Multikolinearita a regresní triplet

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Univerzita Pardubice
Studentská 95, 532 10 Pardubice

Souhrn: **Multikolinearita neznamená v pravém slova smyslu porušení předpokladů klasické metody nejmenších čtverců, souvisí pouze s předpokladem o pozitivní definitnosti matice $X^T X$.**

Chápeme zde sloupce matice X jako sloupcové vektory, které definují nadrovinu L v n-rozměrném eukleidovském prostoru E^n . Podle velikosti úhlu Θ_{jk}, svíraného dvojicí vektorů x_j a x_k, čili sloupci matice X, mohou nastat dva krajní případy:

1. *Ortogonalita*, kdy je kosinus úhlu Θ_{jk} nulový,

$$\cos\Theta_{jk} = \frac{(x_j, x_k)}{\|x_j\| \, \|x_k\|} ,$$

a také skalární součin $(x_j, x_k) = 0$. Zde symbol $\|x_j\|$ = značí délku vektoru x_j. Jsou-li všechny sloupce matice X vzájemně ortogonální, je $X^T X$ diagonální maticí a celá regresní analýza se zjednoduší (viz odst 6.4.1).

2. *Kolinearita*, kdy je kosinus úhlu Θ_{jk} roven jedné ($\cos\Theta_{jk} = 1$), protože úhel mezi vektory x_j a x_k je nulový ($\Theta_{jk} = 0$). Pak jsou oba vektory x_j a x_k rovnoběžné a lineárně závislé

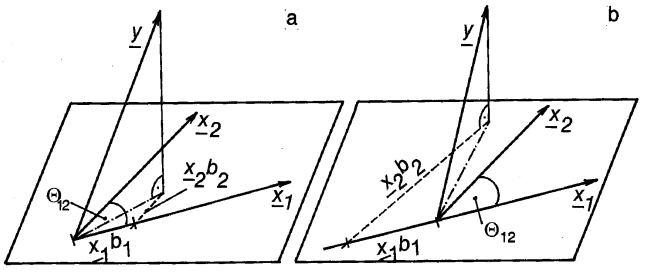
$$c_j x_j + c_k x_k = 0 ,$$

kde c_j, c_k jsou nenulové konstanty. Platí-li tato pro q dvojic sloupců matice X, je její hodnost rovna rozdílu $m - q$ a matice $X^T X$ je singulární.

Vztahy tohoto typu mohou platit i pro více vektorů, kdy je jeden ze sloupců x_j lineární kombinací několika ostatních sloupců. Tato situace se označuje jako *perfektní multikolinearita*. Pojmem multikolinearita se však v oblasti regrese označují i případy, kdy některé sloupce matice X svírají téměř nulový úhel a jsou skoro lineárně závislé

$$\sum_{i=1}^{m} c_j x_j = \delta ,$$

kde δ je vektor, jehož složky jsou blízké nule a vektor c obsahuje nenulové prvky c_j. Platí, že $\|c\| >> \|\delta\|$.



Obr. 1 Geometrické znázornění citlivosti odhadů v případě multikolinearity:
a) odhady b_1 b_2 jsou malé a kladné, b) odhad b_1 je záporný a b_2 velký kladný.

Multikolinearita způsobuje špatnou podmíněnost matice $X^T X$. To má za následek, že:

a) determinant matice $X^T X$ je číslo blízké nule,

b) některá vlastní čísla matice $X^T X$ jsou blízká nule.

Lze tedy očekávat řadu obtíží, když hledáme inverzní matici k matici $X^T X$, a projevů numerických chyb v závislosti na užitém algoritmu a strojové přesnosti počítače.

Kromě numerických problémů způsobuje multikolinearita i řadu statistických obtíží. Z rovnic je patrné, že pro hodnoty λ blízké nule budou odhady i jejich rozptyly neúměrně vysoké. Velký problém představuje citlivost odhadů *b* na malé změny v datech, způsobené například přidáním dalšího bodu.

Na obrázku je znázorněna geometrie metody nejmenších čtverců v případě dvou téměř kolineárních vysvětlujících proměnných. Na obrázku a) je případ, kdy je vektor *y* promítnut do výseče úhlu Θ_{12}. Na obrázku b) je případ, kdy malá změna složek vektoru *y* způsobí, že jeho kolmý průmět padne mimo tuto výseč.

Multikolinearita způsobuje při statistickém zpracování tyto obtíže:

1. N e s t a b i l i t u o d h a d ů , která vzniká citlivostí odhadů na malé změny v datech. Odhady mívají často nesprávné znaménko, což znemožňuje jejich věcnou (fyzikální) interpretaci, a také jsou co do absolutních hodnot příliš veliké.

2. V e l k é r o z p t y l y $D(b_j)$ jednotlivých odhadů, z nichž vyplývá, že výsledky *t*-testů indikují statistickou nevýznamnost parametrů β_j .

3. S i l n o u k o r e l o v a n o s t mezi prvky vektoru odhadů *b*, z čehož vyplývá, že je nelze interpretovat odděleně.

Na druhé straně však v případě multikolinearity vychází koeficient determinace vysoký a regresní model může dobře popisovat experimentální data. Pro účely aproximace dat a konstrukce modelů, které mají „vyhladit" experimentální závislosti, není multikolinearita na obtíž, zůstává pouze numerické hledisko, spočívající ve špatné podmíněnosti matice $X^T X$.

Pokud jsou data získána na základě plánovaných experimentů, je problém multikolinearity odstraněn. Schémata plánovaných experimentů vedou k ortogonalitě sloupců matice *X*.

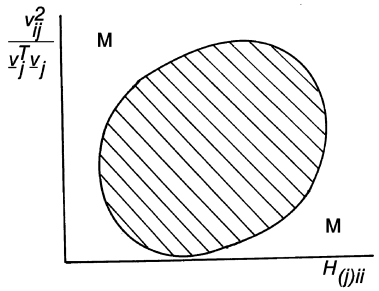
Pokud jsou data získána na základě neplánovaných experimentů, nebo nepocházejí z experimentálních měření, vyskytuje se multikolinearita jistého stupně téměř vždy. Problémem je však jen *silná multikolinearita*, kdy jsou odhady parametrů a testy hypotéz ovlivněny více l i n e á r n í m i v a z b a m i mezi sloupci matice *X* než vlastním regresním modelem.

