Faktory ovlivňující použití váhové funkce v exponenciálním regresním modelu

Bohumil Maroš, FSI VUT Brno

Marie Budíková, PřF MU Brno

Finanční matematika v praxi II, Podlesí, 11. – 13.9.2012

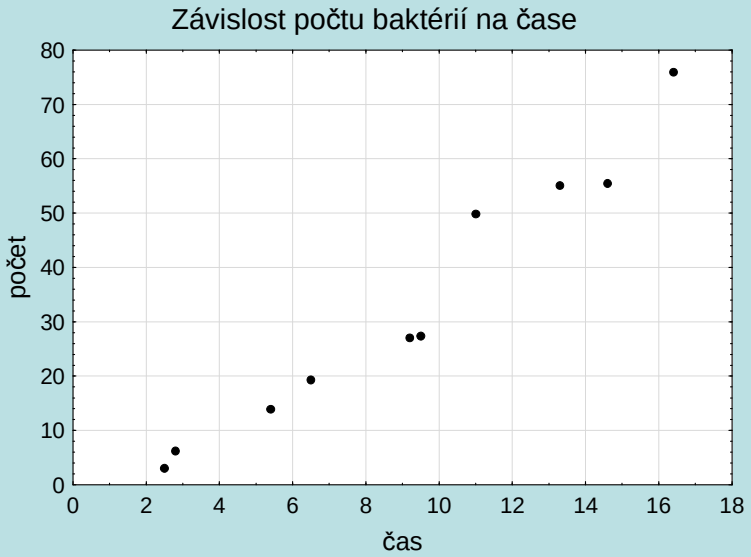
Motivace: Množení baktérií v tekutém prostředí

Byl sledován počet baktérií (závisle proměnná veličina y, v tisících v krychlovém centimetru) v závislosti na čase (nezávisle proměnná veličina x, v hodinách):

Data

x	y
2,5	3,03
2,8	6,213
5,4	13,91
6,5	19,305
9,2	27,037
9,5	27,381
11	49,845
13,3	55,069
14,6	55,453
16,4	75,943

