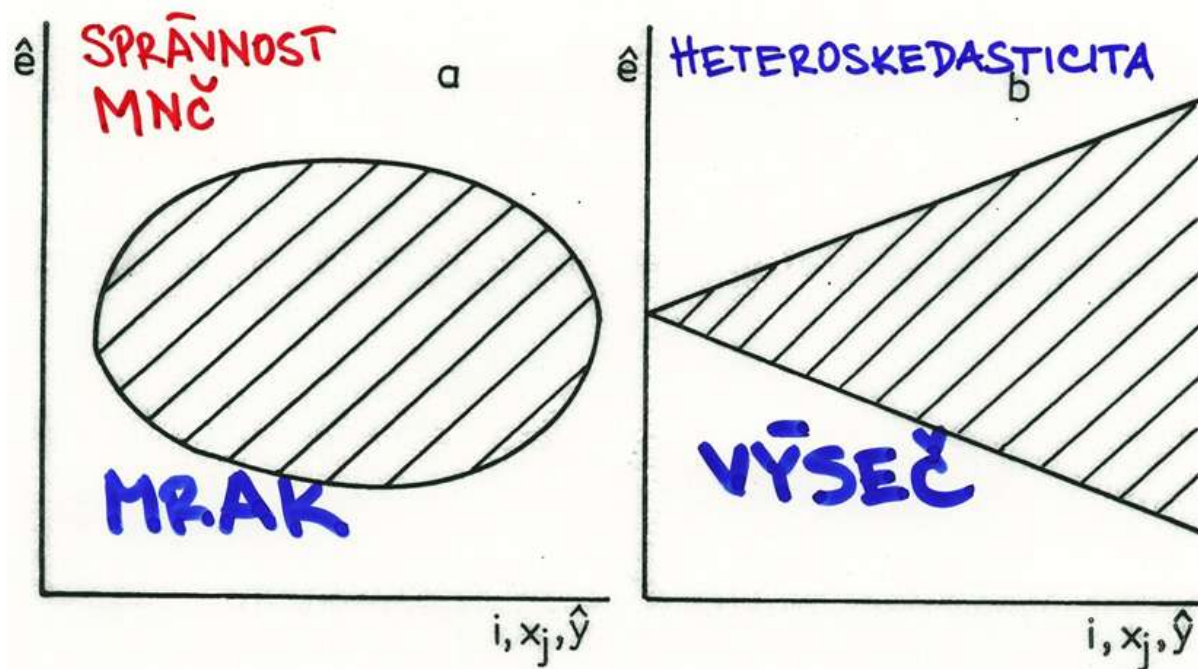


Obrazce v diagnostických grafech

Typ I: Graf reziduí $\hat{e}_i$ proti indexu i .

Typ II: Graf reziduí $\hat{e}_i$ proti proměnné x_j .

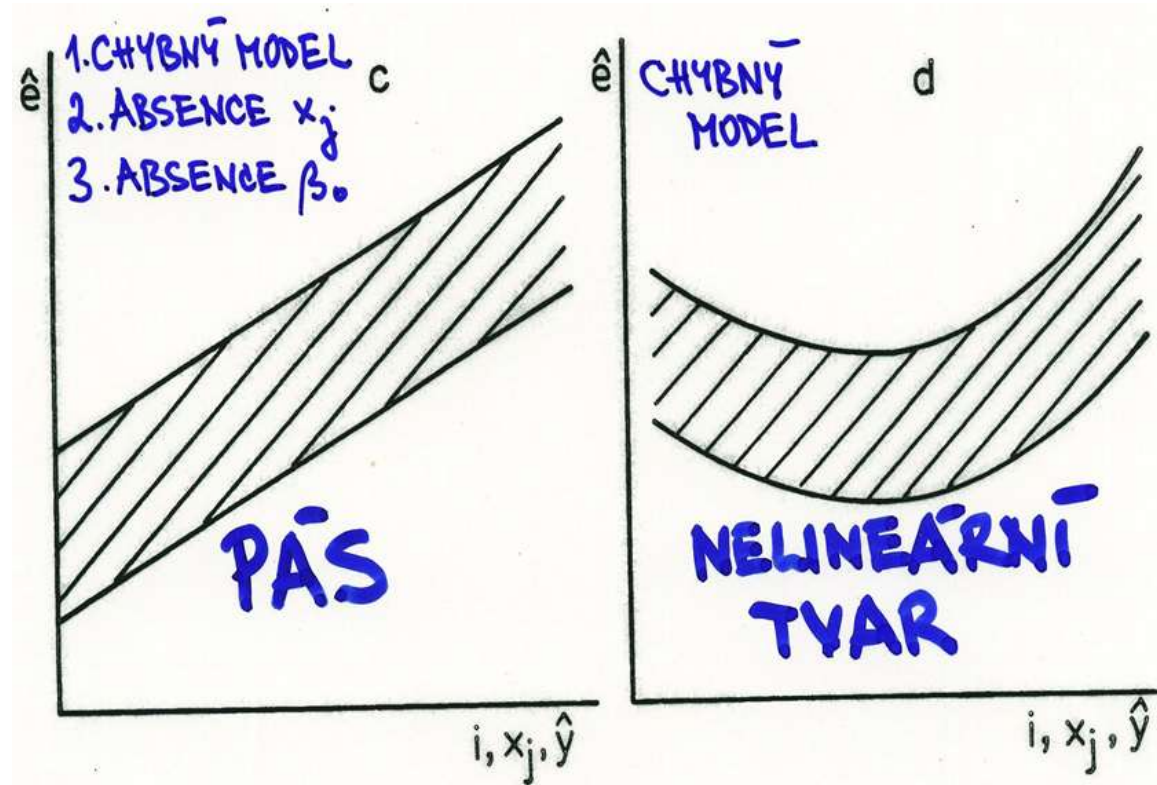
Typ III: Graf reziduí $\hat{e}_i$ proti predikci $\hat{y}_i$.





Obrazce v diagnostických grafech

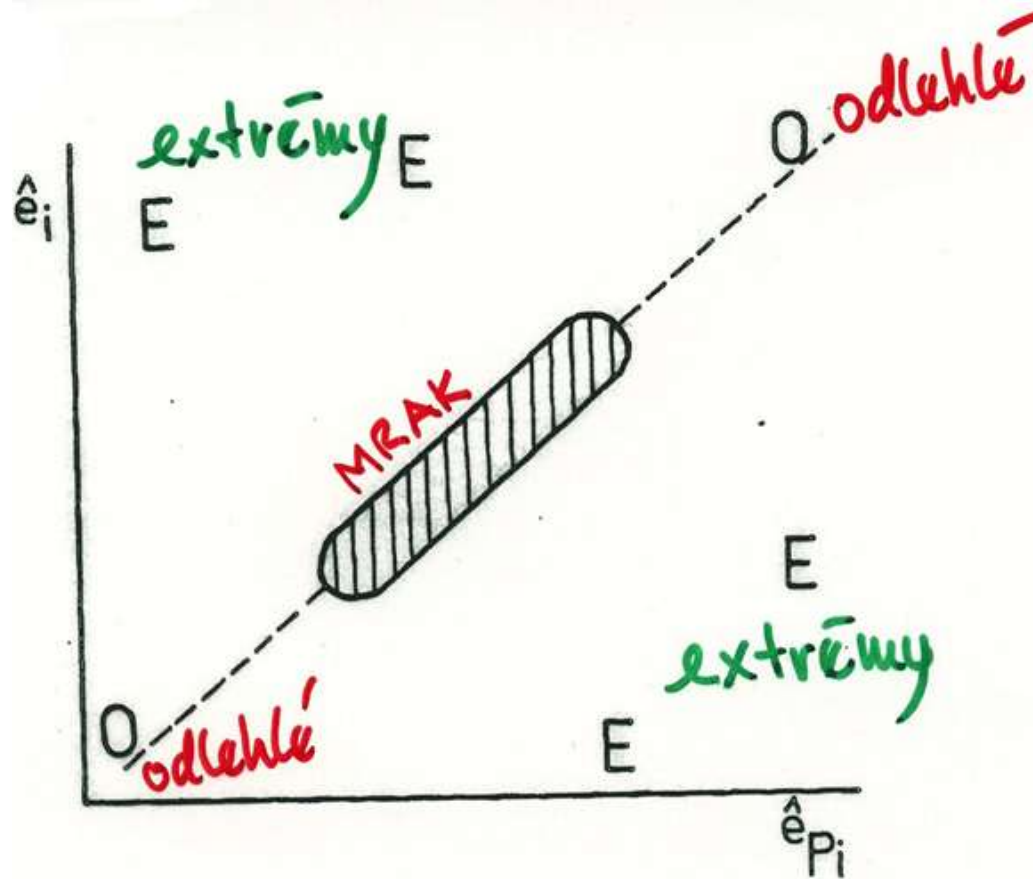
(a) tvar mraku, (b) tvar výseče, (c) tvar pásu a (d) nelineární tvar





Grafy identifikace vlivných bodů

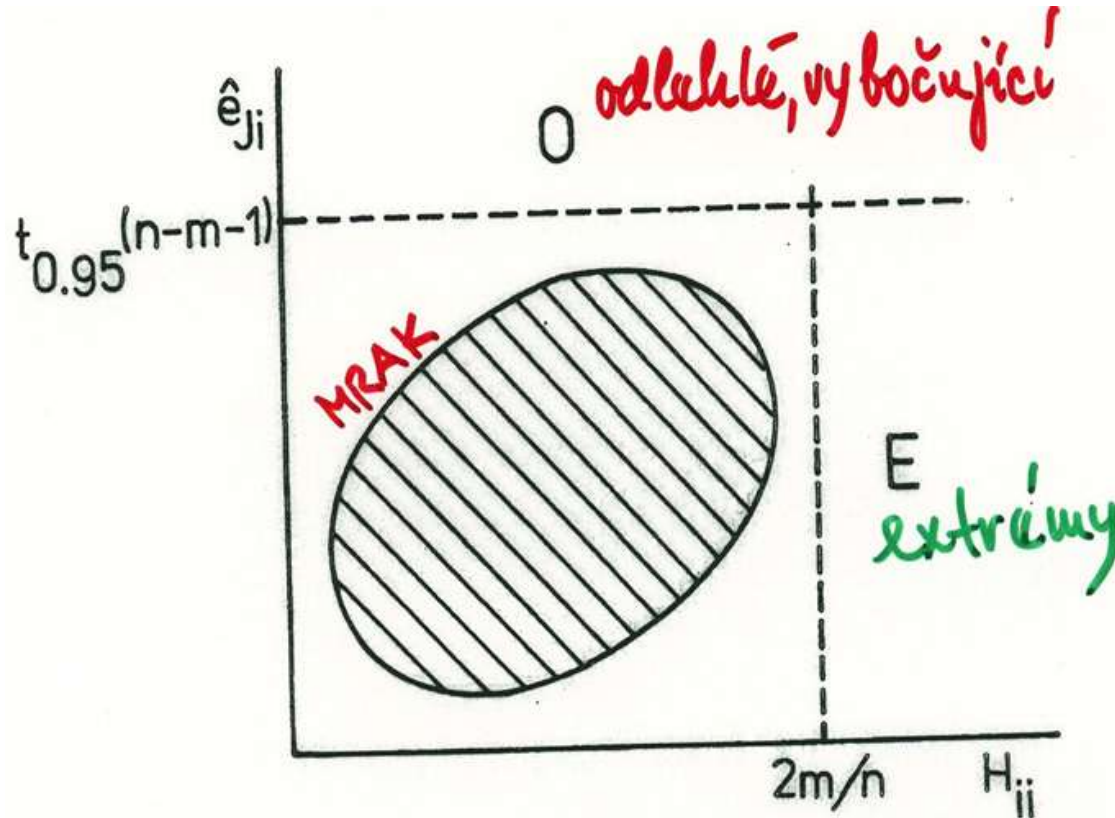
1. Graf predikovaných reziduí (GPR): osa x : $\hat{e}_{Pi}$, osa y : $\hat{e}_i$





Grafy identifikace vlivných bodů

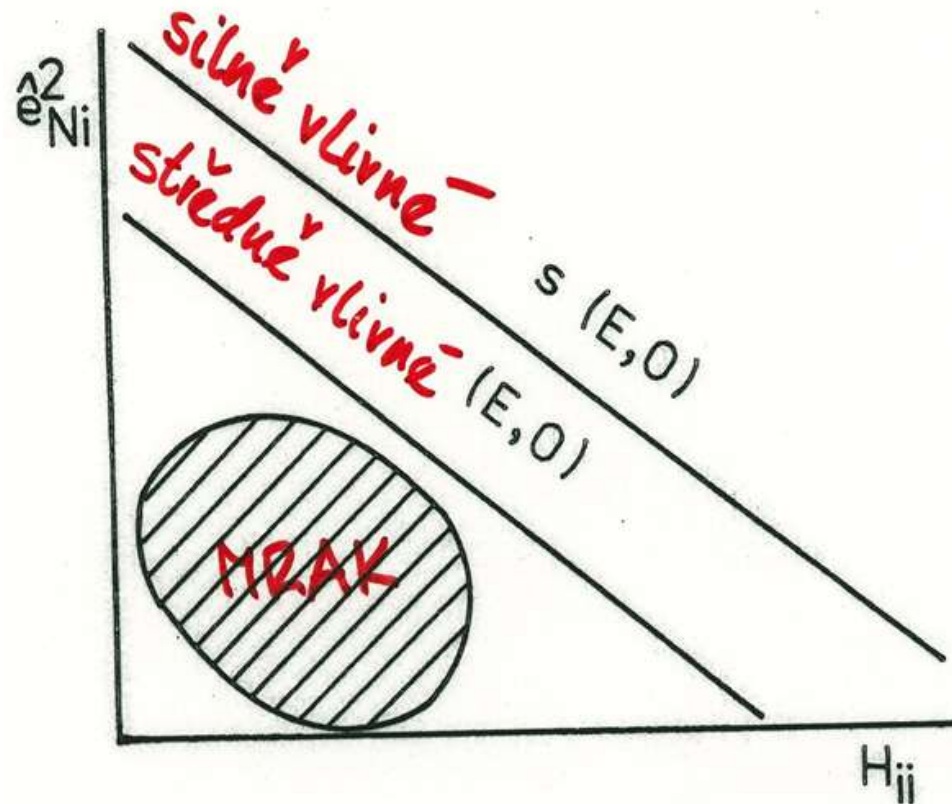
2. Williamsův graf (WG): osa x : prvky H_{ii} , osa y : $\hat{e}_{ji}$