U technických modelů lze obyčejně nastavovat hodnoty vysvětlujících proměnných v experimentu libovolně a problémy multikolinearity lze odstranit již ve fázi získávání experimentálních dat.

Mezi základní příčiny vzniku multikolinearity patří:

1. Přeurčený regresní model obsahující nadměrný počet vysvětlujících proměnných, které vyjadřují stejné základní faktory. Příkladem jsou modely typu „struktura−vlastnosti", ve kterých se vlastnosti látek popisují různými měřitelnými proměnnými struktury. Ty různou měrou souvisejí s menším počtem strukturních faktorů.

2. Nevhodné rozmístění experimentálních bodů, kdy multikolinearita vzniká „uměle" výběrem poloh bodů tak, aby byla dodržena rovnice. Často hodnoty významné vysvětlující proměnné kolísají jen v malém rozmezí, že se jeví jako konstantní a jsou tedy kolineární s vektorem odpovídajícím absolutnímu členu.



3. Fyzikální omezení modelu nebo v datech, kdy vznikají vazby mezi vysvětlujícími proměnnými přímo ve sledovaném systému. Příkladem může být sledování vícekompo-nentních směsí, u kterých vysvětlující proměnné představují obsahy jednotlivých složek. Jelikož součet všech relativních zastoupení složek musí být 100%, bude ve *q*-složkové směsi pouze $q - 1$ nezávislých složek. Při jejich nastavení bude *q*-tá složka doplňkem do 100 % a při sestavování regresního modelu bude třeba uvažovat jenom $q - 1$ proměnných: u dvousložkové směsi jde o jednu proměnnou, u trojsložkové o dvě atd. Obdobná omezení mohou vnášet i stechiometrické poměry atd.

Na základě předběžných znalostí o vysvětlujících proměnných, jejich významu a vzájemných vazbách lze často multikolinearitu omezit nebo zcela vyloučit. V některých případech, jako např. u polynomických modelů, je multikolinearita dána přímo strukturou modelu. V těchto situacích, nebo tam, kde nelze ovlivnit přípravu experimentu, je vhodné užít postupy, při nichž se omezuje působení multikolinearity i za cenu vychýlení odhadů, jako je například metoda racionálních hodností. Přednosti této metody proti klasické metodě nejmenších čtverců v případě silné multikolinearity v datech jsou demonstrovány v příkladu.

Multikolinearitu lze často odhalit na základě rozptylových grafů pro x_j a x_k, kdy zhruba lineární závislost ukazuje na silnou multikolinearitu.

Multikolinearita může být vyvolána nebo naopak maskována přítomností vlivných bodů, a to zejména extrémů. K diagnostickým účelům je vhodné užít reziduí v_j regrese proměnné x_j na ostatních vysvětlujících proměnných v matici $X_{(j)}$, která neobsahuje sloupec s x_j. Označme $H_{(j)}$ projekční matici, která odpovídá projekci do podprostoru sloupců matice $X_{(j)}$. Pro diagnostiku vlivných bodů z hlediska multikolinearity se užívá závislosti $v_{ij}^2/(\boldsymbol{v}_j^{\mathrm{T}}\,\boldsymbol{v}_i)$ na $H_{(j)ii}$. Zde v_{ji} je *i*-tá složka vektoru v_j a $H_{(j)ii}$ je *i*-tý diagonální prvek matice $H_{(j)}$.

V tomto grafu se multikolinearitou ovlivňované body vyskytují jako body izolované v pravém dolním nebo levém horním rohu. Body ležící v levém horním rohu ovlivňují multikolinearitu, jen pokud je x_j zahrnuta v modelu. Na druhé straně body v pravém dolním rohu jsou silně vlivné, jen pokud není x_j zahrnuta v modelu.

Přítomnost multikolinearity lze posuzovat na základě numerických a statistických kritérií. Místo matice $X^T X$ se používá její normovaná verze R, která je formálně shodná s korelační maticí vysvětlujících proměnných. Z numerických kritérií se užívají:

a) *Determinant matice R*

$$\det(\boldsymbol{R}) \quad = \quad \prod_{j=1}^{m} \lambda_j \ ,$$

kde λ_j jsou vlastní čísla matice R. Je-li determinant $\det R$ příliš malý, menší než 10^{-3}, jde o silnou multikolinearitu;

b) *Číslo podmíněnosti*

$$K \quad = \quad \frac{\lambda_{\max}}{\lambda_{\min}} \ ,$$

kde λ_{max}, λ_{min} jsou maximální a minimální vlastní číslo matice R. Hodnota $K > 10^3$ ukazuje na silnou multikolinearitu.

c) *VIF-faktor* (Variance Inflation Factor), jenž je definován jako poměr rozptylu *j*-tého regresního koeficientu a rozptylu tohoto koeficientu pro případ ortogonálních proměnných, kdy je R jednotkovou maticí. Platí, že

$$VIF_j \quad = \quad \tilde{R}_{jj} \ ,$$

kde $\tilde{R}_{jj}$ je j-tý diagonální prvek matice R^{-1}. Parametry *VIF* souvisejí s koeficientem determinace regrese x_j na $X_{(j)}$, kdy se vlastně vyjadřuje x_j jako lineární kombinace ostatních vysvětlujících proměnných. Platí, že

$$VIF_j \quad = \quad \frac{1}{1 \ - \ \hat{R}_{x.}^2} \ .$$

Je-li $VIF_j > 10$, jde o silnou multikolinearitu.

V technických modelech se často provádí hlubší analýza multikolinearity a hledají se vazby mezi vysvětlujícími proměnnými.