Graf

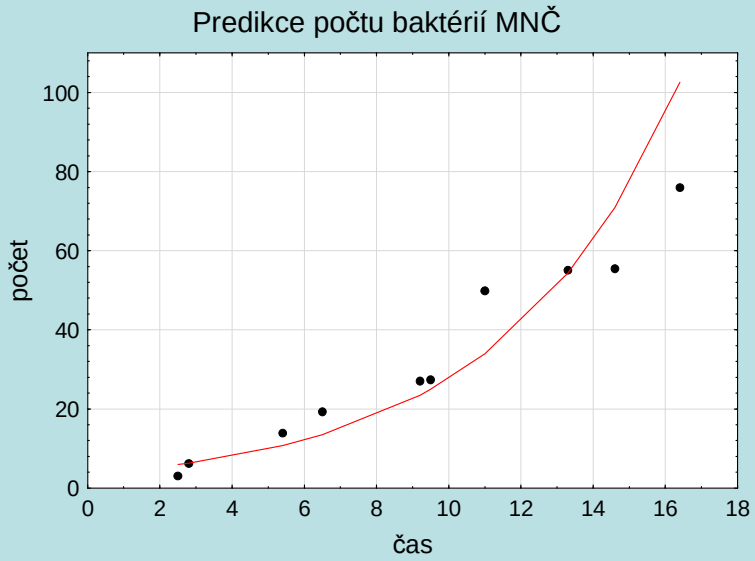


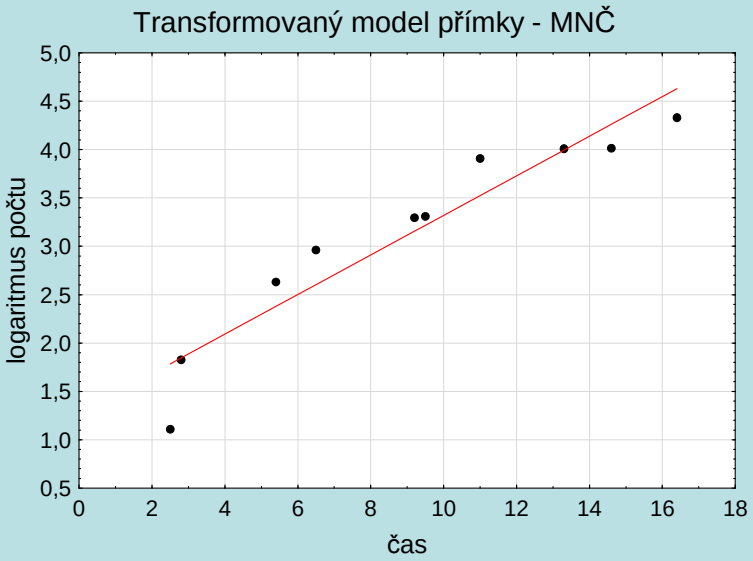
Je známo, že závislost počtu baktérií na čase lze popsat modelem
$y = \exp(\beta_0 + \beta_1 x)$.
a) Odhadneme parametry modelu.
b) Zjistíme, za jak dlouho dojde ke zdvojnásobení počtu baktérií.

Řešení:
Model linearizujeme: $\ln y = \beta_0 + \beta_1 x$. Metodou nejmenších čtverců získáme odhady $b_0 = 1{,}273$ a $b_1 = 0{,}205$ regresních parametrů β_0, β_1. Reziduální součet čtverců je $S_0 = 1272$.
Odhad počtu baktérií v okamžiku $x_0 = 0$ je $N_0 = \exp(b_0) = \exp(1{,}273)$.
Ke zdvojnásobení počtu baktérií dojde v okamžiku x_2, tedy

$$2N_0 = N_0 \exp(b_1 x_2) \Rightarrow \ln 2 = b_1 x_2 \Rightarrow x_2 = \frac{\ln 2}{b_1} = \frac{\ln 2}{0{,}205} = 3{,}38\,\text{h} = 3\,\text{h}\,23\,\text{min}$$

Grafické znázornění výsledků:





Otázka: Jdou dosažené výsledky zlepšit?

Odpověď: ANO, použitím váhové funkce.

Stručná teorie:

Použili jsme model $y = \exp(\beta_0 + \beta_1 x)$, který jsme linearizovali $\ln y = \beta_0 + \beta_1 x$.

Přitom však dojde k porušení normality náhodných odchylek a homogenity jejich rozptylů.

Proto pro odhad regresních parametrů místo obyčejné MNČ použijeme váženou MNČ s váhovou funkcí $w(y) = y^2$.

Získáme reziduální součet čtverců S_1.

Mnohdy zjistíme, že hodnota reziduálního součtu čtverců s použitím váhové funkce poklesla, tzn. $S_1 < S_0$. Iteračním způsobem budeme pokračovat a při r-té iteraci získáme reziduální součet čtverců S_r. Nejmenší dosaženou hodnotu označíme S_t.

Vliv váhové funkce posuzujeme pomocí podílu

$$RI_r = \frac{S_r}{S_0} 100\% ,$$

r = 0, 1, 2, …, t, který nazveme relativní zlepšení reziduálního součtu čtverců.

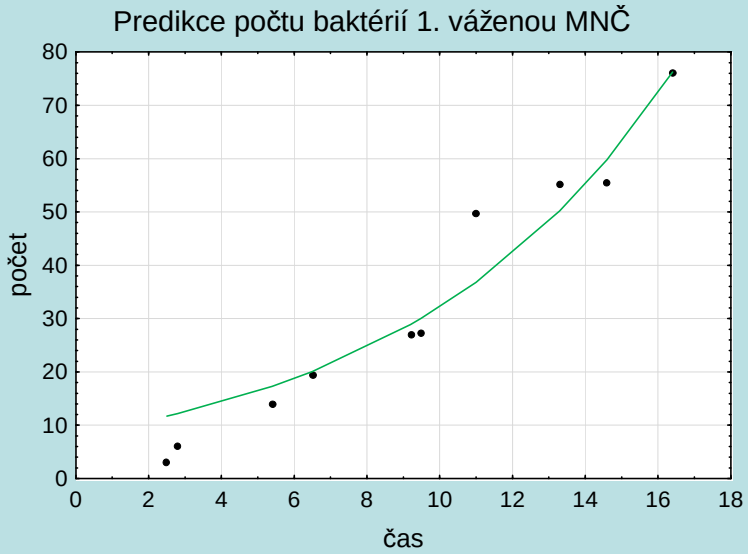
Vraťme se k motivačnímu příkladu.

