6. Lineární regresní modely

MULTIKOLINEARITA

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

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

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.

Vybrané předpoklady MNČ

1. Regresní parametry β mohou teoreticky nabývat **libovolných** hodnot.

2. Regresní model je **lineární v parametrech**.

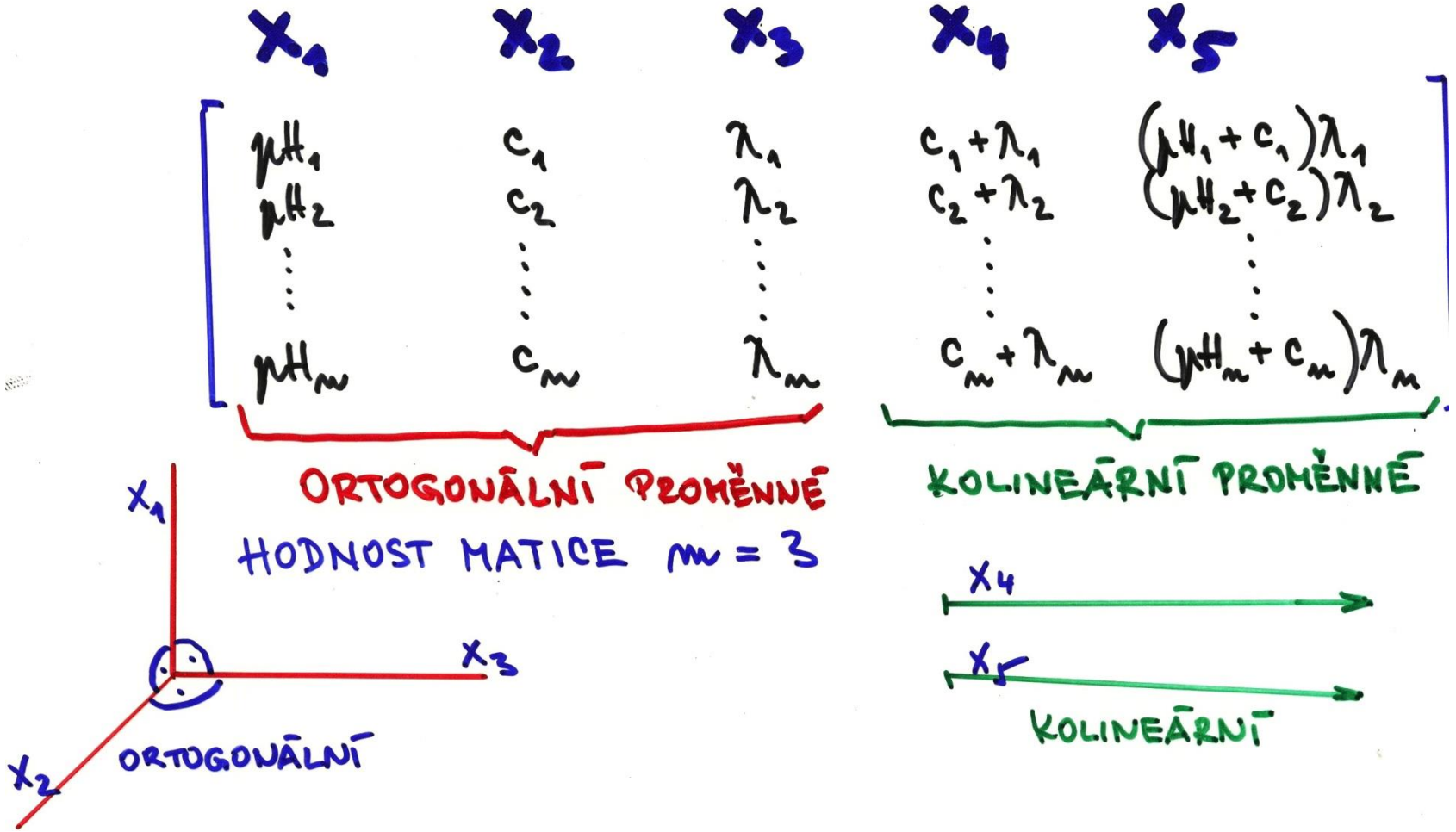
3. Jednotlivé nezávislé proměnné jsou skutečně vzájemně nezávislé, tedy mezi nimi nedochází k tzv. **multikolinearitě**.