Grafy identifikace vlivných bodů

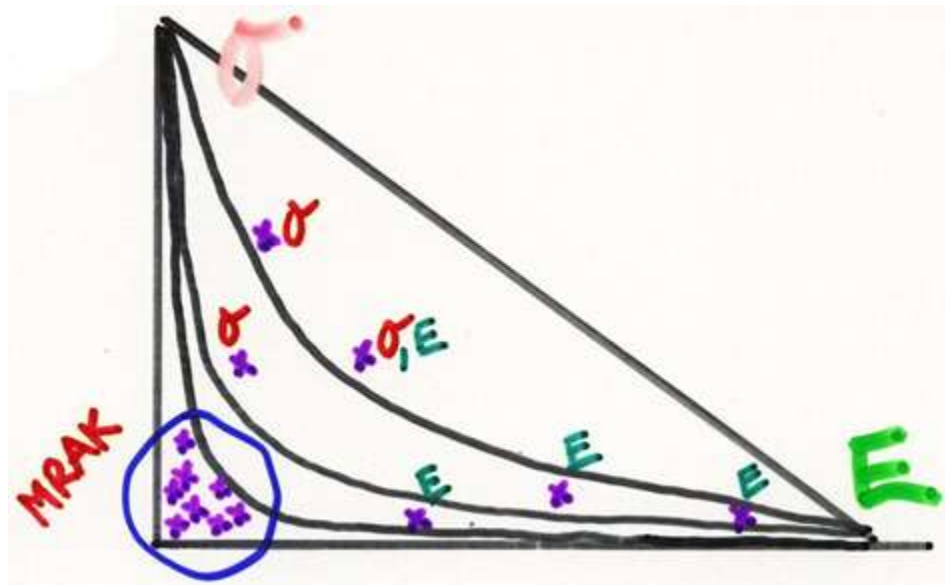
3. Pregibonův graf (PG): osa x : prvky H_{ii} , osa y : $\hat{e}_{Ni}^2$





Grafy identifikace vlivných bodů

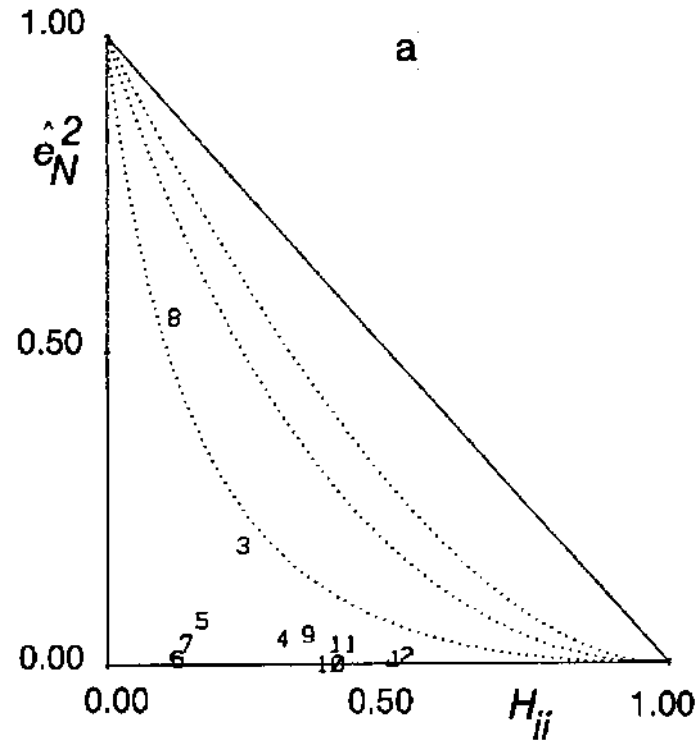
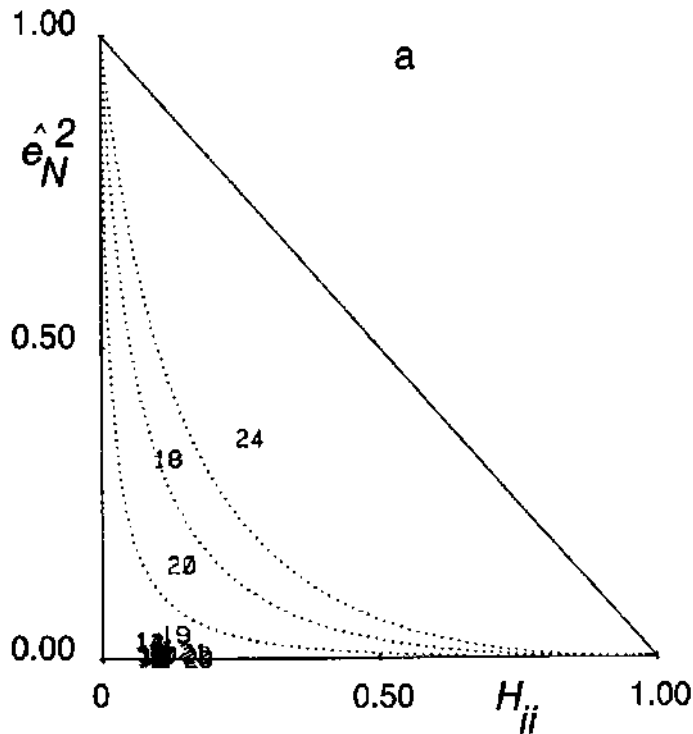
4. L-R graf (PG): osa x : $\hat{e}_{Ni}^2 = \hat{e}_i^2 / RSC$, osa y : prvky H_{ii}





Grafy identifikace vlivných bodů

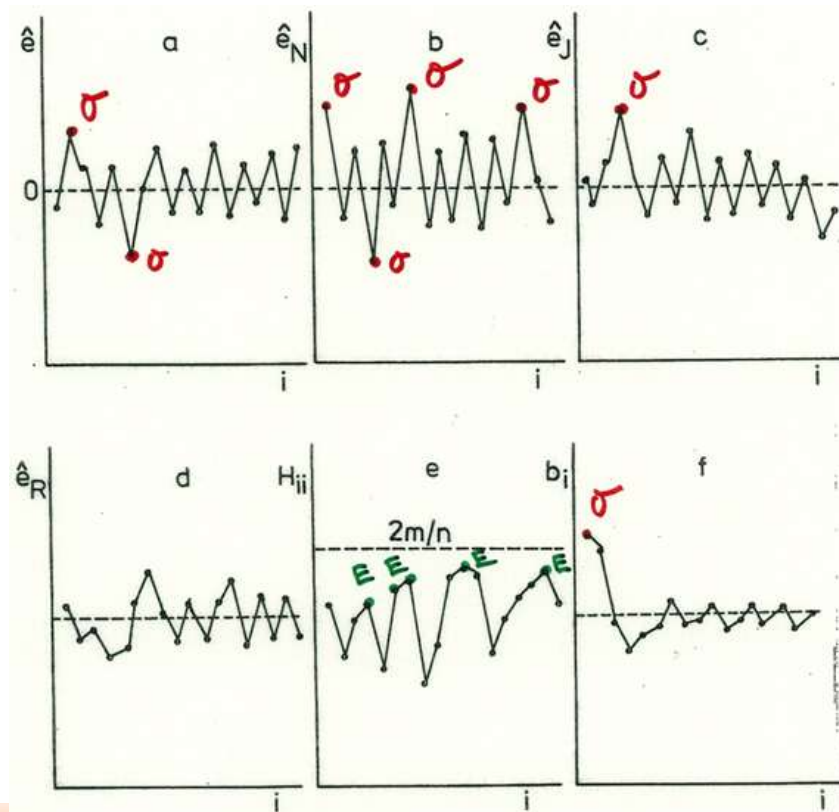
4. L-R graf (PG): osa x : $\hat{e}_{Ni}^2 = \hat{e}_i^2 / RSC$, osa y : prvky H_{ii}





Grafy identifikace vlivných bodů

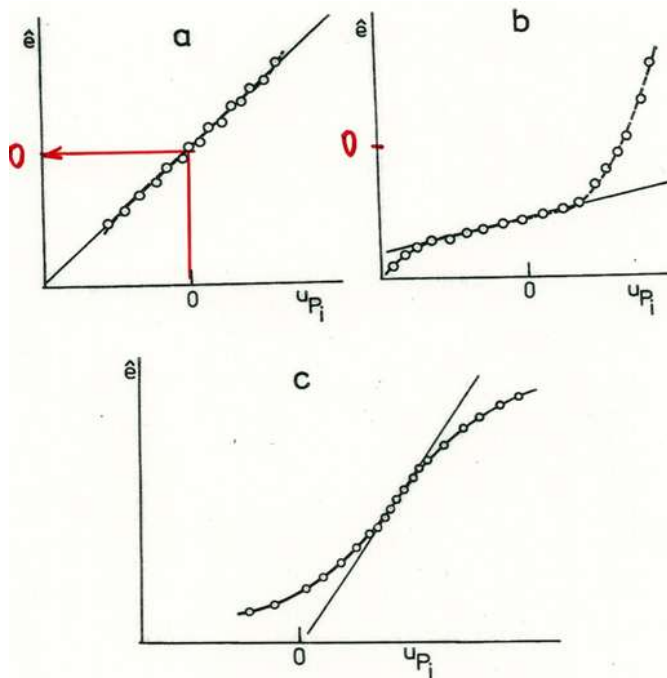
5. Indexové grafy (IG): osa x : index i , osa y : $\hat{e}_i$, $\hat{e}_{Si}$, $\hat{e}_{Ni}$, $\hat{e}_{Pi}$, $\hat{e}_{Ji}$, $\hat{e}_{Ri}$, nebo prvky H_{ii} , H_{ii}^* , a nebo odhady β_i





Grafy identifikace vlivných bodů

6. Rankitové grafy (Q-Q): osa x : u_{P_i} pro $P_i = i/(n + 1)$,
osa y : $\hat{e}_{(i)}$, $\hat{e}_{S(i)}$, $\hat{e}_{N(i)}$, $\hat{e}_{P(i)}$, $\hat{e}_{J(i)}$, $\hat{e}_{R(i)}$

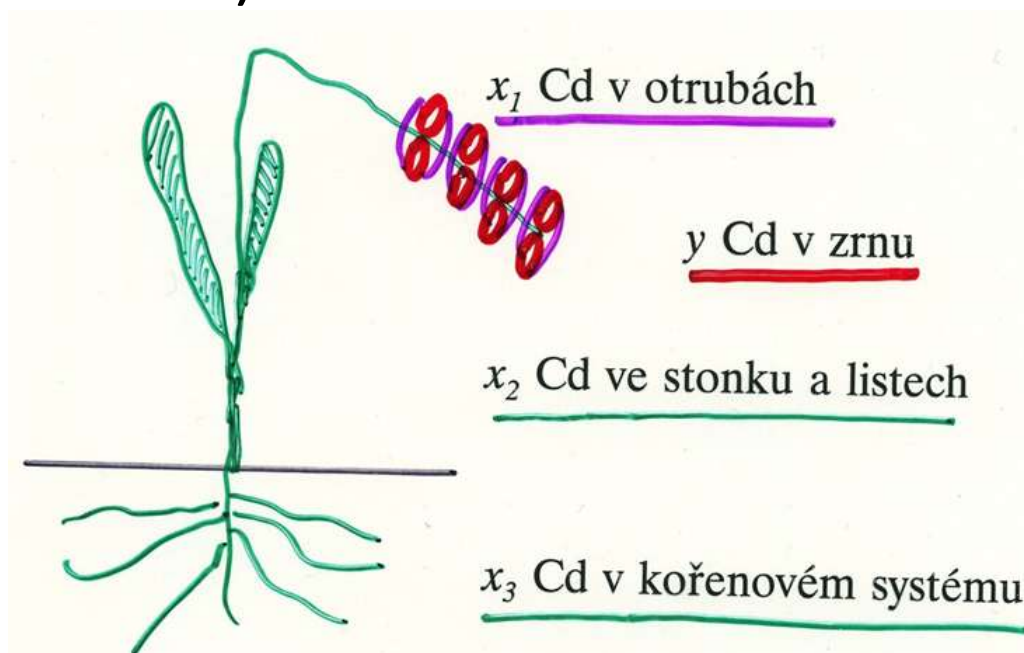


- (a) přibližně normální rozdělení
- (b) rozdělení s dlouhými konci
- (c) rozdělení s krátkými konci



Příklad M619 Vliv tří parametrů na obsah kadmia v potravinářské pšenici

Obsah kadmia v zrně y [mg/l] v závislosti na obsahu kadmia v otrubách x_1 [mg/l], ve stonku s listy x_2 [mg/l] a v kořenovém systému x_3 [mg/l]. Vyšetřete regresní triplet (data, model, metoda) a nalezněte lineární regresní model.