PŘÍKLAD 1. ***Multikolinearita u teplotní závislosti aktivitního koeficientu***

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

○ ***Data:*** **pro $m_{HCl} = 0.1$**

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

○ ***Řešení:*** **Pro model**

$$\ln \gamma_\pm = \beta_1 T + \beta_2 T^2 + \beta_3 T^3 + \beta_4$$

byla klasickou MNČ (tzn. parametr vychýlení P je nastaven na hodnotu $P = 10^{-35}$) určena regresní rovnice

$$\ln \gamma_{\pm} = 0.807\ (\pm 1.06 \cdot 10^{-3}) - 2.654 \cdot 10^{-4}\ (\pm 1.07 \cdot 10^{-4})\ T - \\ - 3.13 \cdot 10^{-6}\ (\pm 2.87 \cdot 10^{-6})\ T^2 + 9.44 \cdot 10^{-9}\ (\pm 2.09 \cdot 10^{-8})\ T^3.$$

V závorce jsou u každého členu polynomu uvedeny směrodatné odchylky odhadů. Koeficient determinace $\hat{R}^2$ = 0.9957, kvadratická chyba predikce $MEP = 3.507 \cdot 10^{-6}$ a kritérium $AIC = -132.26$. V tabulce 1 jsou uvedena vlastní čísla a *VIF*-faktory. Z těchto údajů se vyčíslí det $R = 3.97 \cdot 10^{-4}$ a číslo podmíněnosti $K = 1989.73$. Na základě *t*-testů vycházejí na hladině významnosti $\alpha = 0.05$ parametry β_2 a β_3 statisticky nevýznamné.

Tabulka 1. Charakteristiky multikolinearity

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

Užitím metody racionálních hodností s parametrem vychýlení $P = 0.05$ byl určen regresní model ve tvaru

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

Koeficient determinace $\hat{R}^2 = 0.9955$, střední kvadratická chyba predikce $MEP = 2.364 \cdot 10^{-6}$ a kritérium $AIC = -131.7$. V tabulce 1. jsou uvedeny odpovídající parametry *VIF*. Je patrné, že došlo k odstranění výrazné multikolinearity. Ze závěrů *t*-testů na hladině významnosti $\alpha = 0.05$ se pouze koeficiet β_3 jeví statisticky nevýznamný.

Na obrázku 2 v části a) je znázorněn regresní model určený metodou nejmenších čtverců spolu s 95% pásy spolehlivosti a v části b) je závislost reziduí na proměnné *T*.

Na obrázku 3 v části a) je znázorněn regresní model určený metodou racionálních hodností ($P = 0.05$) spolu s 95% pásy spolehlivosti a v části b) pak závislost reziduí na proměnné *T*. Porováním obou obrázků je vidět, že eliminace multikolinearity vede ke zúžení pásů spolehlivosti.

Obr. 2 a) Regresní model určený metodou nejmenších čtverců spolu s 95% pásy spolehlivosti a b) závislost reziduí na proměnné *T*.

Obr. 3 a) Regresní model určený metodou racionálních hodností spolu s 95% pásy spolehlivosti a b) závislost reziduí na proměnné *T*.

○ ***Závěr:*** **Výrazná multikolinearita, indikovaná hodnotou kritéria $VIF > 10$, způsobuje růst rozptylů odhadů, což se projeví i rozšířením pásů spolehlivosti a dále i ve výsledcích testů významnosti parametrů. Eliminace multikolinearity vede ke snížení přesnosti proložení (poklesu $\hat{R}^2$), ale ke zlepšení predikční schopnosti modelu (kritérium *MEP*), dále k poklesu rozptylů odhadů a zúžení pásu spolehlivosti.** ❑

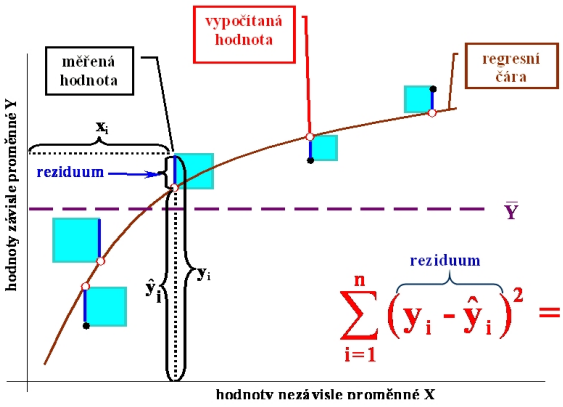
Formulace lineárního regresního modelu

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

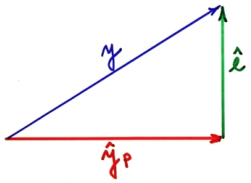
$\mathbf{y}$ — závisle proměnná; $\mathbf{X}$ — nezávisle proměnná; $\boldsymbol{\beta}$ — regresní parametry; $\boldsymbol{\varepsilon}$ — náhodná chyba

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

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



Metoda projekčních matic



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

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

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

Odhady modelových parametrů **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 $\hat{e}$ je kolmý na všechny sloupce matice **X**, a proto odpovídající skalární součiny jsou nulové

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

a maticově $\mathbf{X}^T \hat{e} = 0$

a dosazením $\hat{e} = y - X b$ 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

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

a pomocí vektoru **b**

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

Vlastnosti projekční matice H:

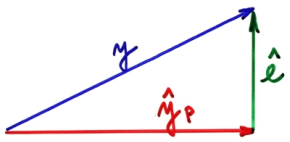
1. Projekční matice $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T$ promítne libovolný vektor **V** do roviny L.
2. Pokud vektor **V** již leží v rovině L, platí $\mathbf{H}\,\mathbf{V} = \mathbf{V}$.
3. Pokud je však vektor **V** kolmý na rovinu L, platí $\mathbf{H}\,\mathbf{V}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

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

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

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 **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 **X 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, $\mathrm{cov}(\epsilon_i \,.\, \epsilon_j) = E(\epsilon_i \,.\, \epsilon_j) = 0$.
7. Náhodné chyby mají normální rozdělení $\epsilon \approx N(0, \sigma^2)$.