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.

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
- řešení

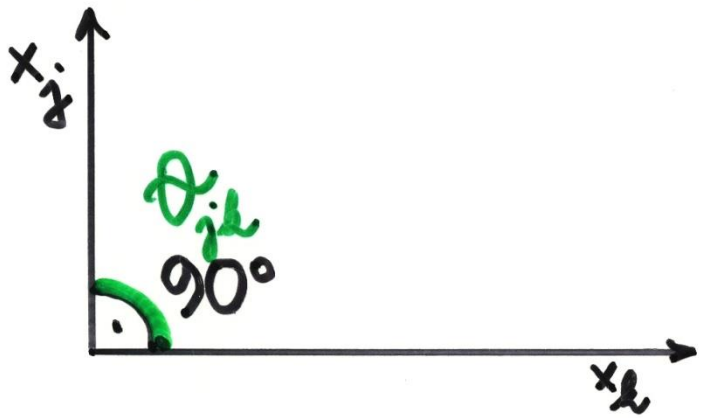
K odstranění nebo zmenšení vlivu multikolinearity může vést:

- snížení počtu nezávisle proměnných
- použití jiného modelu
- použití jiné metody výpočtu (obvykle metody regrese hlavních komponent – PCR)

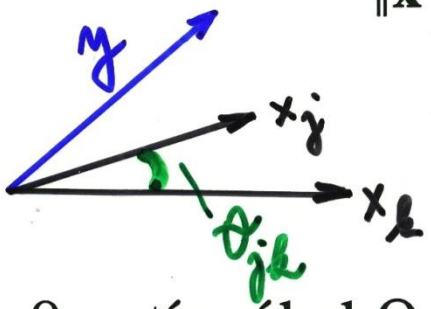
Multikolinearita

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^o.

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

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

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

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 $\mathbf{X}$ 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 $\mathbf{c}$ 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

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).

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

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

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 $F_R = \frac{\sum_{j=1}^{m-1} T_j^2}{m-1} = T_S$, kde

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

Scottova testační charakteristika k posouzení stupně multikolinearity:

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

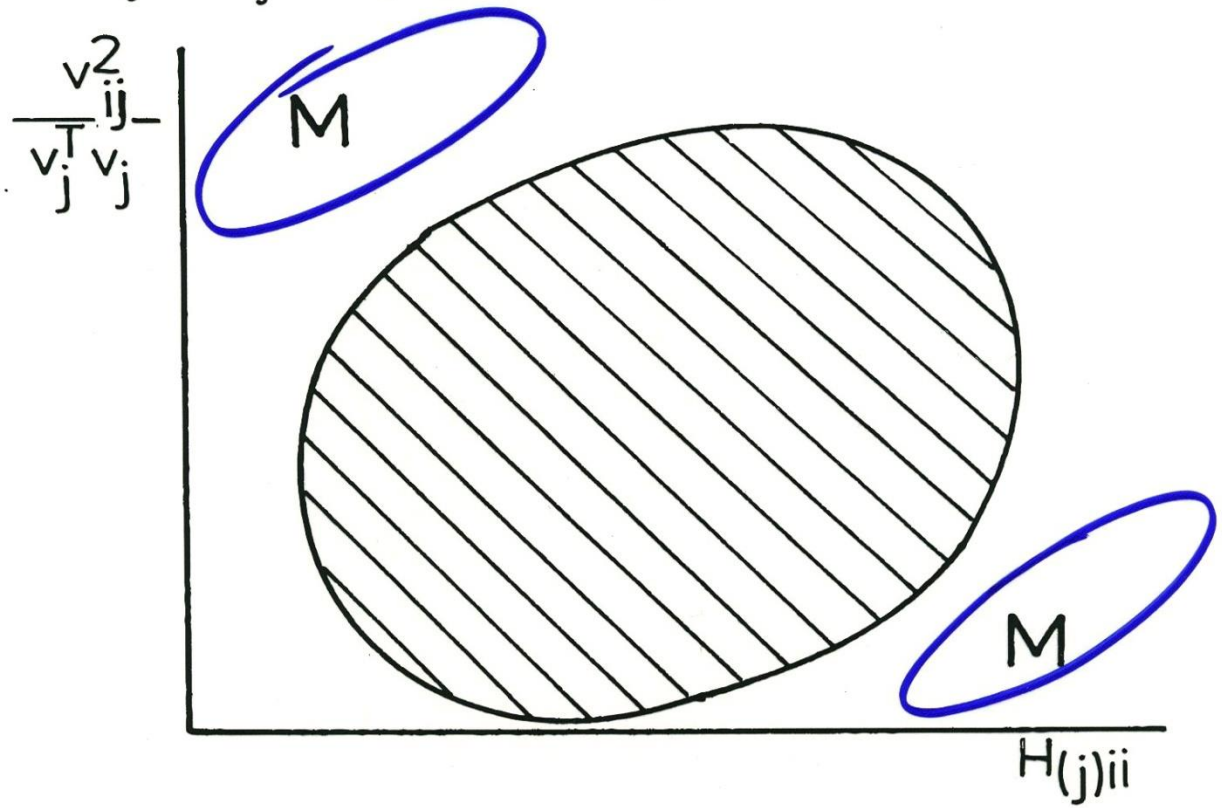
a) $M_T > 0.8$, model je nevyhovující a je ***zapotřebí*** provést **úpravu.**