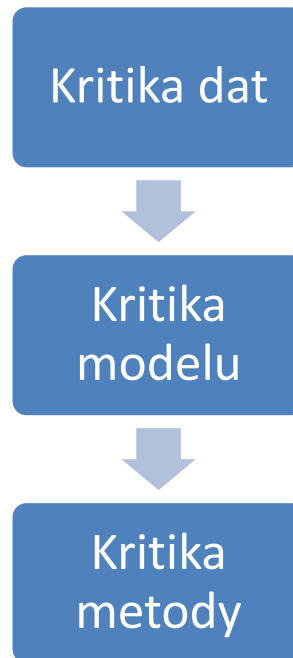




Pokračování příkladu

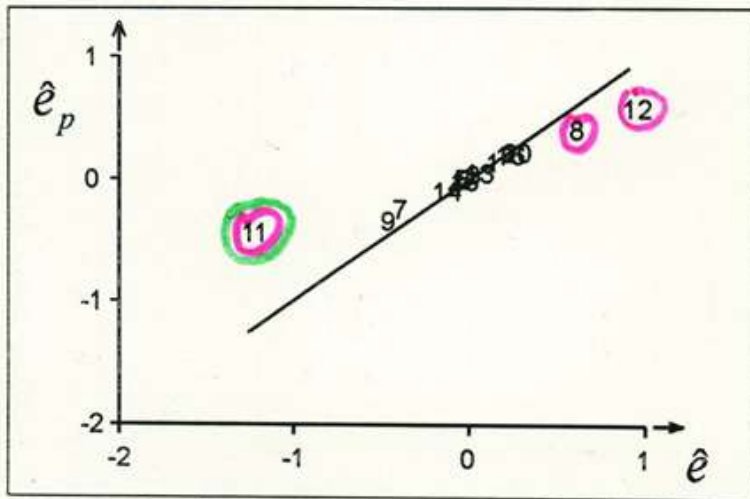
Výstavba lineárního regresního modelu

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

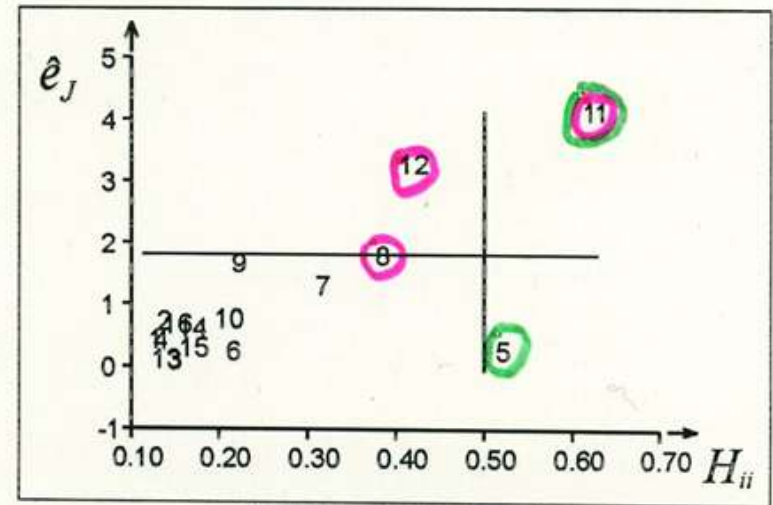




Graf vlivných grafů (odlehklých O a extrémů E)



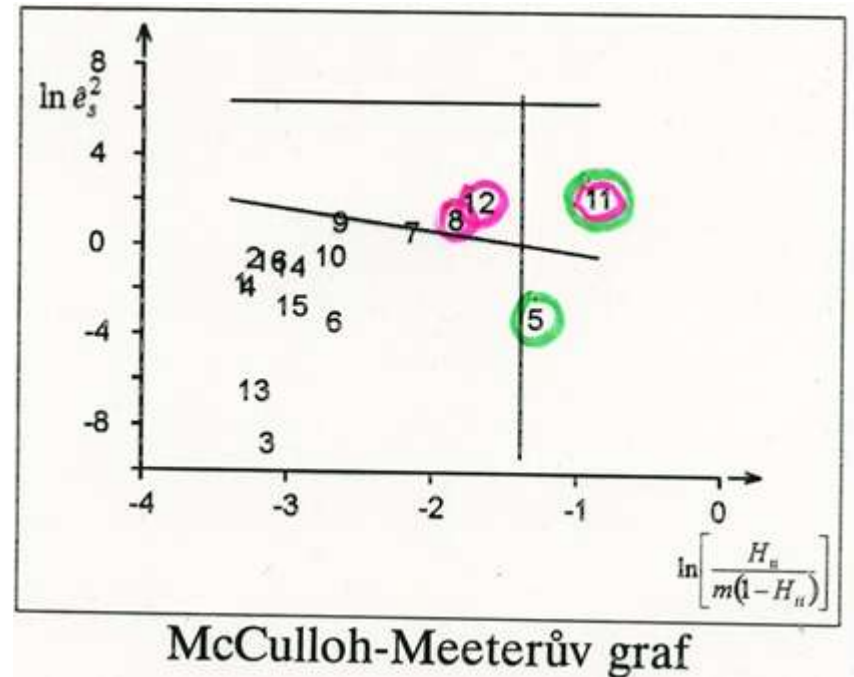
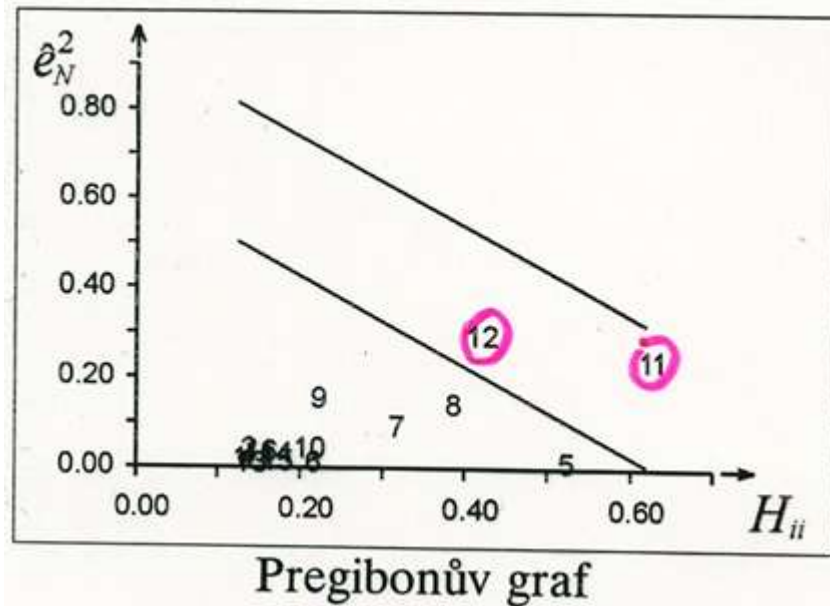
Graf predikovaných reziduí



Williamsův graf

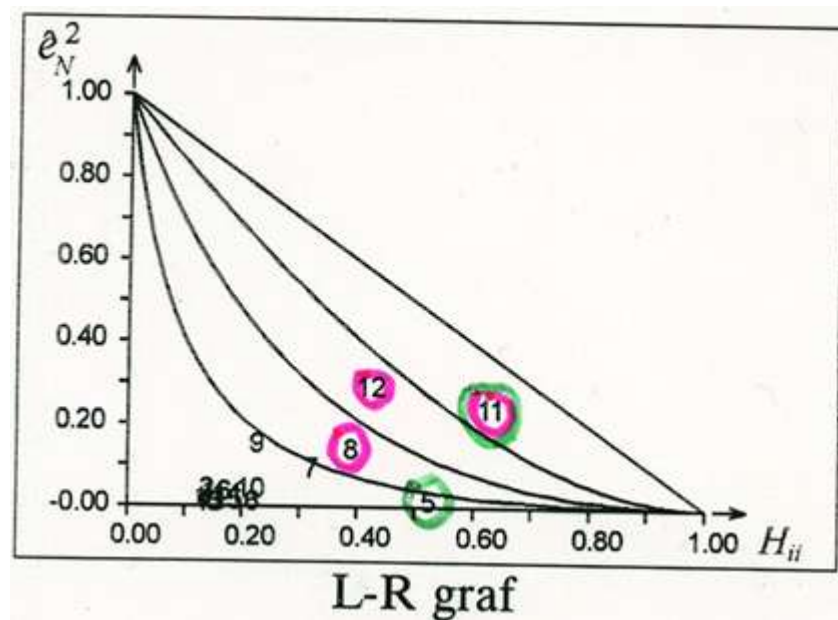


Graf vlivných grafů (odlehklých O a extrémů E)



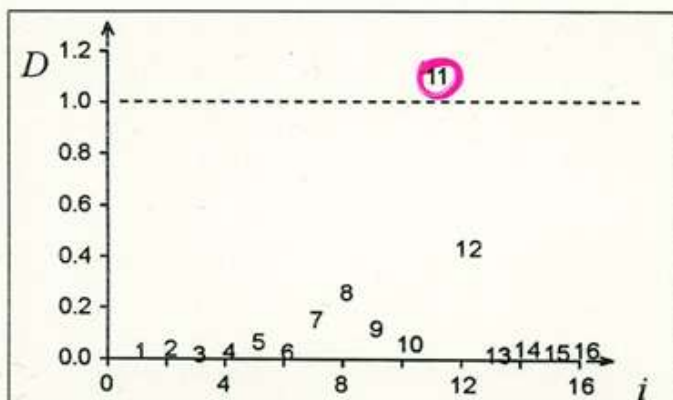


Graf vlivných grafů (odlehklých O a extrémů E)

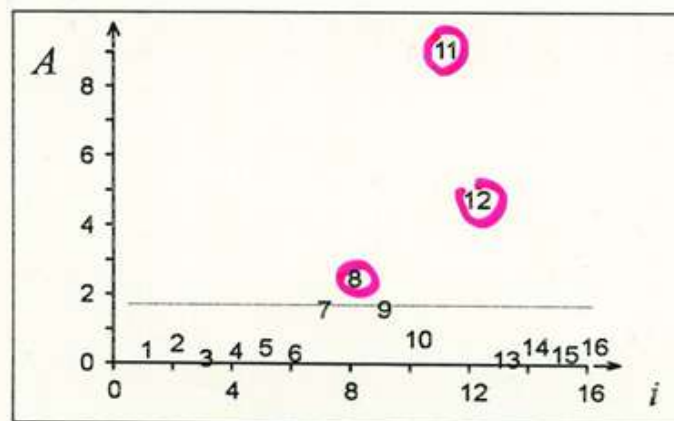




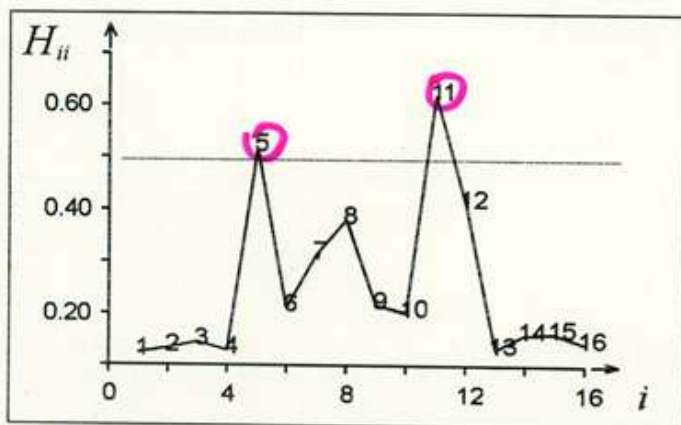
Vybrané regresní diagnostiky vlivných bodů



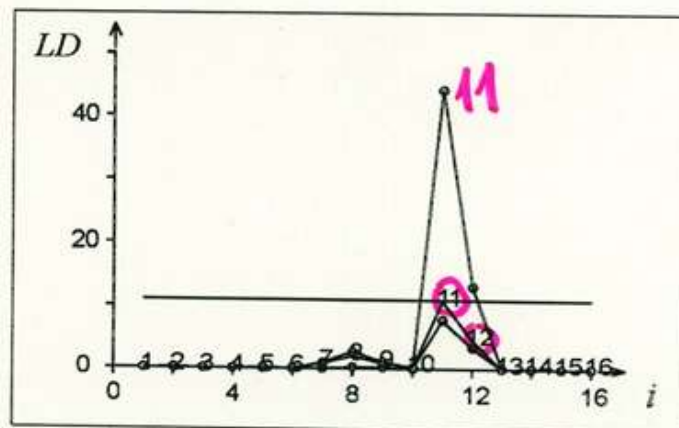
Indexový graf Cookovy vzdálenosti



Indexový graf Atkinsonovy vzdálenosti



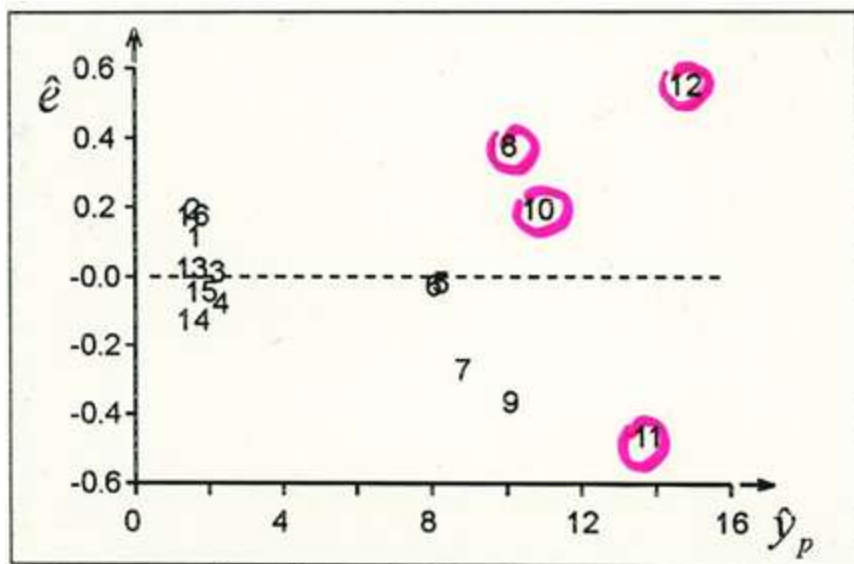
Indexový graf diagonálních prvků
H projekční matice