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)

Kritika dat

Vyšetření vlivných bodů:

(1) Odlehlé body

(2) Extrémy

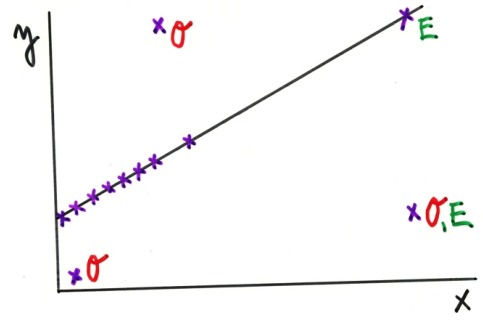
Metoda:

Grafy vlivných bodů

Vybrané diagnostiky vlivných bodů

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,
2. Čím je reziduum $\hat{e}_i$ větší, tím je bod vlivnější, a tím spíše by se měl z dat vyloučit.

$$\hat{\mathbf{e}} = \mathbf{P}\,\mathbf{y} = \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{P}\,\boldsymbol{\varepsilon} = (\mathbf{E} - \mathbf{H})\,\boldsymbol{\varepsilon}$$

Rozepsáním

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

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.

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.

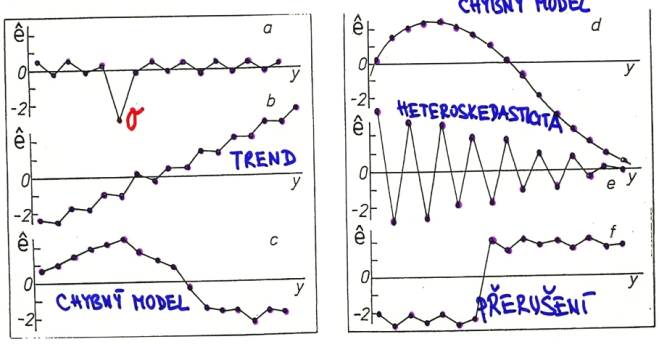
3. Rezidua *neindikují* silně extrémní hodnoty.

4. Rezidua jsou normálnější než chyby:
(effekt supernormality)

5. U malých výběrů nemusí správně indikovat model.

Statistická analýza klasických reziduí

1. Grafická analýza



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^2_{exp} vs. $\chi^2_{crit.}$,

(g) **Hamiltonův R-faktor relativní těsnosti:** $R \approx 0.5\ \%.$

Indikace vlivných dat Software QC-EXPERT 3.1 (TriloByte)

A. Analýza rozličných druhů reziduí

Index	Standardní	Jackknife	Predikované	Diag(Hii)	Diag(H*ii)	Cookova vzdál.
1.0	0.358396	0.349548	12.039487	0.225546	0.231072	0.052188
2.0	-1.065714	-1.069995	-32.439711	0.056782	0.116296	-0.032078
3.0	0.508826	0.498085	15.511562	0.059601	0.073127	0.016124
4.0	-0.18279	-0.177805	-5.926527	0.168641	0.170184	-0.018539
5.0	1.159883	1.171846	38.041795	0.187561	0.248283	0.133886
6.0	-3.174503	-4.650163	-96.457219	0.0534	0.583363	-0.089541
7.0	1.283018	1.308118	39.164901	0.062101	0.147874	0.042476
8.0	0.046714	0.045401	1.552602	0.208837	0.208933	0.006165
9.0	0.614627	0.603678	19.505109	0.132217	0.150429	0.046823
10.0	-0.506712	-0.495985	-15.381984	0.051622	0.06515	-0.013791
11.0	-0.349541	-0.340852	-10.703492	0.067969	0.074296	-0.012745
12.0	-1.03349	-1.035566	-32.279234	0.104119	0.157279	-0.060056
13.0	0.470912	0.460489	14.284453	0.050191	0.061893	0.012442
14.0	0.95392	0.951404	29.417222	0.081021	0.127479	0.042051
15.0	0.581705	0.570705	18.061898	0.093509	0.110551	0.030003
16.0	-0.575594	-0.564597	-17.475813	0.051926	0.069376	-0.015763
17.0	0.92739	0.923596	28.860813	0.097614	0.140731	0.05016
18.0	0.385478	0.376173	11.853575	0.07576	0.08339	0.015799
19.0	0.144322	0.140337	4.386234	0.053834	0.054928	0.004106
20.0	-0.394223	-0.384781	-12.407616	0.117749	0.125366	-0.026307

2. Normovaná rezidua $\hat{\sigma}_N$

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

normálně rozdělené veličiny $\hat{e}_{Ni} \sim N(0, 1)$.
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 $\hat{\sigma}_{(-i)}$,

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

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. Rekurzívní rezidua $\hat{\sigma}_R$

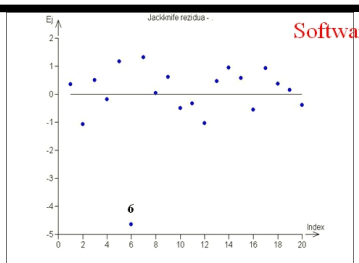
Dopředná rekurzívní rezidua jsou definována vztahy

$$\hat{e}_{Ri} = 0, \qquad i = 1, \ldots, m$$

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

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

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

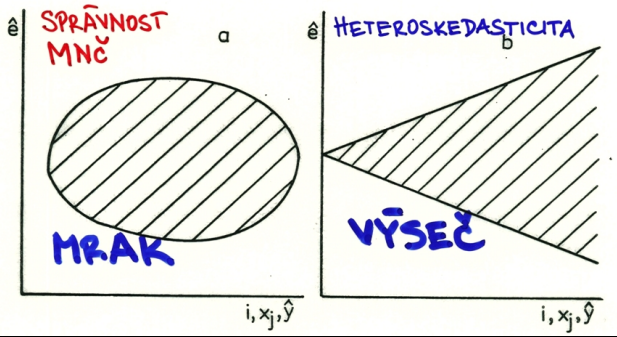


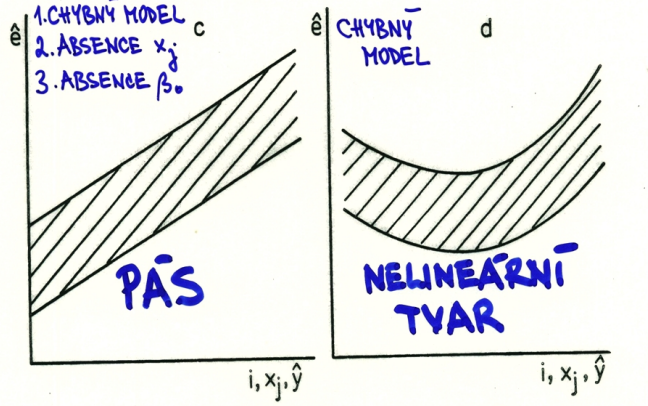
Software QC-EXPERT 3.1 (TriloByte)

Jackknife rezid.	Predikované rezid.
0.349548	12.039487
-1.069995	-32.439711
0.498085	15.511562
-0.177805	-5.926527
1.171846	38.041795
-4.650163*outlier	-96.457219*outlier
1.308118	39.164901
0.045401	1.552602
0.603678	19.505109
-0.495985	-15.381984
-0.340852	-10.703492
-1.035566	-32.279234
0.460489	14.284453
0.951404	29.417222
0.570705	18.061898
-0.564597	-17.475813
0.923596	28.860813
0.376173	11.853575
0.140337	4.386234
-0.384781	-12.407616

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 $\hat{y}_i$.