b) $0.33 \leq M_T \leq 0.8$, 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 vynesením $v_{ij}^2/(\mathbf{v}_j^T \mathbf{v}_i)$ 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

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 blízká nule (multikolinearita) a běžné metody (MNČ) zcela selhávají.

III. skupina regresních úloh: některá vlastní čísla $\lambda_j = 0$ jsou rovna nule.
Pak je matice $\mathbf{X}^T\mathbf{X}$ nebo $\mathbf{R}$ singulární a **nelze** ji invertovat.

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

$\frac{W}{E} > P$

Řešení:

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

$$\mathbf{u} = \frac{W - E\,P}{\lambda_\omega}$$

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

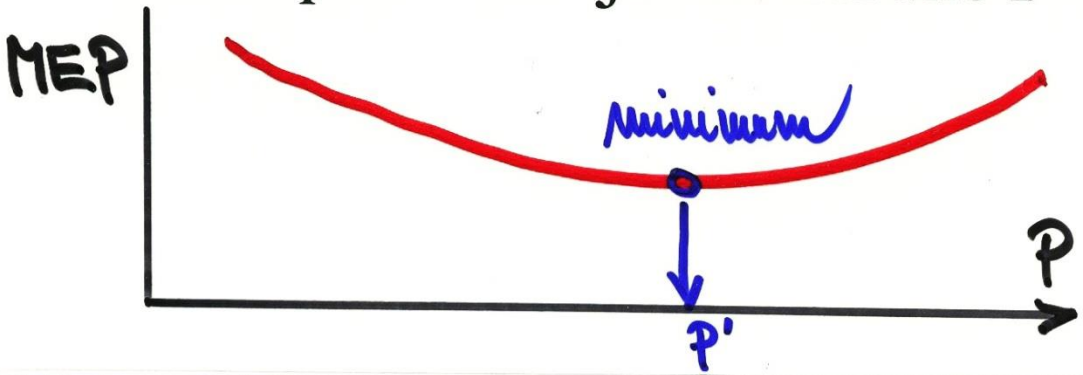
- to je však doprovázeno růstem ***vychýlení odhadů*** a poklesem vícenásobného korelačního koeficientu.

- vychýlení odhadů je zde způsobeno zanedbáním sčítanců v rovnicích normovaných odhadů parametrů při $\omega > 1$.

- 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, $\varepsilon = f(\lambda)$ polynomem druhého stupně

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

Data: n = 15; m = 3

ε [dm^3 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 [dm^3 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.

Příklad 6.9 *Aproximace absorpčního spektra polynomem*

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

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

Data: n = 15; m = 3

ε [dm^3 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 [dm^3 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.

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.