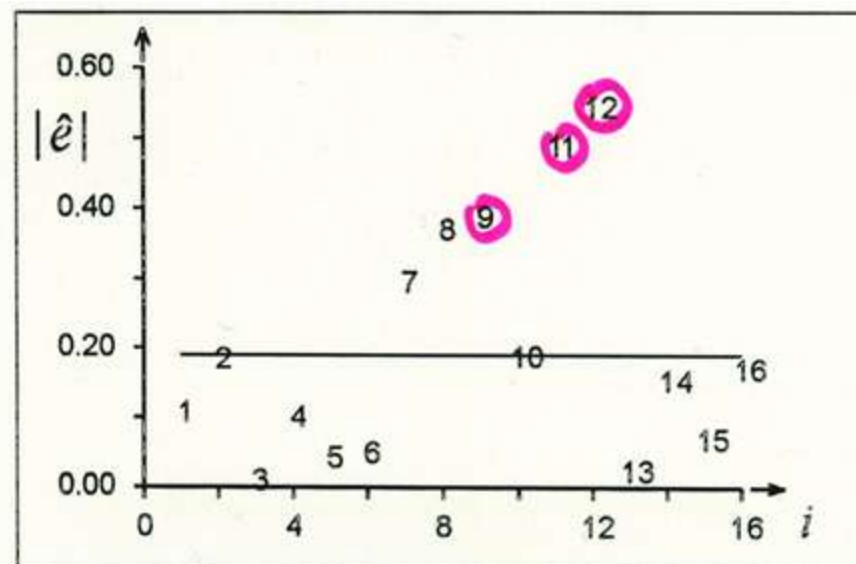
Graf věrohodnostních vzdáleností



Indexové a rozptylové grafy



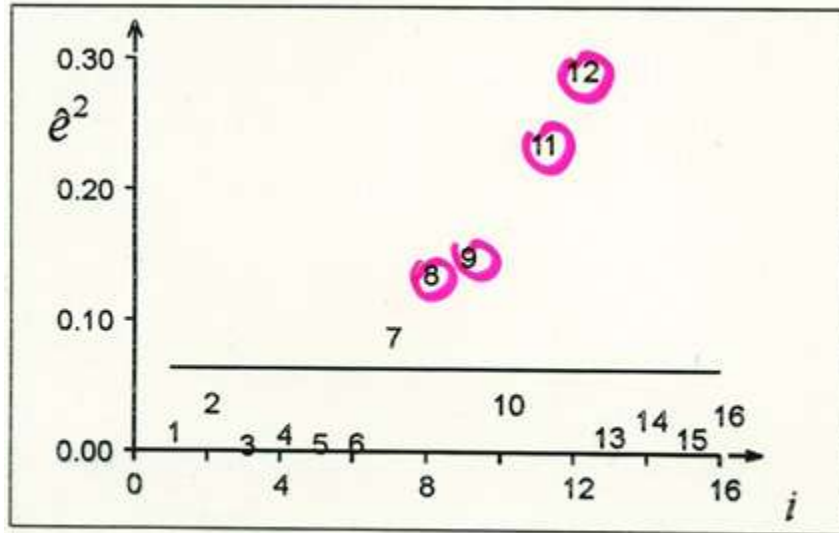
Rozptylový graf klasických reziduí



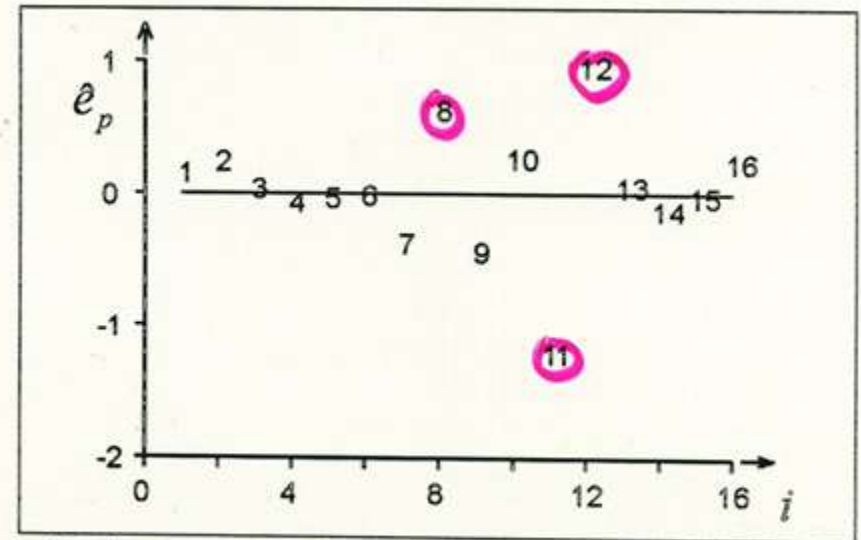
Indexový graf absolutních hodnot klasických reziduí



Indexové a rozptylové grafy



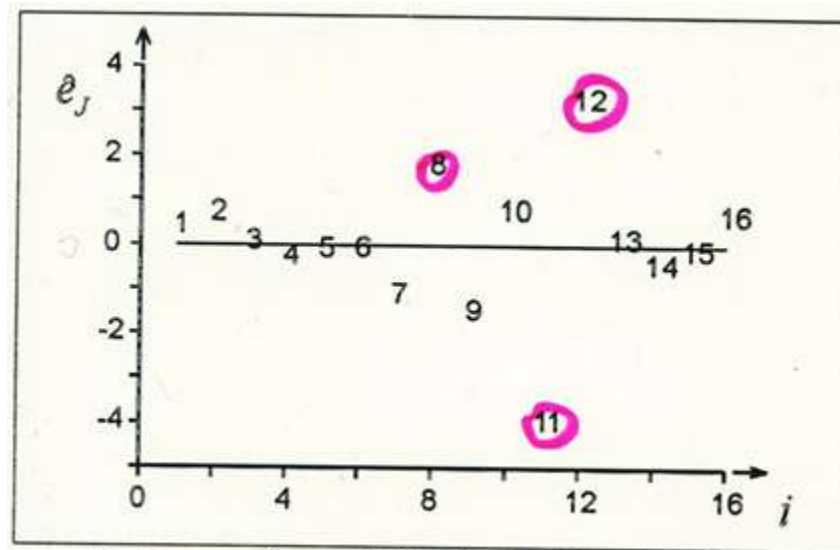
Indexový graf čtverců
klasických reziduí



Indexový graf predikovaných reziduí



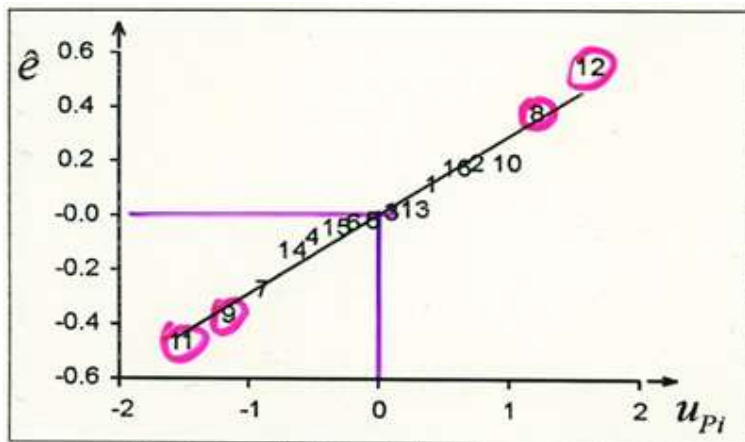
Indexové a rozptylové grafy



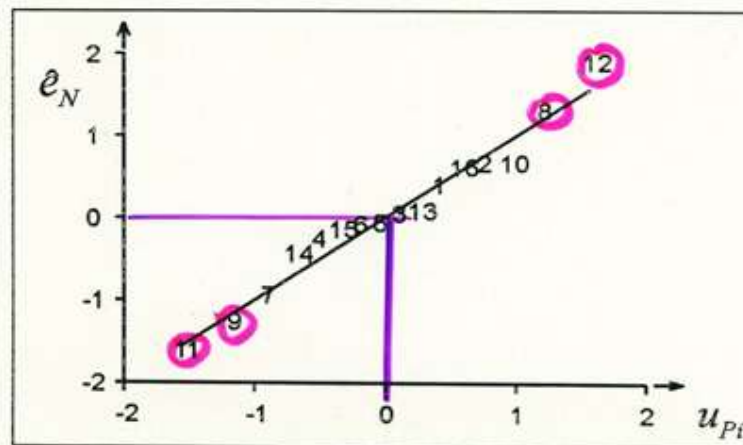
Indexový graf Jack-knife reziduí



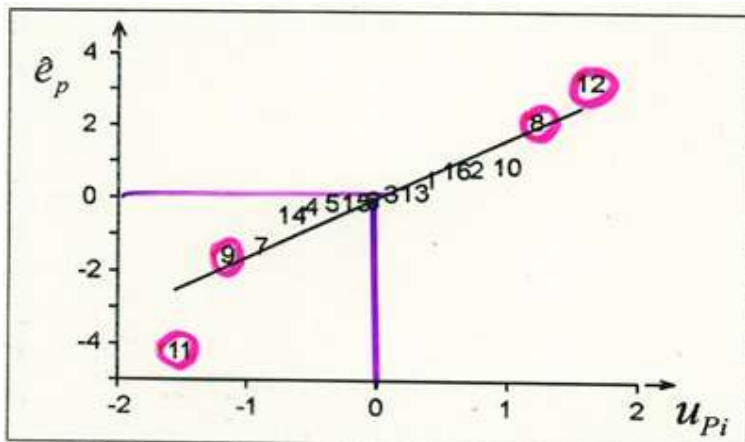
Rankitové Q-Q grafy, normalita rozdělení