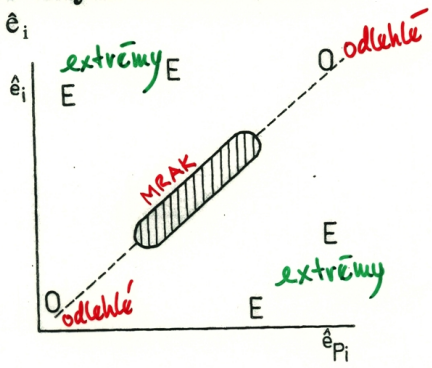


(a) tvar mraku, (b) tvar výseče, (c) tvar pásu a (d) nelineární tvar

Grafy identifikace vlivných bodů

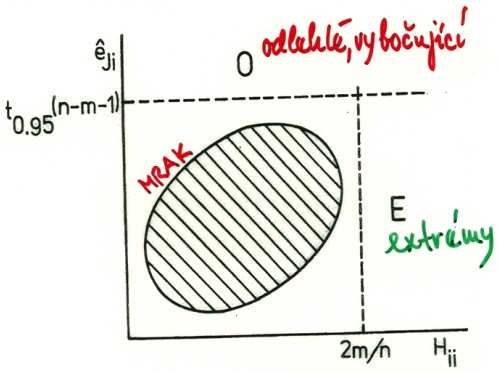
1. Graf predikovaných reziduí (GPR),

osa x: $\hat{e}_{Pi}$, osa y: $\hat{e}_i$



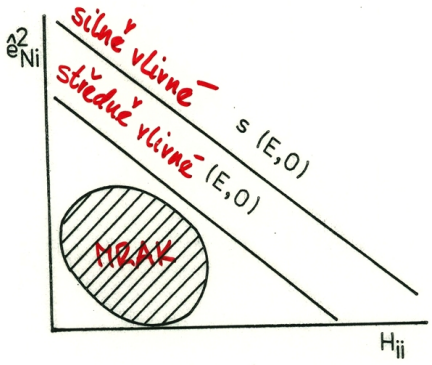
2. Williamsův graf (WG),

osa x: prvky H_{ii}, osa y: $\hat{e}_{Ji}$



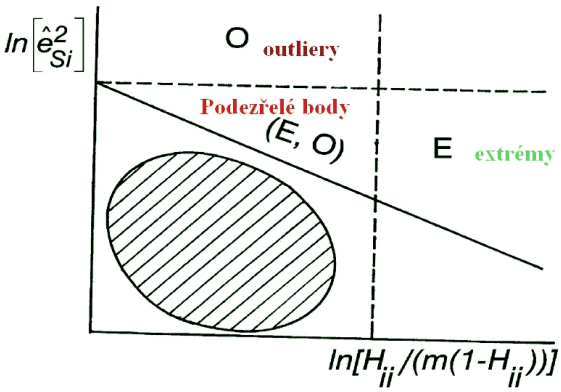
3. Pregibonův graf (PG),

osa x: prvky H_{ii}, osa y: $\hat{e}^2_{Ni}$

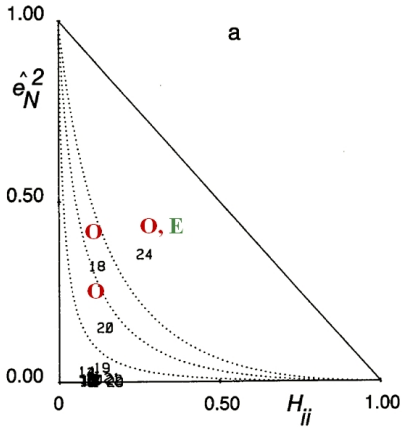


4. McCullohův-Meeterův graf (MMG)

osa *x*: $\ln [H_{ii} /(m(1 - H_{ii}))]$, osa *y*: $\ln e^2_{S_i}$



5. L-R graf (LR)



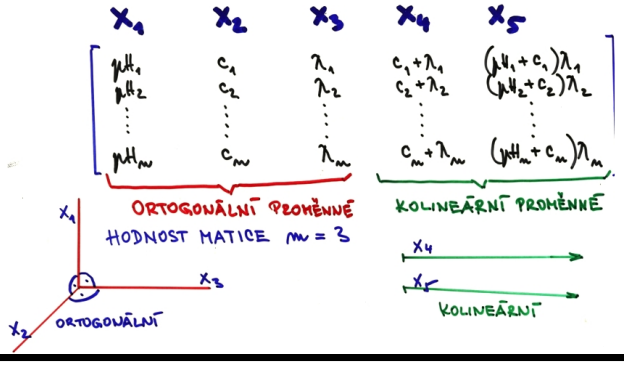
MULTIKOLINEARITA

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

Vektory matice 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.

Test multikolinearity

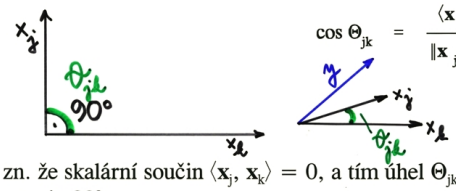
Paradoxní situace: F-test je významný a všechny t-testy jsou nevýznamné, protože je silná multikolinearita mezi sloupci matice **X**, čili existuje rovnoběžnost vektorů $\mathbf{x}_j$ a $\mathbf{x}_k$, $j \neq k$, sloupců matice **X**.