Příklad 6.46 *Multikolinearita u teplotní závislosti aktivitního koeficientu*
Závislost logaritmu středního aktivitního koeficientu $\ln \gamma_{\neq}$ 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
$\ln\gamma_{\neq}$	0.8067	0.8038	0.8000	0.7946	0.7927	0.7867	0.7828

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

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 predikce MEP $= 3.507 \cdot 10^{-6}$,
Kritérium AIC $= -132.26$,
Det($\mathbf{R}$) $= 3.97 \cdot 10^{-4}$ 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

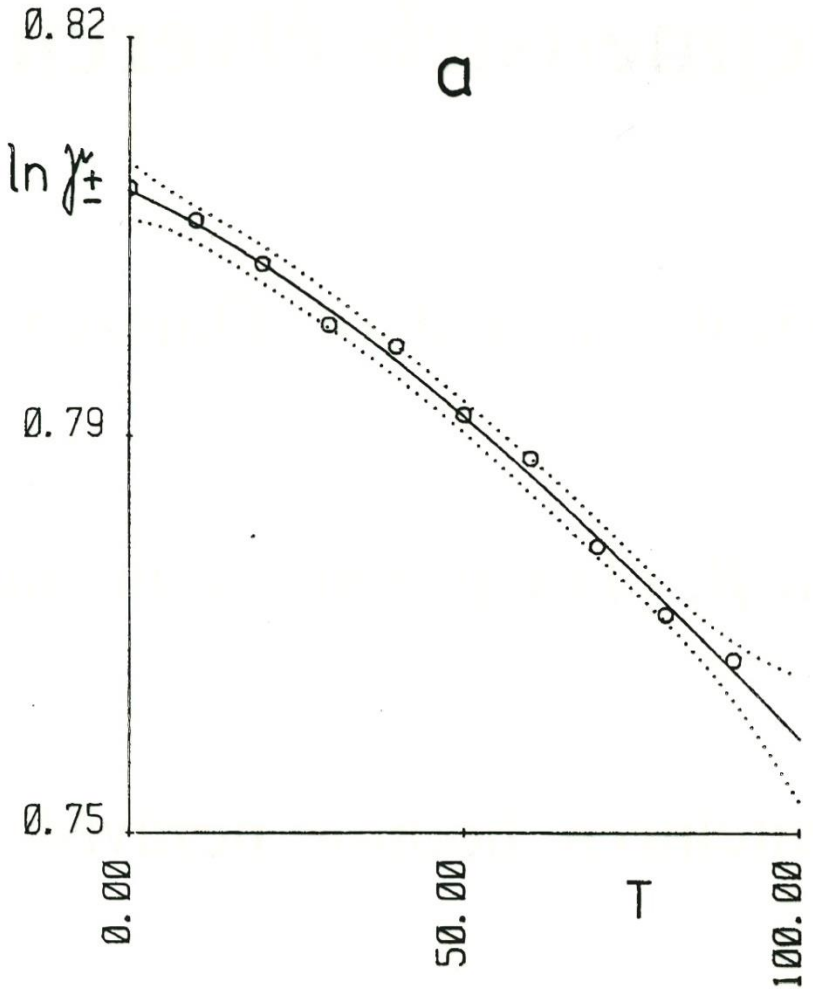
$$\ln \gamma_{\neq} = 0.807\ (\pm 8.72 \cdot 10^{-4}) - 3.22 \cdot 10^{-4}\ (\pm 3.28 \cdot 10^{-5})\ T - 1.476 \cdot 10^{-6}\ (\pm 7.18 \cdot 10^{-8})\ T^2 - 2.837 \cdot 10^{-9}\ (\pm 3.314 \cdot 10^{-9})\ T^3$$