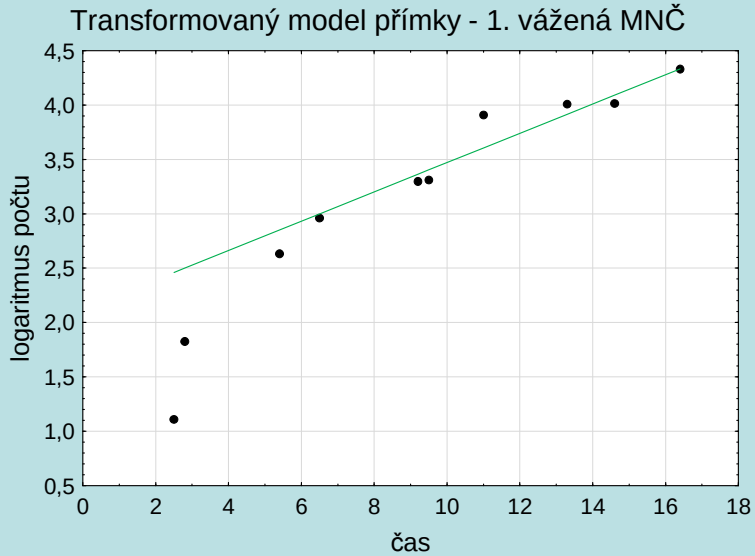
První použití váhové funkce:

$b_0 = 2{,}123$, $b_1 = 0{,}135$, $S_1 = 346{,}375$,

$$RI_1 = \frac{S_1}{S_0}100\% = \frac{346{,}375}{1272}100\% = 27{,}23\%$$

Grafické znázornění výsledků při 1. použití váhové funkce:





Druhé použití váhové funkce:

$b_0 = 1{,}874$ a $b_1 = 0{,}152$

$S_2 = 352{,}708$

$$RI_2 = \frac{S_2}{S_0}100\% = \frac{352{,}708}{1272}100\% = 27{,}74\%$$

Třetí použití váhové funkce:

$b_0 = 1{,}946$ a $b_1 = 0{,}147$

$S_3 = 340{,}55$

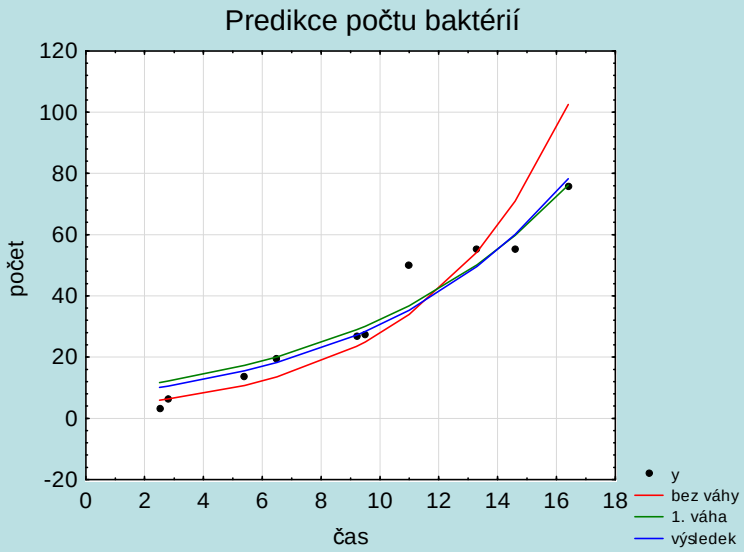
$$RI_3 = \frac{S_3}{S_0}100\% = \frac{340{,}55}{1272}100\% = 26{,}78\%.$$