Multikolinearita

M. neznamená porušení předpokladů MNČ,
M. souvisí pouze s pozitivní definitností matice $\mathbf{X}^T\mathbf{X}$,

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 $\langle \mathbf{x}_j, \mathbf{x}_k \rangle = 0$, a tím úhel Θ_{jk} mezi vektory $\mathbf{x}_j$ a $\mathbf{x}_k$ je 90°.

(Symbol $\|\mathbf{x}_j\| = \sqrt{\langle \mathbf{x}_j, \mathbf{x}_j \rangle}$ značí délku vektoru $\mathbf{x}_j$).

Jsou-li všechny sloupce matice **X** vzájemně ortogonální, je $\mathbf{X}^T\mathbf{X}$ diagonální matice.

2. **Kolinearita**, kdy je kosinus Θ_{jk} roven 1, (cos Θ_{jk} = 1), protože úhel Θ_{jk} mezi vektory $\mathbf{x}_j$ a $\mathbf{x}_k$ je nulový (Θ_{jk} = 0).

Pak jsou oba vektory $\mathbf{x}_j$ a $\mathbf{x}_k$ rovnoběžné čili lineárně závislé

$c_j \mathbf{x}_j + c_k \mathbf{x}_k = \mathbf{0}$ a c_j, c_k jsou nenulové konstanty.

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

- Vazebné vztahy mohou platit i pro více vektorů, kdy je jeden ze sloupců $\mathbf{x}_j$ lineární kombinací několika ostatních sloupců (***perfektní multikolinearita***).

- Pojmem multikolinearita se však označují i případy, kdy některé sloupce matice **X** svírají téměř nulový úhel a jsou prakticky lineárně závislé $\sum_{j=1}^{m} c_j \mathbf{x}_j = \delta$, kde δ je vektor, jehož složky jsou blízké nule a vektor **c** s prvky c_j je nenulový. Platí, že $\|\mathbf{c}\| \gg \|\delta\|$.

- M. způsobuje špatnou podmíněnost matice $\mathbf{X}^T\mathbf{X}$ čili
  a) determinant matice $\mathbf{X}^T\mathbf{X}$ je číslo blízké nule,
  b) některá vlastní čísla matice $\mathbf{X}^T\mathbf{X}$ jsou blízká nule.

Multikolinearitu lze ***odstranit:***

a) Vhodnou volbou poloh experimentálních bodů, kdy sloupce matice **X** budou vzájemně ortogonální, tj. jejich skalární součin bude nulový

$$\langle \mathbf{x}_j, \mathbf{x}_k \rangle = \sum_{i=1}^{n} x_{ij} x_{ik} = 0 \qquad \text{pro} \quad j \neq k$$

b) Použití funkcí ortogonálních vzhledem k daným polohám experimentálních bodů.

Statistické obtíže:

1. ***Nestabilita odhadů*** je způsobená citlivostí odhadů na malé změny v datech. Odhady mívají často nesprávné znaménko, což znemožňuje jejich věcnou (fyzikální) interpretaci a jsou co do absolutních hodnot příliš veliké.

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.

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).

Metoda racionálních hodností

Matici **R** (symetrická) vyjádříme:

- pomocí vlastních čísel $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$
- a odpovídajících vlastních vektorů $\mathbf{P}_j$, j = 1, ..., m, ve tvaru

$$\mathbf{R} = \sum_{j=1}^{m} \lambda_j \mathbf{P}_j \mathbf{P}_j^T$$

a inverzní matici $\mathbf{R}^{-1}$ vztahem

$$\mathbf{R}^{-1} = \sum_{j=1}^{m} \lambda_j^{-1} \mathbf{P}_j \mathbf{P}_j^T$$

a ***vektor normovaných odhadů parametrů*** bude

$$\mathbf{b}_N = \sum_{j=\omega}^{m} \left(\lambda_j^{-1} \mathbf{P}_j \mathbf{P}_j^T\right) \mathbf{r}$$

a ***kovarianční matice normovaných odhadů*** bude

$$D(\mathbf{b}_N) = \hat{\sigma}_N^2 \sum_{j=\omega}^{m} \lambda_j^{-1} \mathbf{P}_j \mathbf{P}_j^T$$

V případě MNČ bude $\omega = 1$.

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

Jsou-li všechny sloupce matice X vzájemně ortogonální, matice $\mathbf{X}^T\mathbf{X}$ je diagonální a odhad b lze vyjádřit

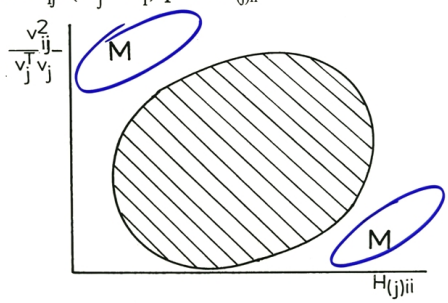
$$b_j = \frac{\sum_{i=1}^{n} x_{ij} y_i}{\sum_{i=1}^{n} x_{ij}^2} \qquad j = 1, \ldots, m$$

a definuje se $F_R = \dfrac{\sum_{j=1}^{m-1} T_j^2}{m-1} = T_S$, **kde**

T_S je průměrná hodnota čtverců testačních statistik T_j^2, definovaných pro případ $\beta_{0j} = 0$ a β_m je absolutní člen.

Diagnostika:

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



Diagnostika multikolinearity: M značí multikolinearitu

2. Numerická kritéria:

a) **Determinant matice R,** $\det(\mathbf{R}) = \prod_{j=1}^{m} \lambda_j$, 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 $\mathbf{R}^{-1}$. Platí vztah

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

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

	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

Řešení problémů II. a III.: užitím racionálních hodností

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

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

kde

- P je zvolená přesnost (obyčejně 10^{-5}),
- číslo ω určuje spodní mez, od které se norm. odhadů provádí sčítání.