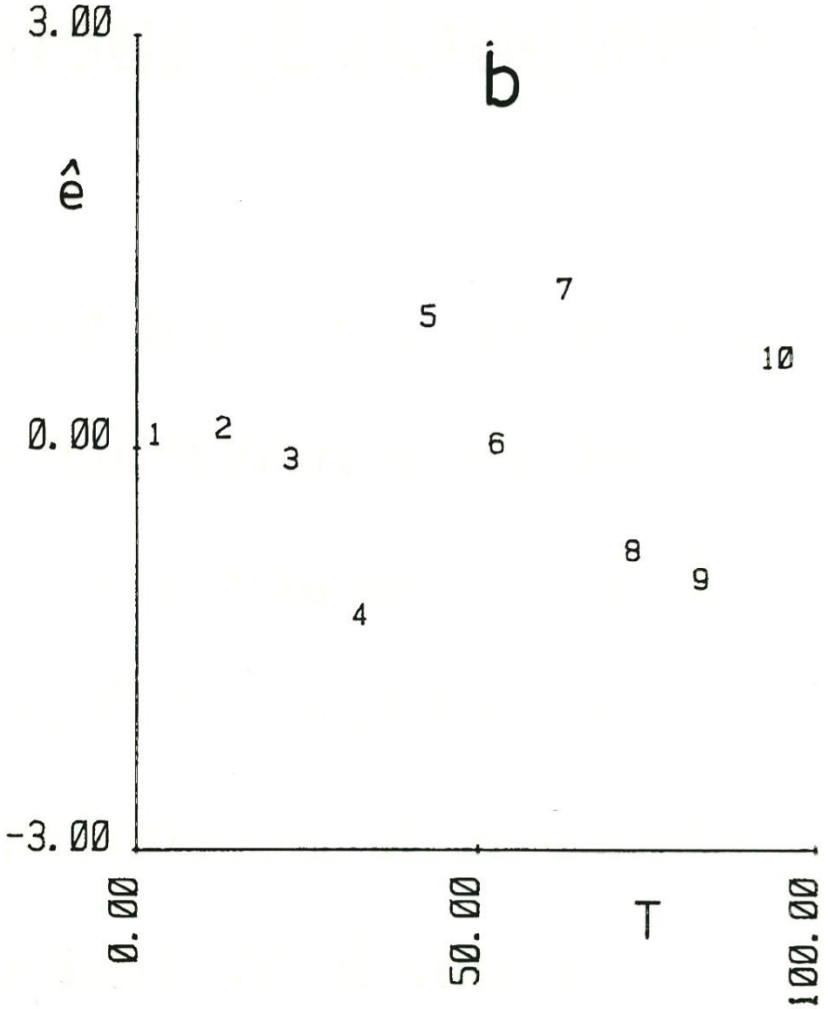
Koeficient determinace $\hat{R}^2 = 0.9955$,

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

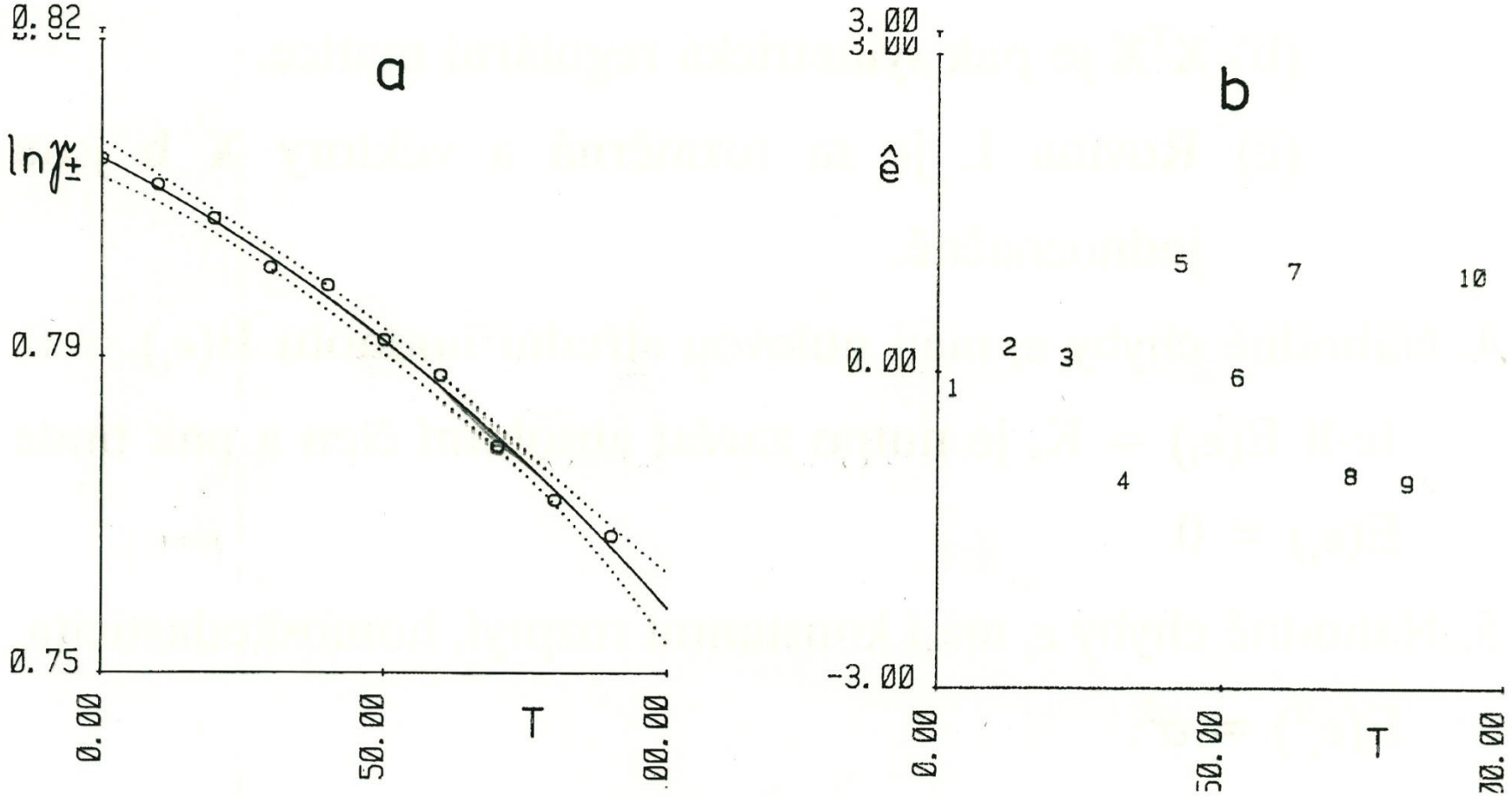
Kritérium AIC = -131.7.