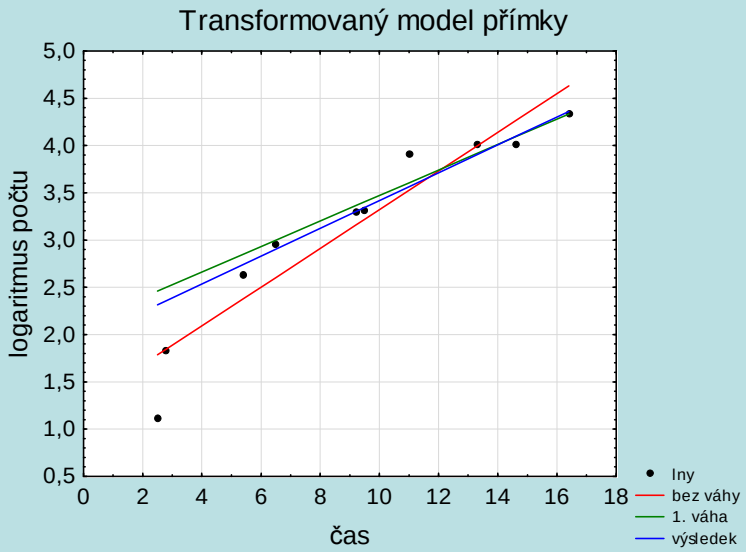
Při dalších iteracích už dochází ke zvětšování reziduálního součtu čtverců. Třetí iteraci budeme tedy považovat za konečnou.

Ke zdvojnásobení počtu baktérií dojde v okamžiku

$$x_2 = \frac{\ln 2}{b_1} = \frac{\ln 2}{0{,}147} = 4{,}72\text{h} = 4\text{h}43\,\text{min}$$

Grafické znázornění výsledků: bez použití váhové funkce, při 1. a 3. použití váhové funkce





Pro porovnání: Odhad parametrů Levenbergovou - Marquardtovou metodou

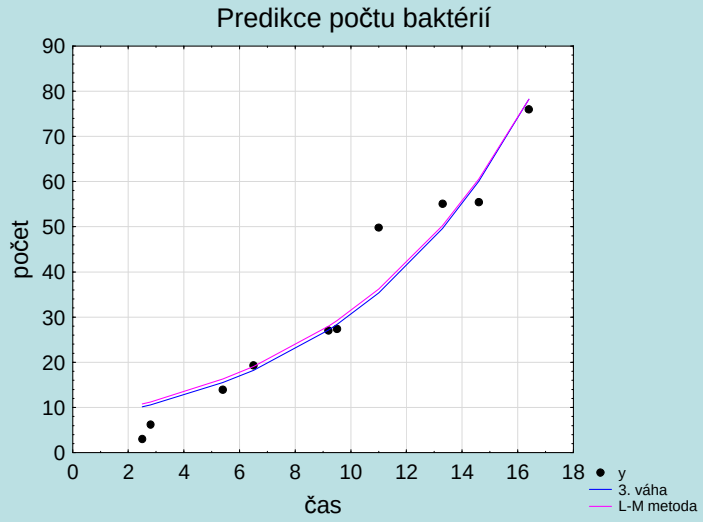
(počáteční aproximace $b_0 = b_1 = 0{,}1$, 10 iterací):

$b_0 = 2{,}020$ a $b_1 = 0{,}143$, $S_E = 335{,}08$

$$RI = \frac{S_E}{S_0}100\% = \frac{335{,}08}{1272}100\% = 26{,}34\%$$

Ke zdvojnásobení počtu baktérií dojde v okamžiku $x_2 = \frac{\ln 2}{b_1} = \frac{\ln 2}{0{,}143} = 4{,}85\text{h} = 4\text{h}51\text{min}$

Srovnání výsledků při 3. použití váhové funkce a při L-M metodě:



Shrnutí výsledků:

Model $y = \exp(\beta_0 + \beta_1 x)$	Bez váhy	1. váha	2. váha	3. váha	L-M metoda
Odhad b_0	1,273	2,123	1,874	1,946	2,020
Odhad b_1	0,205	0,135	0,152	0,147	0,143
Reziduální součet čtverců	1272	346,38	352,71	340,55	335,08
Relativní zlepšení (%)	100	27,24	27,74	26,78	26,34
Odhad N_0	3,57	8,36	6,51	7	7,54
Čas x_2	3 h 23 min	5 h 8 min	4 h 34 min	4 h 43 min	4 h 51 min

Jaké faktory mohou ovlivnit použití váhové funkce v exponenciálním modelu $y = \exp(\beta_0 + \beta_1 x)$, kde $a \leq x \leq b$?