Označme $W = \sum_{j=1}^{\omega}\lambda_j$

a $E = \sum_{j=1}^{m}\lambda_j$

$W/E > P$

Řešení:

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

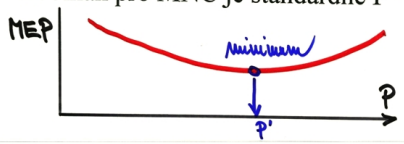
$$u = \frac{W - EP}{\lambda_\omega}$$

- tím je zajištěno, že lze spojitě v závislosti na růstu přesnosti P snižovat délku odhadů $\|\mathbf{b}_N\|$ a jejich rozptyly.
- 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 ω > 1.
- pro čtverec vychýlení $h_v^2(\mathbf{b}_N) = [\beta - E(\mathbf{b})]^2$, získaný metodou racionálních hodností, platí

$$h_V^2(\mathbf{b}_N) = \beta_N^T \left(\sum_{j=1}^{\omega}\mathbf{P}_j\mathbf{P}_j^T\right)\beta_N$$

- 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, ε = f(λ) polynomem druhého stupně

$$E(\varepsilon/\lambda) = \beta_2\lambda + \beta_3\lambda^2 + \beta_1$$

Data: n = 15; m = 3

ε [dm³ mol⁻¹ cm⁻¹]	3	3.4	4.3	5	6	6.8	8.1	9.2	10.7	11.6
λ [nm]	460	470	480	490	500	510	520	530	540	550

ε [dm³ mol⁻¹ cm⁻¹]	12.9	13.6	14.6	15.3	15.5
λ [nm]	560	570	580	590	600

Řešení:

a) Pro $\beta_j = 0$: $F_R = 696 > F_{0.95}(2, 12) = 3.885$, je model jako celek statisticky významný.

b) $t_{0.975}(12) = 2.179$ je vyšší než T_2 a T_3, a oba parametry β_2 a β_3 vycházejí statisticky nevýznamné.

c) $M_T = 0.989$ ukazuje na silnou multikolinearitu.

Řešení:

a) Pro $\beta_j = 0$: $F_R = 696 > F_{0.95}(2, 12) = 3.885$, je model jako celek statisticky významný.

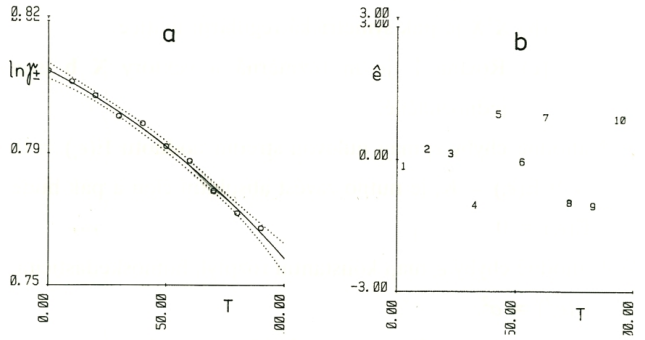
b) $t_{0.975}(12) = 2.179$ je vyšší než T_2 a T_3, a oba parametry β_2 a β_3 vycházejí statisticky nevýznamné.

c) $M_T = 0.989$ ukazuje na silnou multikolinearitu.

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	$-2.505 \cdot 10^{-6}$	$6.925 \cdot 10^{-5}$	-0.0361

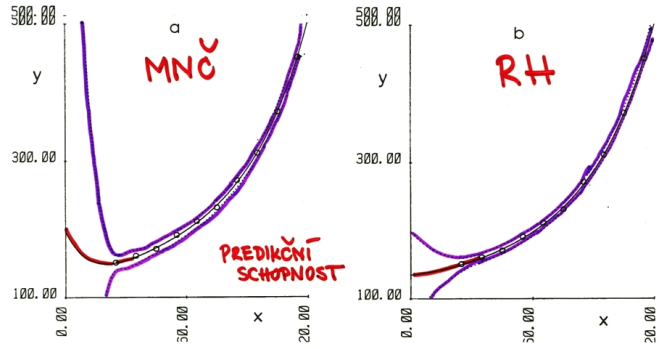
Závěr: U polynomických regresních modelů ***nelze*** z výsledků Studentova t-testu usuzovat na významnost jednotlivých členů polynomu. Příčinou je silná multikolinearita.



Model určený RH

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.



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.

Kritika modelu

Vyšetření významnosti parametrů

Metoda:

Parciální regresní grafy

Parciální reziduální grafy

Studentův *t*-test významnosti parametru β

KRITÉRIA HLEDÁNÍ A ROZLIŠENÍ NEJLEPŠÍHO REGRESNÍHO MODELU

Střední kvadratická chyba predikce (MEP)

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

e_i^2 čtverec reziduí modelu
H_{ii} *i*-tý diagonální prvek projekční matice **H**

Akaikovo informační kritérium (AIC)

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

RSC reziduální součet čtverců
m počet parametrů

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

OBECNÝ POSTUP VÝSTAVBY REGRESNÍHO MODELU

1. Kvalita nalezených odhadů parametrů

a) Podle **intervalů spolehlivosti** (čím menší interval spolehlivosti, tím lépe)

$$\beta_j = b_j \pm \sqrt{C_{mm}\cdot m\cdot s^2\cdot F_{1-\alpha;m;n-m}}$$

b) Podle **rozptylů parametrů**, kde pro kvalitní odhad musí platit

$$2\cdot\sqrt{D(b_j)} < |b_j|$$

2. Kvalita dosažené těsnosti proložení

a) Podle **reziduálního rozptylu** $s(y)$.

b) Podle **regresního rabatu** D (= koeficient determinace v %: čím více se blíží 100 %, tím lepší je proložení).

3. Vhodnost navrženého modelu

Akaikovo informační kritérium *AIC* (čím je menší nebo zápornější, tím vhodnější je navržený model).

(3) ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

Parametr	Odhad	Směrodatná odchylka	Test H0: B[j] = 0 vs. HA: B[j] <> 0 t-kriterium	hypoteza H0 je	Hlad. výz.
B[0]	-7.2666E-02	1.3791E-01	-5.2692E-01	Akceptována	0.608
B[1]	-6.8505E-01	1.9165E-01	-3.5746E+00	Zamítnuta	0.004
B[2]	8.9619E-01	1.6072E-01	5.5761E+00	Zamítnuta	0.000
B[3]	8.3769E-01	1.3322E-01	6.2879E+00	Zamítnuta	0.000

(4) STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	: 9.9858E-01
Koeficient determinace, R^2	: 9.9716E-01
Predikovaný korelační koeficient, Rp^2	: 9.9527E-01
Střední kvadratická chyba predikce, MEP	: 2.1014E-01
Akaikeho informační kritérium, AIC	:-3.6180E+01

4. Predikční schopnost modelu

Střední kvadratická chyba predikce *MEP* (čím je MEP menší, tím je predikční schopnost navrženého modelu lepší).