T-testy významnosti parametrů 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 $\hat{\mathbf{R}}^2$),
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

$$E(y/x) = \sum_{j=1}^{6} b_j x^j + b_7$$

Cílem je vyčíslení hodnoty y v počátku čili odhadu parametru β_7.

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	$9.2 \cdot 10^{-3}$	$3.2 \cdot 10^{-5}$

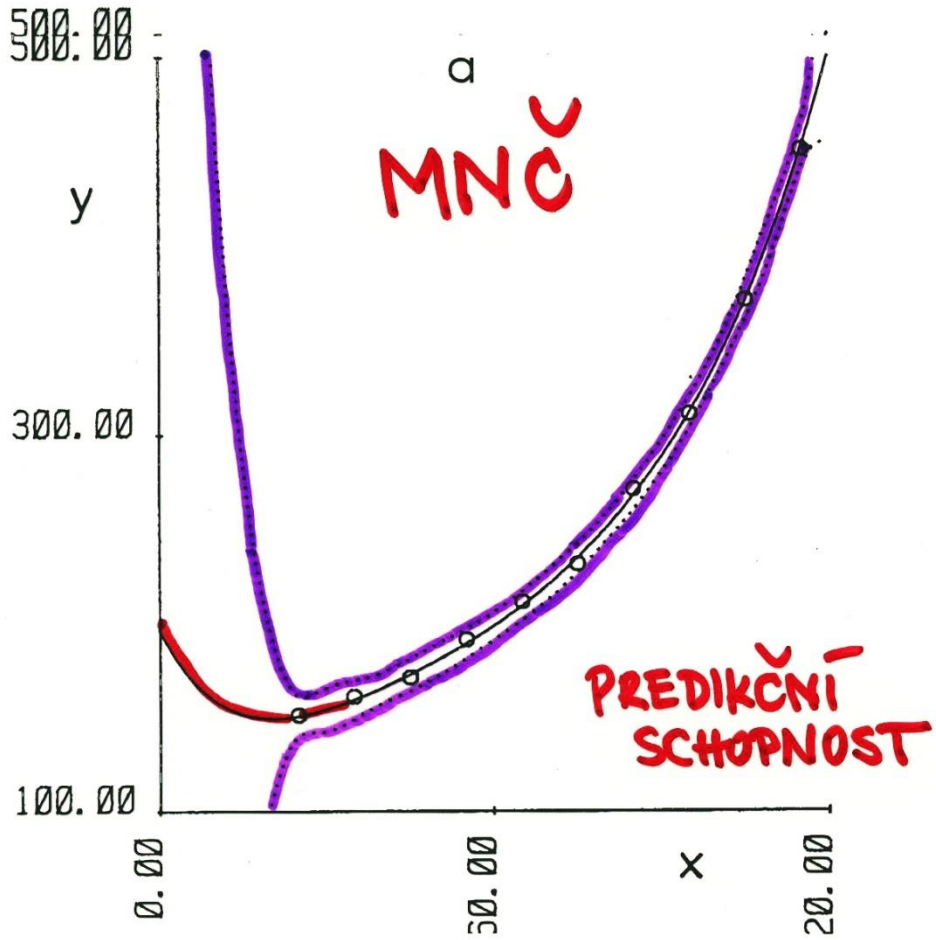
Metoda	b_4	b_5	b_6
MNČ	$5.30 \cdot 10^{-5}$	$-2.9 \cdot 10^{-7}$	$6.94 \cdot 10^{-10}$
RH	$-5.3 \cdot 10^{-8}$	$3.9 \cdot 10^{-9}$	$4.5 \cdot 10^{-11}$