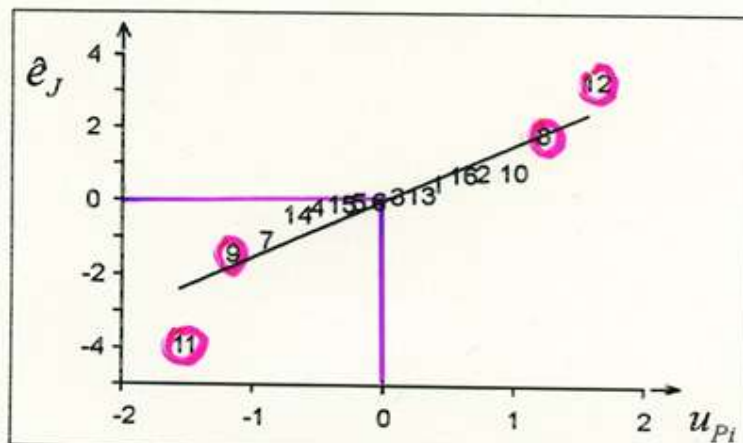
Q-Q graf klasických reziduí



Q-Q graf normovaných reziduí



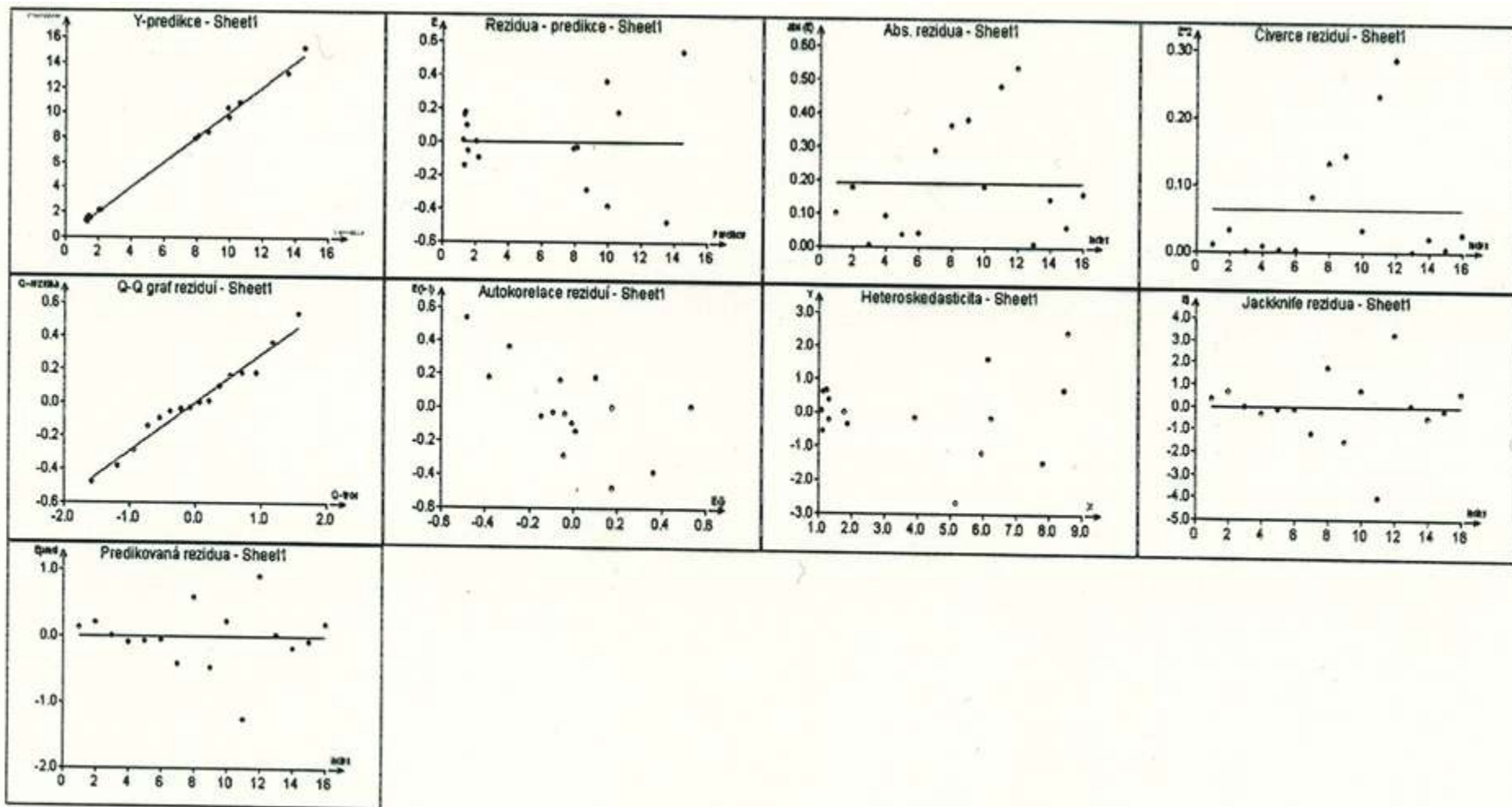
Q-Q graf predikovaných reziduí



Q-Q graf Jack-knife reziduí



Rezidua

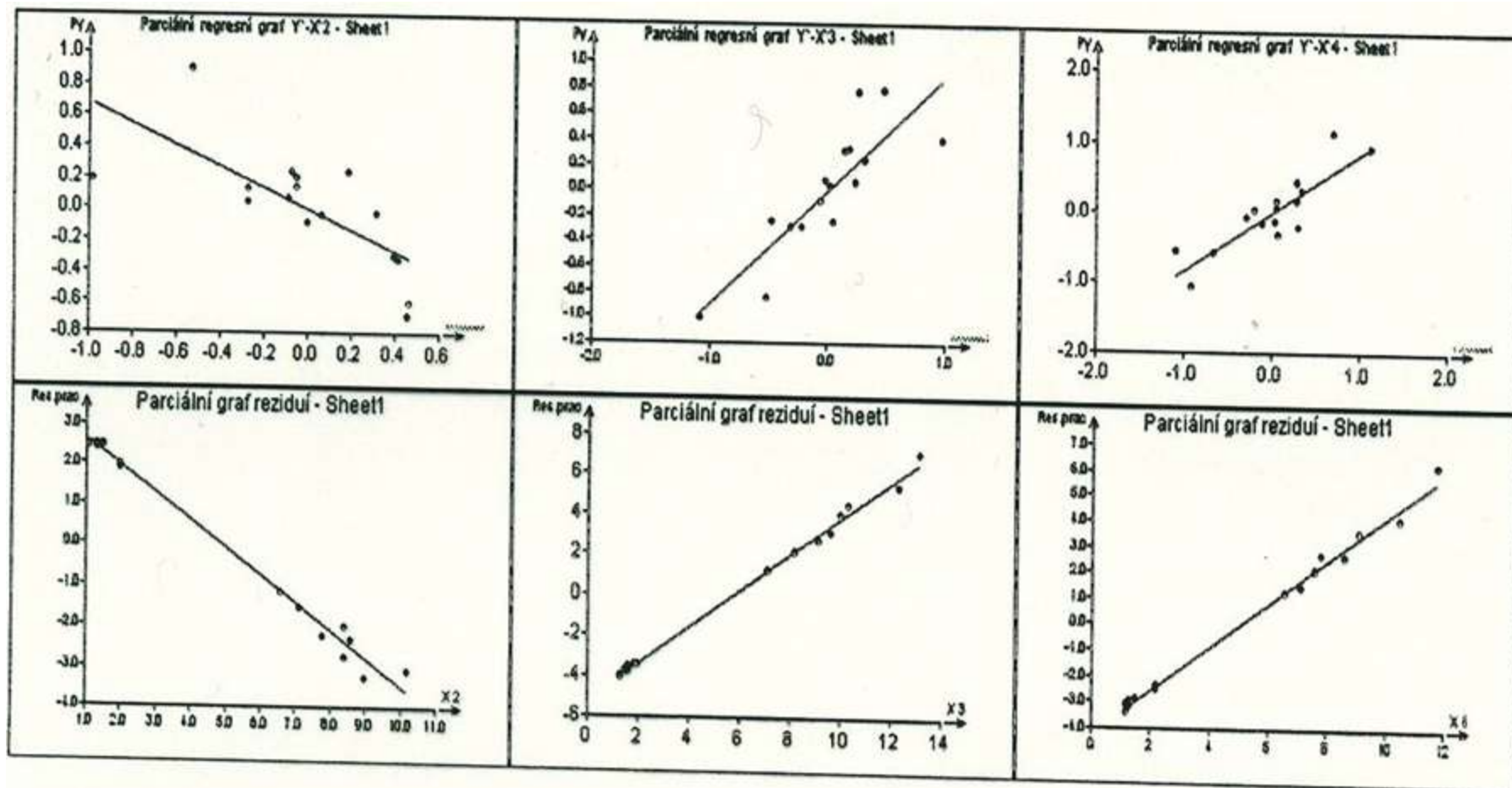


31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



Parciální grafy



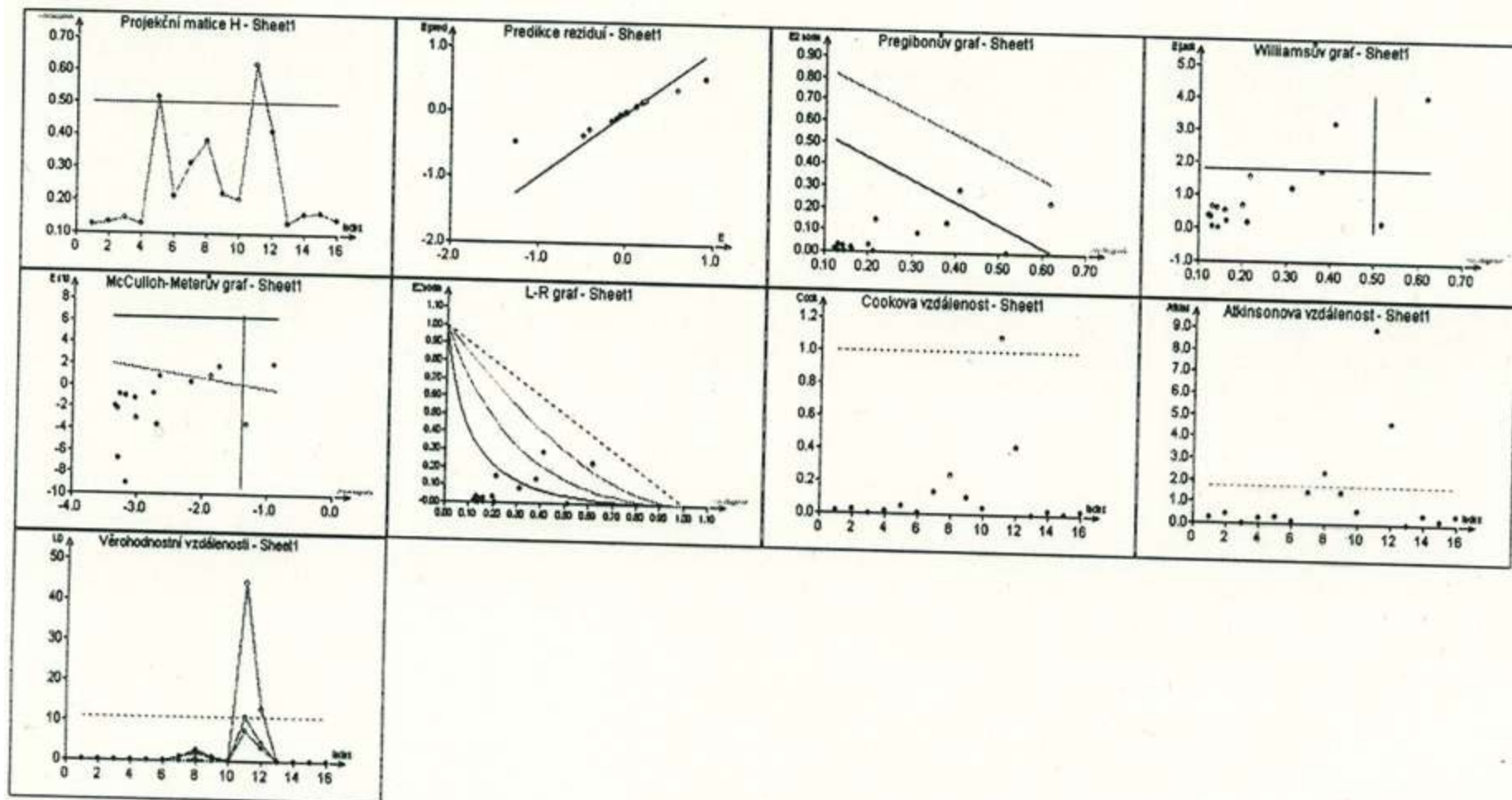
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Vlivná data



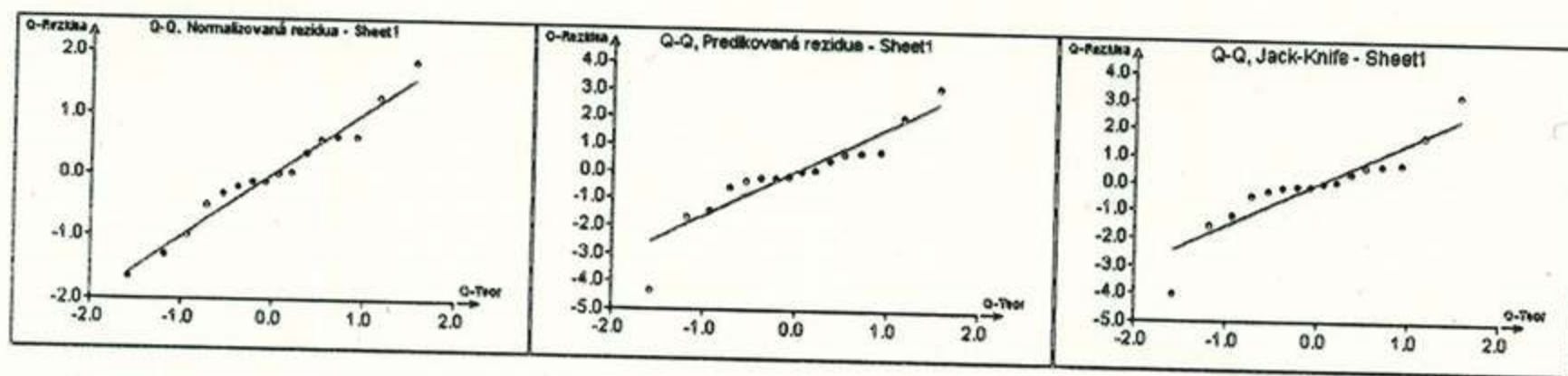
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Rankitové grafy



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



Lineární regrese, regresní diagnostika

Název: m619

V S T U P

(1) ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:

Omezení, P	: 1.0000E-34
Transformace	: Ne
Váhy	: Ne
Absolutní člen zahrnut	: Ano

(2) PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa	: 0.050
Počet bodů, n	: 16
Počet parametrů, m	: 3
Kvantil Studentova rozdělení $t(1-\alpha/2, n-m)$	: 2.179
Kvantil rozd. Chí-kvadrát $\text{Chi-square}(1-\alpha, m)$	: 9.488
Jméno výstupního souboru	: M619.txt



Pokračování příkladu

V Ý S T U P

(1) PŘEDBĚŽNÁ STATISTICKÁ ANALÝZA:

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina výz.
y	6.0125E+00	4.8734E+00	1.0000	-----
x1 OTRUBY	4.8937E+00	3.5692E+00	0.9837	0.000
x2 LISTY	5.7812E+00	4.5296E+00	0.9935	0.000
x3 KOŘEN	5.0812E+00	3.8782E+00	0.9948	0.000
Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných				Spočtená hladina výz.
x1 versus x2 :	9.9344E-01			0.000
x1 versus x3 :	9.8693E-01			0.000
x2 versus x3 :	9.8847E-01			0.000



Pokračování příkladu

(2) INDIKACE MULTIKOLINEARITY:

Č [j]	Vlastní čísla korel. matice $l[j]$	Čísla podmí- něnosti $K[j]$	Variance inflation factor VIF[j]	Vícenás.korel. koef pro $X[j]$
1	6.4568E-03	4.6141E+02	8.3272E+01	0.9940
2	1.4307E-02	2.0823E+02	9.4324E+01	0.9947
3	2.9792E+00	1.0000E+00	4.7508E+01	0.9894
Maximální číslo podmíněnosti K			: 4.6141E+02	
(K[j], $K > 1000$ indikuje silnou multikolinearitu)				
(VIF[j] > 10 indikuje silnou multikolinearitu)				



Pokračování příkladu

Kritika dat: původní data obsahují vlivné body 8, 11, 12

Kritika modelu $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$: technika MEP, AIC, $s(e)$) našla, že absolutní člen β_0 statisticky nevýznamný

Odhad parametru	Směrodatná odchylka	Student t-test $t_{0.95}(16-4) = 2.18$	Parameter je	Hladina významnosti
-----------------	---------------------	---	--------------	---------------------

A. Původní data a původní model vedou k závěru

$$y = -0.07(0.14) - 0.69(0.19)x_1 + 0.90(0.16)x_2 + 0.84(0.13)x_3$$

(LS: $t_{1-0.05/2}(16-4) = 2.179$, $D = 99.72\%$, $MEP = 0.21014$, $AIC = -36.18$, $s(e) = 0.290$)

β_0	-0.073	0.138	-0.5269	Nevýznamný	0.608
β_1	-0.685	0.192	-3.5746	Významný	0.004
β_2	0.896	0.161	5.5761	Významný	0.000
β_3	0.838	0.133	6.2879	Významný	0.000

B. Data bez 8, 11, 12 a původní model vedou k závěru

$$y = -0.02(0.10) - 0.80(0.50)x_1 + 0.92(0.34)x_2 + 0.91(0.15)x_3$$

(LS: $t_{1-0.05/2}(13-4) = 2.262$, $D = 99.83\%$, $MEP = 0.06118$, $AIC = -41.58$, $s(e) = 0.178$)

β_0	-0.017	0.103	-0.1637	Nevýznamný	0.874
β_1	-0.802	0.504	-1.5919	Nevýznamný	0.146
β_2	0.920	0.338	2.7193	Významný	0.024
β_3	0.906	0.152	5.9698	Významný	0.000