5. Kvalita experimentálních dat

a) Na základě analýzy rozličných druhů **reziduí**.

b) Na základě analýzy **vlivných bodů** (Jackknife rezidua, standardizovaná r., normovaná r., predikovaná r., rekurzivní r., Cookova vzdálenost, diagonální prvky projekční matice a věrohodnostní vzdálenosti).

(5) ANALÝZA KLASICKÝCH REZIDUÍ:

Bod i	Měřená hodnota yexp[i]	Predikovaná hodnota yvyp[i]	Směrodatná odchylka s(yvyp[i])	Klasické reziduum e[i]	Relativní reziduum er[i]
1	1.6000E+00	1.5006E+00	1.0203E-01	9.9415E-02	6.2135E+00
2	1.6000E+00	1.4227E+00	1.0587E-01	1.7733E-01	1.1083E+01
3	2.1000E+00	2.1029E+00	1.1045E-01	-2.9205E-03	-1.3907E-01
4	2.1000E+00	2.1925E+00	1.0412E-01	-9.2540E-02	-4.4067E+00
5	8.1000E+00	8.1354E+00	2.0871E-01	-3.5421E-02	-4.3729E-01
6	7.9000E+00	7.9410E+00	1.3352E-01	-4.1016E-02	-5.1919E-01
7	8.4000E+00	8.6869E+00	1.6249E-01	-2.8690E-01	-3.4155E+00
8	1.0300E+01	9.9377E+00	1.7913E-01	3.6232E-01	3.5176E+00
9	9.6000E+00	9.9805E+00	1.3555E-01	-3.8050E-01	-3.9636E+00
10	1.0800E+01	1.0621E+01	1.3061E-01	1.7918E-01	1.6591E+00
11	1.3100E+01	1.3581E+01	2.2832E-01	-4.8080E-01	-3.6703E+00
12	1.5100E+01	1.4565E+01	1.8609E-01	5.3530E-01	3.5450E+00
13	1.3000E+00	1.2908E+00	1.0514E-01	9.1821E-03	7.0632E-01
14	1.2000E+00	1.3441E+00	1.1635E-01	-1.4406E-01	-1.2005E+01
15	1.5000E+00	1.5597E+00	1.1723E-01	-5.9675E-02	-3.9784E+00
16	1.5000E+00	1.3389E+00	1.0987E-01	1.6110E-01	1.0740E+01

Rezidualní součet čtverců, RSC:	1.0114E+00
Průměr absolutních hodnot reziduí, Me:	1.9048E-01
Průměr relativních reziduí, Mer:	4.3750E+00
Odhad reziduálního rozptylu, s^2(e):	8.4282E-02
Odhad směrodatné odchylky reziduí, s(e):	2.9031E-01
Odhad šikmosti reziduí, g1(e):	9.2356E-02
Odhad špičatosti reziduí, g2(e):	2.8755E+00

Kritika metody

Ověření předpokladů metody nejmenších čtverců

Metoda:

- Vyšetření heteroskedasticity
- Vyšetření autokorelace
- Vyšetření multikolinearity
- Vyšetření normality náhodných chyb
- Vyšetření omezení parametrů
- Vyšetření trendů reziduí

REGRESNÍ DIAGNOSTIKA
testy vybraných předpokladů klasické MNČ

Multikolinearita: VIF diagnostika indikuje

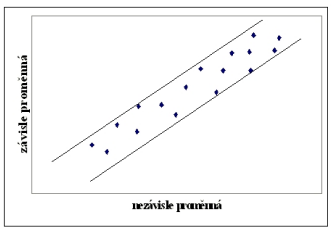
Heteroskedasticita: testy heteroskedasticity (např. Cook-Weisberg)

Autokorelace reziduí: test významnosti autokorelačního koeficientu

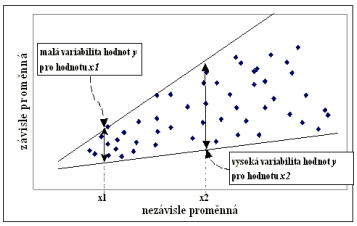
Normalita reziduí: testy normality

HOMOSKEDASTICITA *vs.* ETEROSKEDASTICITA
Grafické vysvětlení principu

Homoskedasticita znamená, že hodnoty závisle proměnné y mají pro všechny hodnoty nezávisle proměnné x **konstantní rozptyl** (variabilitu, proměnlivost).



Homoskedasticita



Heteroskedasticita

Cookův - Weisbergův test homoskedasticity

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

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

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

2. Autokorelace

Data časových řad mají chyby ε_i vzájemně korelované.

Nejčastější je případ autokorelace prvního řádu

$$\varepsilon_i = \rho_1\,\varepsilon_{i-1} + u_i$$

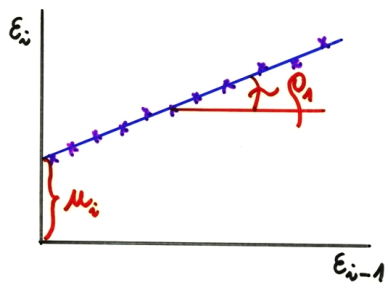
kde $u_i \sim N(0, \sigma^2)$.

a) Pro $\rho_1 = 1$ případ kumulativních chyb, který se v chemii vyskytuje často.

b) Pro $\rho_1 \leq 1$ jde o autokorelační koeficient 1. řádu.

Test:

1. Grafická indikace autokorelace:



2. Waldův test pro ρ_1: H_0: $\rho_1 = 0$ *vs.* H_A: $\rho_1 \neq 0$

Je-li Waldovo kritérium $W_a = \dfrac{n\,\hat{\rho}_1^2}{1 - \hat{\rho}_1^2} < \chi^2(1)$, H_0 je přijata.