6. Lineární regresní modely

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

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.

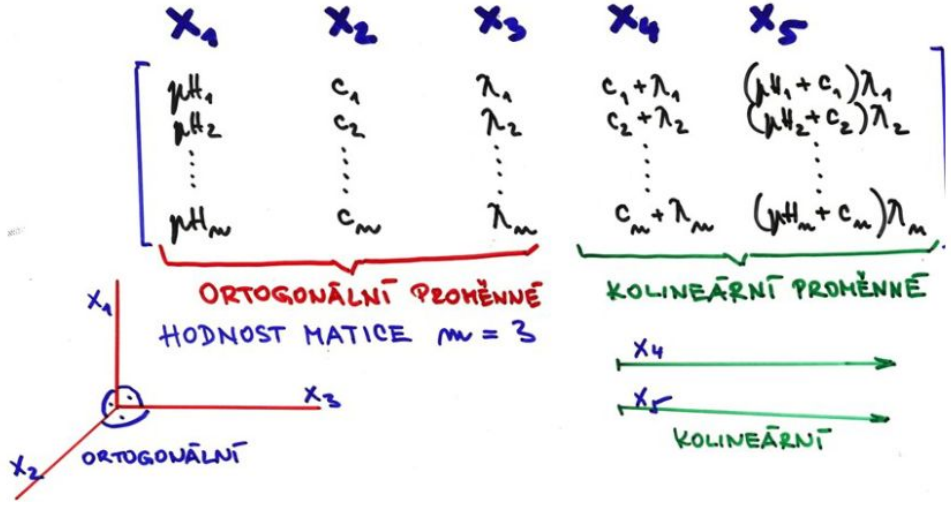
MULTIKOLINEARITA

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

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**.



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.

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.

Metoda racionálních hodností

Matici **R** (symetrická) vyjádříme:

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

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

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

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

Scottova testační charakteristika 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) $0.33 \le M_T \le 0.8$, model je málo vyhovující a je ***vhodná*** **jeho úprava.**

c) $M_T < 0.33$, není model ovlivněn multikolinearitou a není **třeba ho upravovat.**

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

Příklad 6.46 *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	50	60
ln$\gamma_{\pm}$	0.8067	0.8038	0.8000	0.7946	0.7927	0.7867	0.7828

T [°C]	70	80	90
ln$\gamma_{\pm}$	0.775	0.769	0.765

Řešení: Pro model

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

1. **Klasická MNČ** (parametr vychýlení P je nastaven na $P = 10^{-35}$): určila

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

Koeficient determinace $\hat{R}^2 = 0.9957$,
Kvadratická chyba predikce MEP $= 3.507 \cdot 10^{-6}$,
Kritérium AIC = -132.26,
Det(**R**) $= 3.97 \cdot 10^{-4}$ a číslo podmíněnosti K = 1989.73.

T-testy významnosti parametrů β_1, β_2, β_3 (pro $\alpha = 0.05$): β_2 a β_3 jsou statisticky nevýznamné.

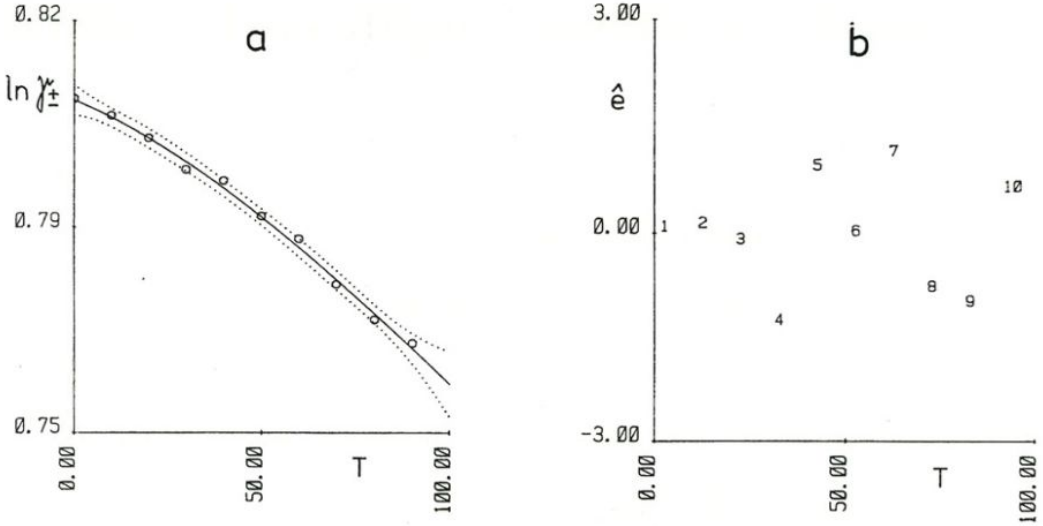
Charakteristiky multikolinearity: (platí pro $VIF_j > 10$ jde o silnou multikolinearitu)

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

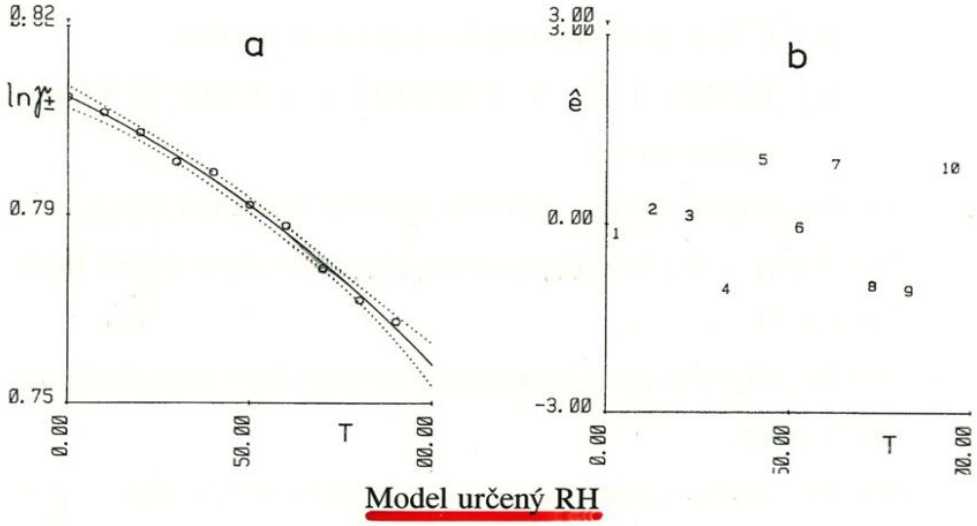
2. **Metoda racionálních hodností** (parametr vychýlení $P = 0.05$): určil

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

Koeficient determinace $\hat{R}^2 = 0.9955$,
Střední kvadratická chyba predikce MEP $= 2.364 \cdot 10^{-6}$
Kritérium AIC = -131.7.
T-testy významnosti parametrů β_1, β_2, β_3 (pro $\alpha = 0.05$): pouze β_3 je statisticky nevýznamný.



Model určený MNČ



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.