Pokračování příkladu

C. Data bez 11, 12 a opravený model vedou k závěru

$$y = -1.18(0.38)x_1 + 1.25(0.24)x_2 + 0.92(0.15)x_3$$

(LS: $t_{1-0.05/2}(14-3) = 2.201$, $D = 99.80\%$, $MEP = 0.06078$, $AIC = -43.47$, $s(e) = 0.193$)

β_0	0.000	---	---	---	---
β_1	-1.181	0.383	-3.0854	Významný	0.010
β_2	1.245	0.236	5.2751	Významný	0.000
β_3	0.917	0.151	6.091	Významný	0.000

D. Data bez 8, 11, 12 a opravený model vedou k závěru

$$y = -0.86(0.37)x_1 + 0.95(0.25)x_2 + 0.92(0.13)x_3$$

(LS: $t_{1-0.05/2}(13-3) = 2.228$, $D = 99.83\%$, $MEP = 0.05101$, $AIC = -43.55$, $s(e) = 0.170$)

β_0	0.000	---	---	---	---
β_1	-0.855	0.372	-2.3002	Významný	0.044
β_2	0.954	0.251	3.8038	Významný	0.003
β_3	0.916	0.132	6.9263	Významný	0.000

Model: $y = -0.86(0.37)x_1 + 0.95(0.25)x_2 + 0.92(0.13)x_3$



Pokračování příkladu

Numerická diagnostika vlivných bodů (odlehklých bodů a extrémů)
Odehlé body jsou tištěny tučně

i	y_i	$y_{P,i}$	$s(y_i)$	e_i	$e_{rel,i}$	$e_{S,i}$	$e_{L,i}$	$e_{P,i}$	H_{ii}	$Hm_{i,i}$	D_i	A_i	AP_i	DF_i	$LD(b)_i$	$LD(s^2)_i$	$LD(b, s^2)_i$
1	1.60	1.50	0.10	0.10	6.21	0.37	0.35	0.11	0.12	0.13	0.00	0.23	0.87	0.13	0.03	0.02	0.05
2	1.60	1.42	0.11	0.18	11.08	0.66	0.64	0.20	0.13	0.16	0.02	0.43	0.84	0.25	0.09	0.01	0.09
3	2.10	2.10	0.11	0.00	-0.14	-0.01	-0.01	0.00	0.14	0.14	0.00	0.01	0.86	0.00	0.00	0.03	0.03
4	2.10	2.19	0.10	-0.09	-4.41	-0.34	-0.33	-0.11	0.13	0.14	0.00	0.22	0.86	-0.13	0.02	0.02	0.05
5	8.10	8.14	0.21	-0.04	-0.44	-0.18	-0.17	-0.07	0.52	0.52	0.01	0.30	0.48	-0.17	0.04	0.03	0.07
6	7.90	7.94	0.13	-0.04	-0.52	-0.16	-0.15	-0.05	0.21	0.21	0.00	0.14	0.79	-0.08	0.01	0.03	0.04
7	8.40	8.69	0.16	-0.29	-3.42	-1.19	-1.22	-0.42	0.31	0.39	0.16	1.42	0.61	-0.82	0.84	0.03	0.95
8	10.30	9.94	0.18	0.36	3.52	1.59	1.71	0.59	0.38	0.51	0.39	2.32	0.49	1.34	1.94	0.25	2.69
9	9.60	9.98	0.14	-0.38	-3.96	-1.48	-1.57	-0.49	0.22	0.36	0.15	1.44	0.64	-0.83	0.80	0.16	1.10
10	10.80	10.62	0.13	0.18	1.66	0.69	0.68	0.22	0.20	0.23	0.03	0.59	0.77	0.34	0.16	0.00	0.16
11	13.10	13.58	0.23	-0.48	-3.67	-2.68	-4.06	-1.26	0.62	0.85	2.91	8.94	0.15	-5.16	10.86	7.83	44.18
12	15.10	14.57	0.19	0.54	3.55	2.40	3.19	0.91	0.41	0.69	1.01	4.62	0.31	2.67	4.63	3.44	13.13
13	1.30	1.29	0.11	0.01	0.71	0.03	0.03	0.01	0.13	0.13	0.00	0.02	0.87	0.01	0.00	0.03	0.03
14	1.20	1.34	0.12	-0.14	-12.01	-0.54	-0.53	-0.17	0.16	0.18	0.01	0.40	0.82	-0.23	0.07	0.01	0.08
15	1.50	1.56	0.12	-0.06	-3.98	-0.22	-0.22	-0.07	0.16	0.17	0.00	0.16	0.83	-0.10	0.01	0.03	0.04
16	1.50	1.34	0.11	0.16	10.74	0.60	0.58	0.19	0.14	0.17	0.02	0.41	0.83	0.24	0.08	0.01	10.09

Závisle proměnná
Predikce
Směrodatná odchylka predikce
Klasické reziduum
Relativní reziduum
Standardizované reziduum
Jack-knife reziduum
Predikované reziduum
Diagon. prvky H projekce
Modif. diagon. prvky H projekce
Cookova vzdálenost
Atkinsonova vzdálenost
Andrews-Pregibonova vzdálenost
Belseyho DF statistika
Věrohodnostní vzdálenost
Věrohodnostní vzdálenost
Věrohodnostní vzdálenost



MULTIKOLINEARITA

31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Multikolinearita

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

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

Vektory matice $\mathbf{X}$ musí být skutečně navzájem nezávislé (jejich párové R musí být nulové nebo statisticky nevýznamné). Pokud tomu tak není, dochází k **multikolinearitě**, která způsobuje početní i statistické problémy.



Vybrané předpoklady MNČ

1. Regresní parametry β mohou teoreticky nabývat **libovolných** hodnot.
2. Regresní model je **lineární v parametrech**.
3. Jednotlivé nezávislé proměnné jsou skutečně vzájemně nezávislé, tedy mezi nimi nedochází k tzv. **multikolinearitě**.
4. Podmíněný rozptyl $D(y/x) = \sigma^2$ je konstantní (tzv. podmínka **homoskedasticity**).
5. Náhodné chyby mají **nulovou střední hodnotu** $E(\varepsilon_i) = 0$, mají konečný rozptyl $E(\varepsilon_i^2) = \sigma^2$ a jsou nekorelované.



Multikolinearita – proč je „nebezpečná“

Početní problémy:

- způsobuje špatnou podmíněnost matice $X^T X$, (determinant této matice je nula nebo číslo blízké nule)
- potíže při invertaci matice (regresní model není jednoznačně řešitelný (singularita matice)).

Statistické problémy:

- nelze odděleně sledovat skutečný vliv jednotlivých vysvětlujících vstupních proměnných na vysvětlovanou (závislou) proměnnou.
- nespolehlivé určení parametrů regresního modelu (interval spolehlivosti parametrů je tak velký, že odhad parametrů ztrácí smysl).
- nestabilita odhadů regresních parametrů (např. malá změna hodnot závisle proměnné znamená zásadní změnu parametrů)



Multikolinearita - příčiny

- **Přeurčenost** regresního modelu („zbytečně“ mnoho nezávislých proměnných),
- Skutečně existující **závislost mezi „nezávislými“** proměnnými,
- **Povaha** modelu (např. polynom),
- **Nevhodné rozmístění** experimentálních bodů, (např. malá variabilita hodnot nezávisle proměnné).



Multikolinearita

M. neznamená porušení předpokladů MNČ

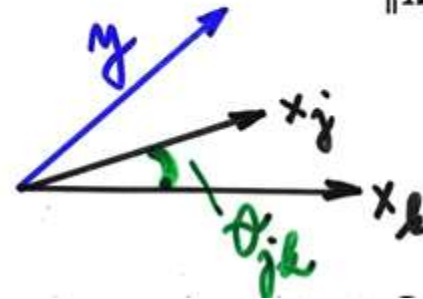
M. souvisí pouze s pozitivní definitností matice $X^T X$

Dle úhlu Θ_{jk} mohou nastat dva krajní případy:

1. **Ortogonalita**, kdy je kosinus Θ_{jk} nulový, ($\cos \Theta_{jk} = 0$)



$$\cos \Theta_{jk} = \frac{\langle \mathbf{x}_j, \mathbf{x}_k \rangle}{\|\mathbf{x}_j\| \cdot \|\mathbf{x}_k\|} = 0$$



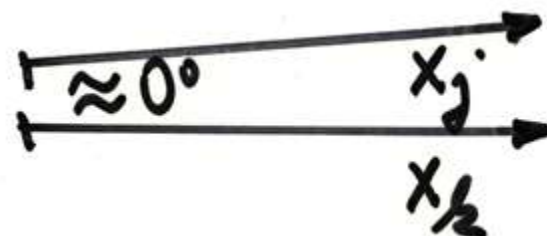
t. zn. že skalární součin $\langle x_j, x_k \rangle = 0$, a tím úhel Θ_{jk} mezi vektory x_j a x_k je 90° .



Multikolinearita

Jsou-li všechny sloupce matice X vzájemně ortogonální, je $X^T X$ diagonální matice.

2. **Kolinearita**, kdy je kosinus Θ_{jk} roven 1, ($\cos \Theta_{jk} = 1$), protože úhel Θ_{jk} mezi vektory x_j a x_k je nulový $\Theta_{jk} = 0$.



Pak jsou oba vektory x_j a x_k rovnoběžné, čili lineárně závislé $c_j x_j + c_k x_k = 0$ a c_j a c_k jsou nenulové konstanty.

Platí-li tato rovnice pro q dvojic sloupců matice X , je její hodnost rovna rozdílu $(m - q)$ a matice $X^T X$ je singulární.

31.1.2011



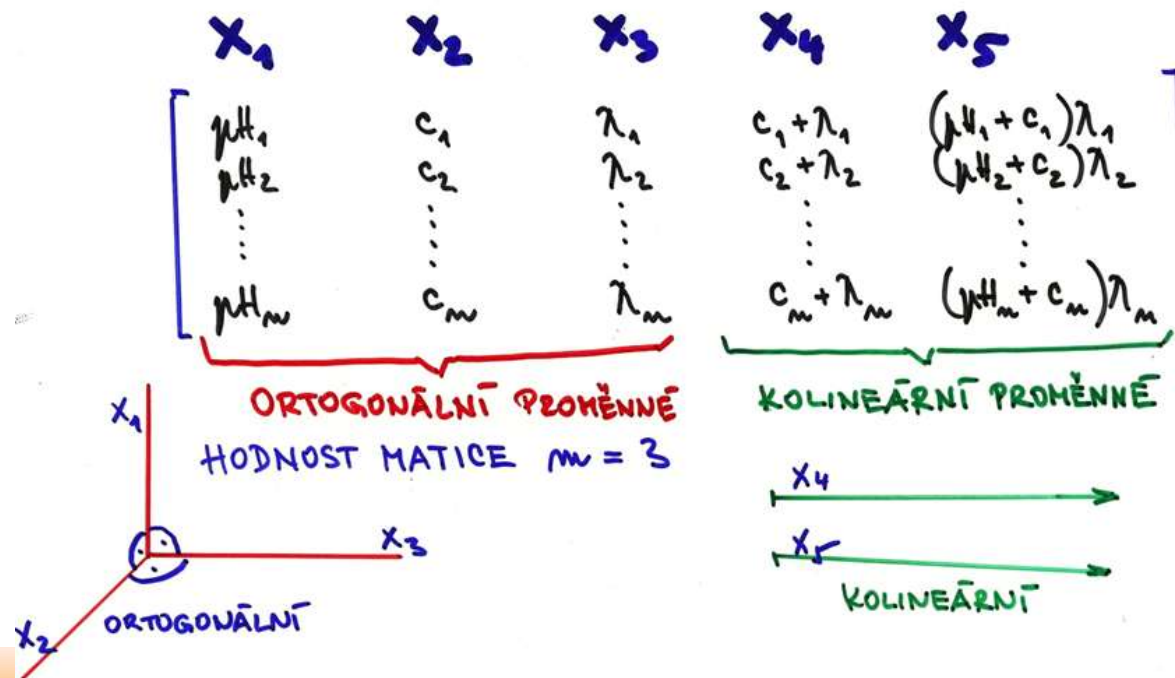
Základní příčiny vzniku

1. **Přeurčený regresní model** obsahuje nadměrný počet nezávislých proměnných, které vyjadřují stejné faktory.
2. **Nevhodné rozmístění bodů**: hodnoty nezávisle proměnné kolísají jen v malém rozmezí a jsou proto kolineární s vektorem odpovídajícím absolutnímu členu.
3. **Fyzikální omezení v modelu nebo v datech**, kdy vznikají vazby mezi nezávisle proměnnými přímo ve studovaném systému.



Test multikolinearity

Paradoxní situace: T-test je významný a všechny t-testy jsou nevýznamné, protože je silná multikolinearita mezi sloupci matice X , čili existuje rovnoběžnost vektorů x_j a $x_k, j \neq k$ sloupců matice X .





Statistické obtíže

1. **Nestabilita odhadů** je způsobena citlivostí odhadů na malé změny v datech. Odhady mívají často nesprávné znaménko, což znemožňuje jejich věcnou (fyzikální) interpretaci a jsou do absolutních hodnot příliš velké.
2. **Velké rozptyly** $D(b_j)$ jednotlivých odhadů způsobují, že t-testy indikují statistickou nevýznamnost β_j .
3. **Silná korelovanost** mezi prvky vektoru odhadů b způsobuje, že odhady b_j nelze interpretovat odděleně.
4. **Koeficient determinace vysoký** a regresní model může dobře popisovat experimentální data.