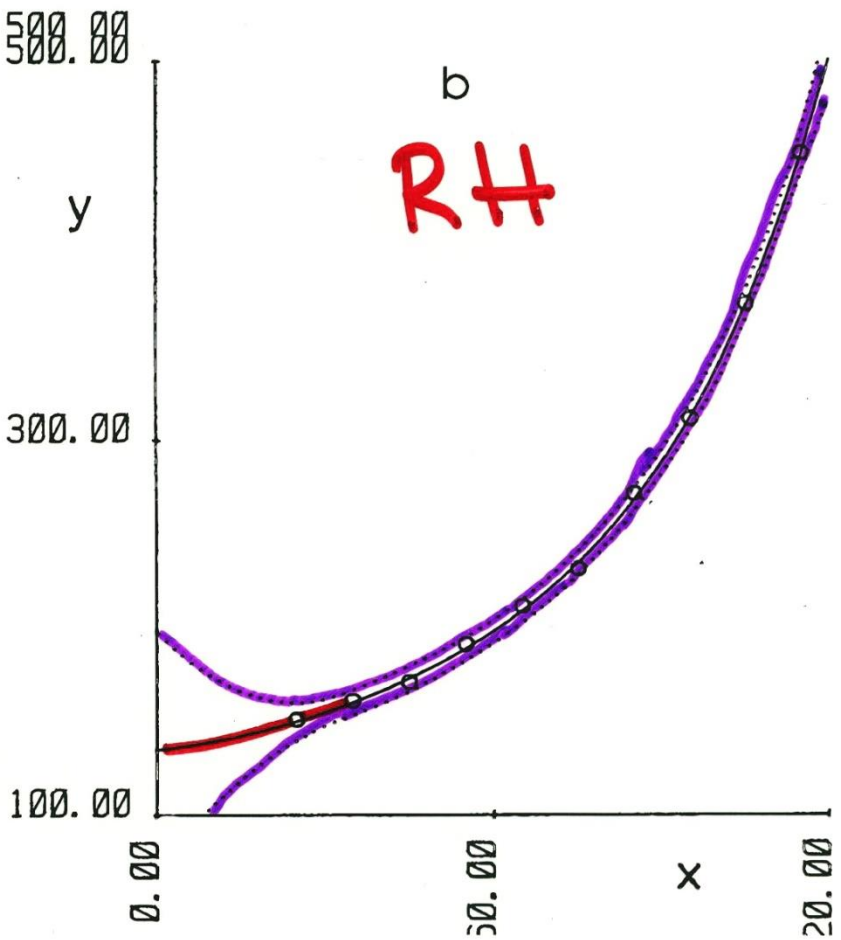
- Odhad parametru b_7 je větší než hodnota y_1, což ukazuje, že model prochází mezi počátkem a bodem (x_1, y_1) minimem.

- Konfidenční intervaly hned za rozsahem experimentálních dat jsou příliš široké, takže neumožňují predikci y mimo rozsah měření.

- Odhady určené metodou racionálních hodností jsou vychýlené.

- Parametr b_7 je však nižší než y_1 a konfidenční intervaly ukazují na možnost predikce i mimo rozsah měření.





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.