Metoda racionálních hodnotí

Matici R (symetrická) vyjádříme

- pomocí vlastních čísel $\lambda_1 \leq \lambda_1 \leq \dots \leq \lambda_m$
- a odpovídajících vlastních vektorů $P_j, j = 1, \dots, m$, ve tvaru

$$R = \sum_{j=1}^m \lambda_j P_j P_j^T$$

a inverzní matici R^{-1}

$$R^{-1} = \sum_{j=1}^m \lambda_j^{-1} P_j P_j^T$$



Metoda racionálních hodností

a vektor normovaných odhadů parametrů bude

$$b_N = \sum_{j=\omega}^m (\lambda_j^{-1} P_j P_j^T) r$$

a kovarianční matice normovaných odhadů bude

$$D(b_N) = \hat{\sigma}_N^2 \sum_{j=\omega}^m \lambda_j^{-1} P_j P_j^T$$

V případě MNČ bude $\omega = 1$.

Platí důsledek: pokud budou vlastní čísla λ_j malá, budou odhady b_N i jejich rozptyly neúměrně vysoké.



Scottova testační charakteristika

Používá se k posouzení stupně multikolinearity:

$$M_T = \frac{\frac{F_R}{T_S} - 1}{\frac{F_R}{T_S} + 1}$$

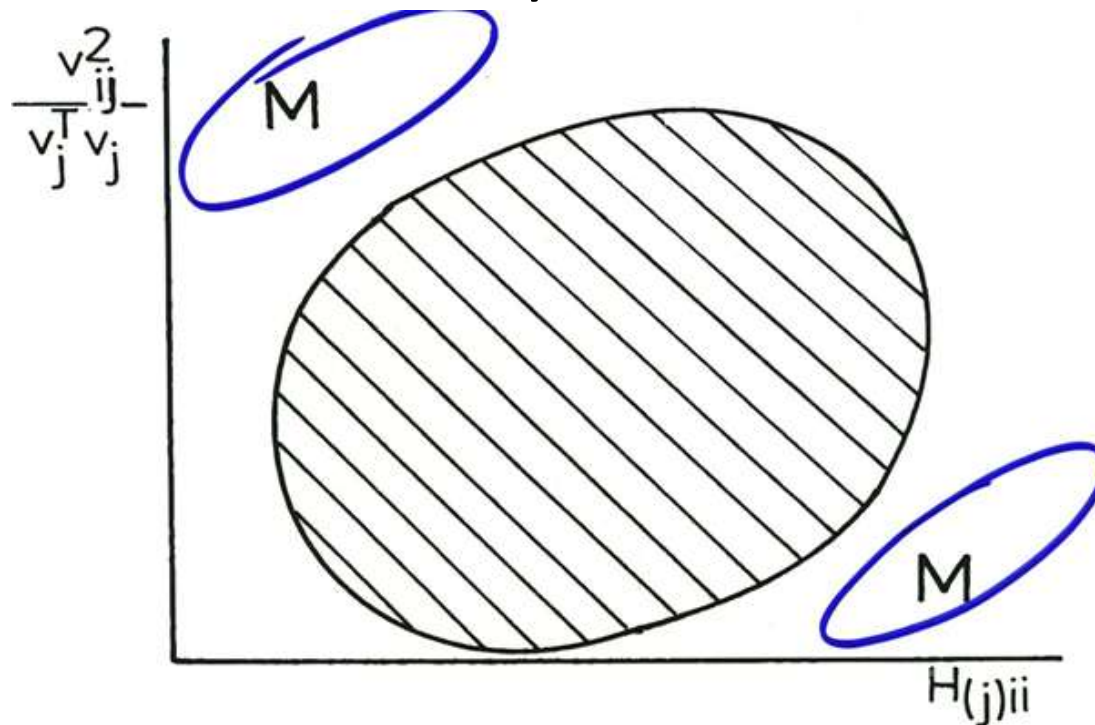
- a) $M_T > 0.8$, model je nevyhovující a je zapotřebí provést úpravu.
- b) $0.33 \leq M_T \leq 0.8$, model je málo vyhovující a je vhodná jeho úprava
- c) $M_T < 0.33$, není model ovlivněn multikolinearitou a není třeba ho upravovat.



Diagnostika

- 1. Grafická analýza:** vlivné body (multikolinearitu) odhalíme vynesáním $v_{ij}^2 / (v_j^T v_i)$ proti $H_{(j)ii}$

Diagnostika multikolinearity: M značí multikolinearitu





2. Numerická kritéria:

- a) **Determinant matice R** , $\det(R) = \prod_{j=1}^m \lambda_j$, kde λ_j jsou vlastní čísla matice R . Je-li determinant $\det(R)$ příliš malý, tj. menší než 10^{-3} , jde o silnou multikolinearitu.
- b) **Číslo podmíněnosti K** $= \frac{\lambda_{max}}{\lambda_{min}}$, kde λ_{max} , λ_{min} jsou maximální a minimální vlastní číslo matice R . Je-li číslo podmíněnosti $K > 10^3$, jde o silnou multikolinearitu.



c) **VIF-faktor** (Variance Inflation Factor) je $VIF_j = \tilde{R}_{jj}$, kde $\tilde{R}_{jj}$ je j-tý diagonální prvek matice R^{-1} . Platí vztah

$VIF_j = \frac{1}{1 - \hat{R}_{x_j}^2}$. Je-li $VIF_j > 10$, jde o silnou multikolinearitu.



Multikolinearita – testování

VIF – variance inflation factor – diagonální prvky
inverzní matice ke korelační matici nezávisle
proměnných ($\text{diag}(R^{-1})$)



Multikolinearita – testování

	A	B	C	D	E	F
1		X1	X2	X3	X4	X5
2	X1	1	0.23	-0.15	0.07	0
3	X2	0.23	1	0.08	0.25	0.34
4	X3	-0.15	0.08	1	0.73	0.67
5	X4	0.07	0.25	0.73	1	0.98
6	X5	0	0.34	0.67	0.98	1
7						
8						
9		X1	X2	X3	X4	X5
10	X1	2.25	-1.28	1.51	-10.2	9.44
11	X2	-1.28	2.15	-0.88	9.05	-9.03
12	X3	1.51	-0.88	3.38	-11.3	9.15
13	X4	-10.2	9.05	-11.3	89.6	-83.5
14	X5	9.44	-9.03	9.15	-83.5	79.9

VIF > 10 ⇒ kritická multikolinearita

korelační matice R

=INVERZE(B2..F6)

Ctrl+Shift+Enter

inverzní matice R⁻¹

kriticky vysoké hodnoty VIF



Podle velikosti vlastních čísel λ_j

- I. skupina regresních úloh: všechna vlastní čísla $\lambda_j \geq 0$ a MNČ nečiní žádné obtíže.
- II. skupina regresních úloh: některá vlastní čísla $\lambda_j \approx 0$ jsou blízka nule (multikolinearita) a běžné metody (MNČ) zcela selhávají.
- III. skupina regresních úloh: některá vlastní čísla $\lambda_j = 0$ jsou rovna nule. Pak je matice $X^T X$ nebo R singulární a nelze ji invertovat.



Řešení problémů II. a III.

Užitím racionálních hodnot

- zanedbávají se sčítance (resp. jejich části) o malých hodnotách vlastních čísel λ_j
- kritériem pro vypuštění sčítanců malých vlastních čísel je vztah

$$abs \left(\frac{\sum_{j=1}^{\omega} \lambda_j}{\sum_{j=1}^m \lambda_j} \right) = P$$

kde

- P je zvolená přesnost (obvyčejně)
- číslo ω určuje spodní mez, od které se norm. odhadů provádí sčítání.



Řešení problémů II. a III.

Označme $W = \sum_{j=1}^{\omega} \lambda_j$ a $E = \sum_{j=1}^m \lambda_j$ ($\frac{W}{E} > P$)

Řešení:

- pokud je $\frac{W}{E} > P$ (tj. by mělo být ω necelé), provádí se sumace od $(\omega - 1)$ a vlastní číslo $\lambda_{\omega-1}$ se „váží“ faktorem

$$u = \frac{W - EP}{\lambda_{\omega}}$$

- tím je zajištěno, že lze spojitě v závislosti na růstu přesnosti P snižovat délku odhadů $\|b_N\|$ a jejich rozptyly.



Řešení problémů II. a III.

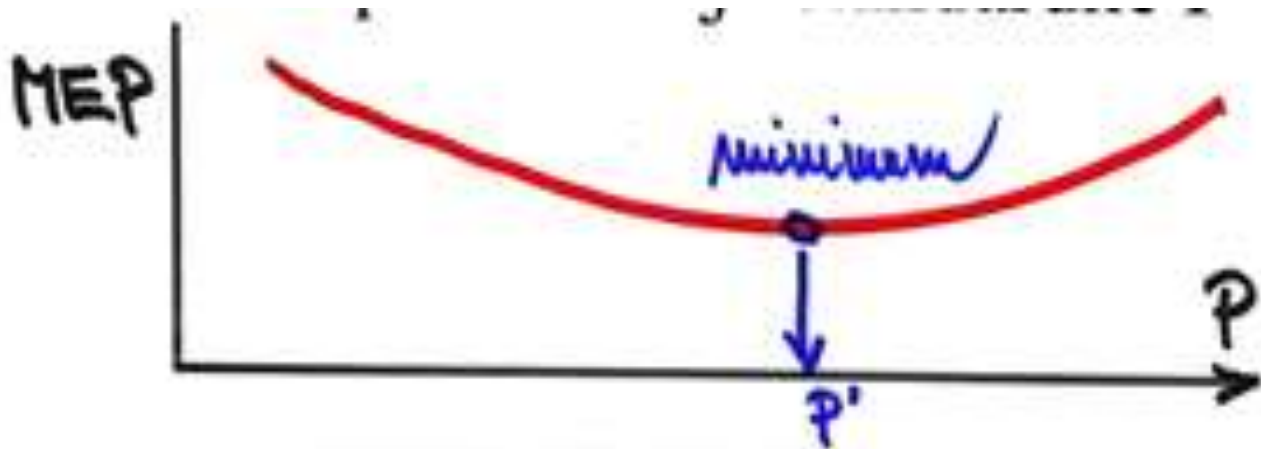
- to je však doprovázeno růstem vychýlení odhadů a poklesem vícenásobného korelačního koeficientu.
- vychýlení odhadů je zde způsobeno zanedbáním sčítanců v rovnicích normovaných odhadů parametrů při $\omega > 1$.
- pro čtverec vychýlení $h_v^2(b_N) = [\beta - E(b)]^2$, získaný metodou racionálních hodnot platí

$$h_v^2(b_N) = \beta_N^T \left(\sum_{j=1}^{\omega} P_j P_j^T \right) \beta_N$$



Řešení problémů II. a III.

- optimální velikost P : z požadavku minima MEP v závislosti na P . Jinak pro MNČ je standardně $P = 10^{-32}$.





Příklad 6.9 Aproximace absorpčního spektra polynomem

Popište závislost molárního absorpčního koeficientu ε na vlnové délce λ , $\varepsilon = f(\lambda)$ polynomem druhého stupně

$$E\left(\frac{\varepsilon}{\lambda}\right) = \beta_2\lambda + \beta_3\lambda^2 + \beta_1$$

Data: $n = 15$; $m = 3$

$\varepsilon [\text{dm}^3 \text{mol}^{-1} \text{cm}^{-1}]$	3	3.4	4.3	5	6	6.8	8.1	9.2	10.7	11.6
$\lambda [\text{nm}]$	460	470	480	490	500	510	520	530	540	550

$\varepsilon [\text{dm}^3 \text{mol}^{-1} \text{cm}^{-1}]$	12.9	13.6	14.6	15.3	15.5
$\lambda [\text{nm}]$	560	570	580	590	600



Příklad 6.9 Aproximace absorpčního spektra polynomem

Řešení:

- a) Pro $\beta_j = 0$: $F_R = 696 > F_{0.95}(2,12) = 3.885$, je model jako celek statisticky významný.
- b) $t_{0.975}(12)$ je vyšší než T_2 a T_3 , a oba parametry β_1 a β_2 vycházejí statisticky nevýznamné.
- c) $M_T = 0.989$ ukazuje na silnou multikolinearitu.



Příklad 6.9 Aproximace absorpčního spektra polynomem

Tabulka 6.2 Odhady parametrů polynomu

j	Parametr	Odhad b_j	$\sqrt{D(b_j)}$	T_j
1	1	-43.93	19.38	-2.267
2	2	0.1018	0.0735	1.386
3	3	-2.505e-6	6.925e-5	-0.0361

Závěr:

U polynomických regresních modelů nelze z výsledků Studentova t-testu usuzovat významnost jednotlivých členů polynomu. Příčinou je silná multikolinearita.



Příklad 6.46 Multikolinearita u teplotní závislosti aktivního koeficientu

Závislost logaritmu středního aktivního koeficientu $\ln \gamma_{\pm}$ na teplotě T lze vyjádřit polynomem třetího stupně. Posuďte míru multikolinearity a využijte metodu racionálních hodnotí ke snížení stupně multikolinearity.

Data: pro $m_{HCl} = 0.1$.

$T [^{\circ}\text{C}]$	0	10	20	30	40	50	60	70	80	90
$\ln \gamma_{\pm}$	0.8067	0.8038	0.8000	0.7946	0.7927	0.7867	0.7828	0.775	0.769	0.765



Příklad 6.46 Multikolinearita u teplotní závislosti aktivního koeficientu

Řešení: Pro model

$$\ln \gamma_{\neq} = \beta_1 T + \beta_2 T^2 + \beta_3 T^3 + \beta_4$$

1. **Klasická MNČ** (parametr vychýlení P je nastaven na $P = 10^{-35}$): určila $\ln \gamma_{\neq} = 0.807(\pm 1.06 \cdot 10^{-3}) - 2.654 \cdot 10^{-4}(\pm 1.07 \cdot 10^{-4})T - 3.13 \cdot 10^{-6}(\pm 2.87 \cdot 10^{-6})T^2 + 9.44 \cdot 10^{-9}(\pm 2.09 \cdot 10^{-8})T^3$

Koeficient determinace $\hat{R}^2 = 0.9957$,

Kvadratická chyba predice $MEP = 3.507 \cdot 10^{-6}$,

Kritérium $AIC = -132.26$,

$\det(R) = 3.97 \cdot 10^{-4}$ a číslo podmíněnosti $K = 1989.73$.



Příklad 6.46 Multikolinearita u teplotní závislosti aktivního koeficientu

T-testy významnosti parametrů $\beta_1, \beta_2, \beta_3$ (pro $\alpha = 0.05$): β_2 a β_3 jsou statisticky nevýznamné.

Charakteristiky multikolinearity: (platí pro $VIF_j > 10$ jde o silnou multikolinearitu)

P	Charakteristika	j		
		1	2	3
10^{-35}	VIF_j	70.42	439.1	184
	λ_j	$1.46 \cdot 10^{-3}$	$9.35 \cdot 10^{-2}$	2.905
0.05	VIF_j	6.204	0.260	4.373

2. Metoda racionálních hodnotí (parametr vychýlení $P = 0.05$): určil

$$\ln \gamma_{\neq} = 0.807 (\pm 8.72 \cdot 10^{-4}) - 3.22 \cdot 10^{-4} (\pm 3.28 \cdot 10^{-5}) T - 1.476 \cdot 10^{-6} (\pm 7.18 \cdot 10^{-8}) T^2 - 2.837 \cdot 10^{-9} (\pm 3.314 \cdot 10^{-9}) T^3$$

Koeficient determinace $\hat{R}^2 = 0.9955$,

Střední kvadratická chyba predikce MEP = $2.364 \cdot 10^{-6}$

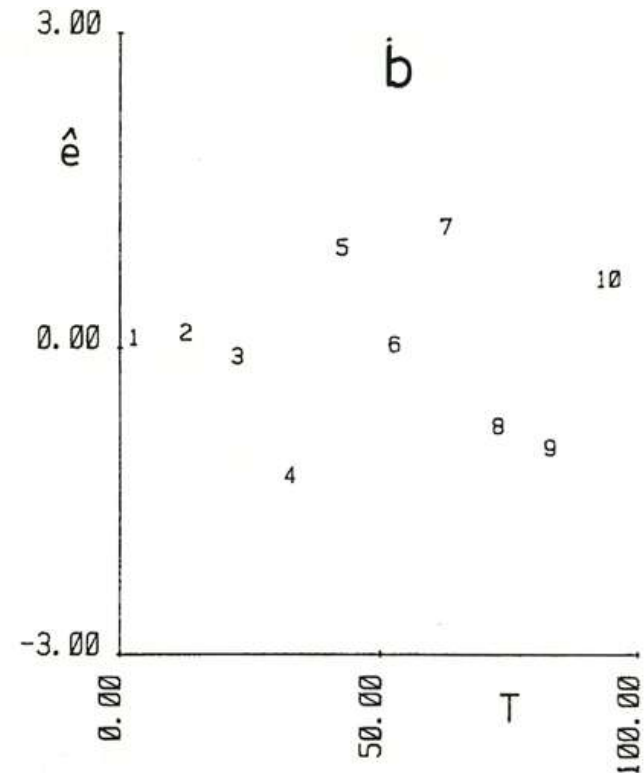
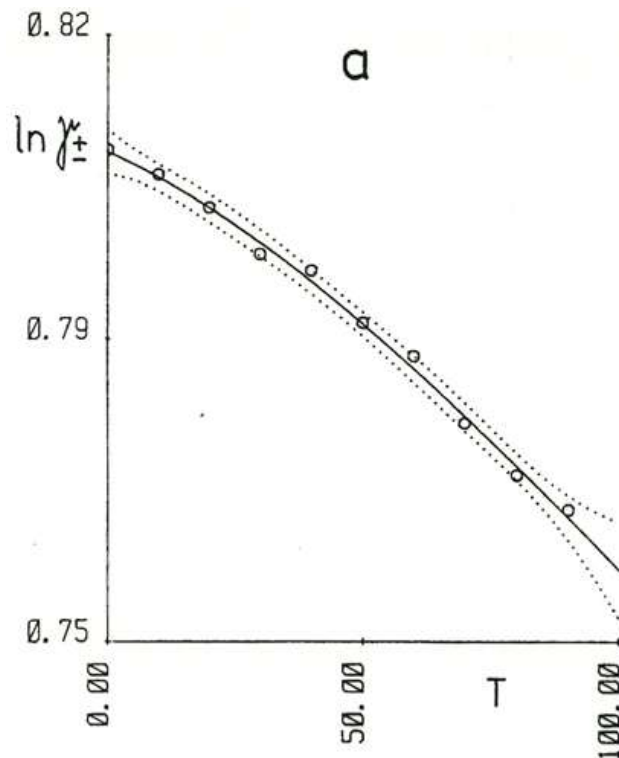
Kritérium AIC = -131.7.

T-testy významnosti parametrů $\beta_1, \beta_2, \beta_3$ (pro $\alpha = 0.05$): pouze β_3 je statisticky nevýznamný.



Příklad 6.46 Multikolinearita u teplotní závislosti aktivního koeficientu

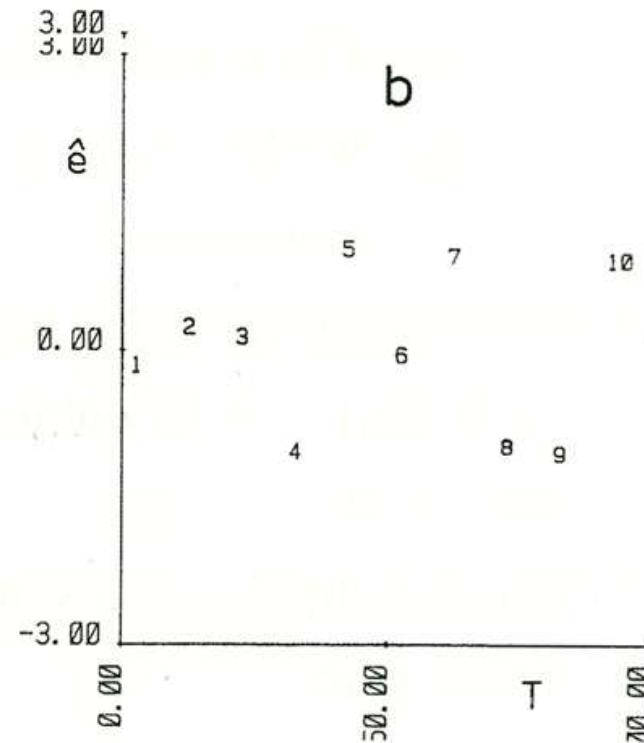
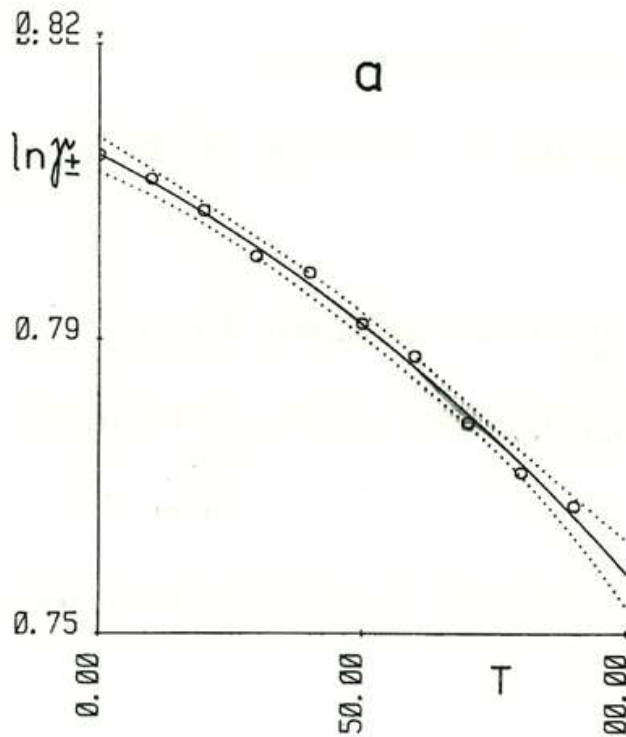
Model určený MNČ





Příklad 6.46 Multikolinearita u teplotní závislosti aktivního koeficientu

Model určený RH





Příklad 6.46 Multikolinearita u teplotní závislosti aktivního koeficientu

Závěr: Eliminace multikolinearity vede:

1. ke snížení přesnosti proložení (poklesu $\hat{R}^2$)
2. ke zlepšení predikční schopnosti modelu (kritérium MEP),
3. k poklesu rozptylů odhadů,
4. k zúžení pásu spolehlivosti.



Příklad 6.23 Aproximace konvexně rostoucí závislosti polynomem

Aproximování experimentálních dat polynomem vhodného stupně tak, aby dokonale splňoval podmínku zachování tvaru křivky dat. Pro aproximaci zvolte polynom šestého stupně.

$$E\left(\frac{y}{x}\right) = \sum_{j=1}^6 b_j x_j + b_7$$

Cílem vyčíslení hodnoty y v počátku čili odhadu parametru β_7 .



Příklad 6.23 Aproximace konvexně rostoucí závislosti polynomem

Data: $n = 10$; $m = 7$

x	25	35	45	55	65	75	85	95	105	115
y	150	160	170	190	210	230	270	310	370	450

Řešení: odhady MNČ při $P = 10^{-29}$ a RH pro $P = 3.5 \cdot 10^{-4}$ (MEP nejmenší).

Odhady parametrů pro různé volby P

Metoda	P	MEP	b_7	b_1	b_2	b_3
MNČ	10^{-29}	380.2	195.7	-5.90	0.257	$-4.9 \cdot 10^{-3}$
RH	$3.5 \cdot 10^{-4}$	8.596	134.7	0.35	$9.2 \cdot 10^{-3}$	$3.2 \cdot 10^{-5}$

Metoda	b_4	b_5	b_6
MNČ	$5.30 \cdot 10^{-5}$	$-2.9 \cdot 10^{-7}$	$6.94 \cdot 10^{-10}$
RH	$-5.3 \cdot 10^{-8}$	$3.9 \cdot 10^{-9}$	$4.5 \cdot 10^{-11}$

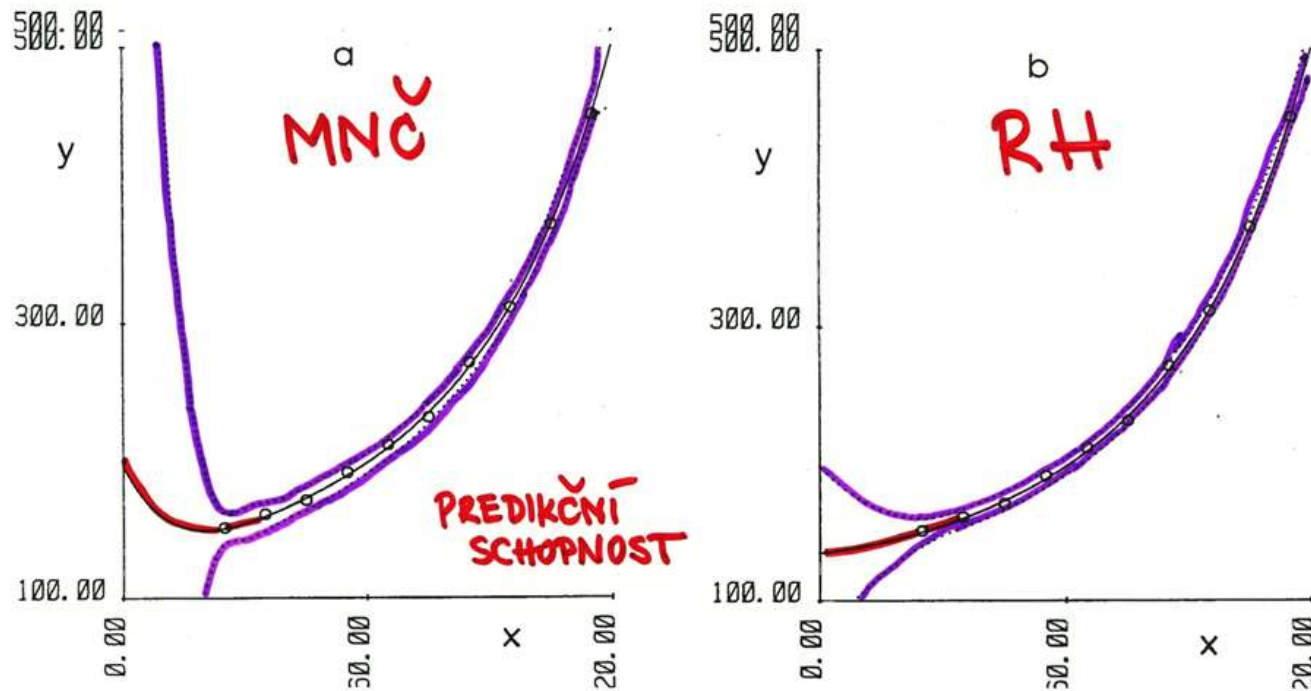


Příklad 6.23 Aproximace konvexně rostoucí závislosti polynomem

- Odhad parametru b_7 je větší než hodnota y_1 , což ukazuje, že model prochází mezi počátkem a bodem (x_1, y_1) minimem.
- Konfidenční intervaly hned za rozsahem experimentálních dat jsou příliš široké, takže neumožňují predikci y mimo rozsah měření.
- Odhady určené metodou racionálních hodnotostí jsou vychýlené.
- Parametr b_7 je však nižší než y_1 a konfidenční intervaly ukazují na možnost predikce i mimo rozsah měření.



Příklad 6.23 Aproximace konvexně rostoucí závislosti polynomem



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Příklad 6.23 Aproximace konvexně rostoucí závislosti polynomem

Závěr:

1. RH poskytne odhady parametrů, které zajišťují průběh modelu odpovídající trendům dat a nemá nadbytečné extrémy či inflexy.
2. Při použití klasické MNČ by úloha byla řešitelná pouze při zavedení omezení na regresní parametry.