Ovlivňující faktory jsou např.:

1. minimální teoretická hodnota závisle proměnné veličiny Y
2. maximální teoretická hodnota závisle proměnné veličiny Y
3. střední hodnota křivosti exponenciály na $\langle a, b \rangle$,

tj. $\overline{K} = \frac{1}{b-a} \int_a^b K(x)dx$, kde $K(x) = \frac{\frac{d^2 f(x)}{dx^2}}{\left(1 + \left(\frac{df(x)}{dx}\right)^2\right)^{\frac{3}{2}}}$

4. kolísání naměřených hodnot
5. počet opakování naměřených hodnot na každé úrovni pozorování
6. počet úrovní měření
7. rozpětí hodnot na x-ové ose, tj. rozdíl b – a.

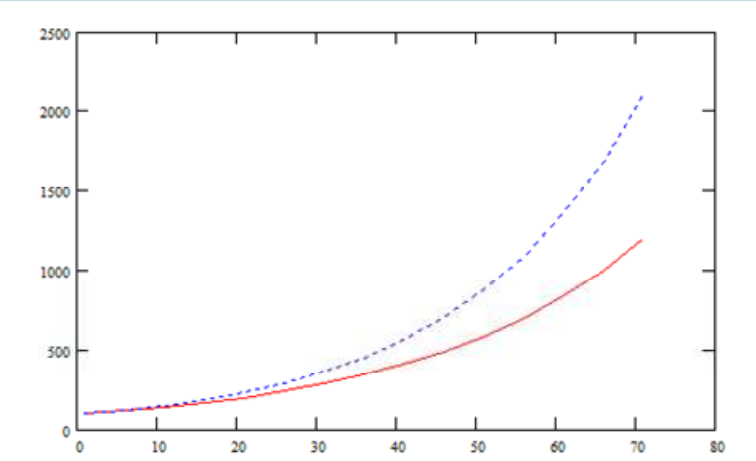
Zaměřili jsme na faktory 1 – 5 a jejich vliv jsme studovali pomocí simulací.

Popis simulací

Předpokládejme, že jistý proces se řídí přesně modelem $y = \exp(\beta_0 + \beta_1 x)$, kde $a \le x \le b$, přičemž víme, že pro $x = a$ je $y = \min$ a pro $x = b$ je $y = \max$. Z těchto dvou podmínek se dají jednoznačně určit parametry modelu: $\beta_0 = \ln(\min) - \frac{a}{b-a}\ln\left(\frac{\max}{\min}\right)$, $\beta_1 = \frac{1}{b-a}\ln\left(\frac{\max}{\min}\right)$.

Budeme nyní pro jednoduchost uvažovat stále hodnoty a = 1, b = 71 a min = 100, 200, …, 500 (5 hodnot) postupně pro max = 1200, 1300, …, 2100 (10 hodnot), takže máme 50 různých modelů.

Pro ilustraci zvolíme min = 100. Dostaneme 10 exponenciálních modelů, jejichž grafy se nacházejí mezi dvěma krajními možnostmi pro max = 1200 a max = 2100, jak vidíme na obrázku.



Simulační studie spočívá v tom, že k přesným hodnotám modelu

$y_i = \exp(\beta_0 + \beta_1 x_i)$ pro $x_i = 5*i - 4$, i=1, 2, …, 15,

se přičte náhodná složka $\varepsilon_i \approx N(0, \sigma^2)$.

Hodnoty směrodatné odchylky σ jsou ve všech případech zvoleny postupně σ = 25, 50, 75, …, 500 (20 hodnot).

To znamená, že obdržíme simulované naměřené hodnoty z 5*10*20 = 1000 různých modelů.

Tento způsob simulační studie předpokládá, že pro každou hodnotu x_i, i = 1, …, 15 obdržíme jedinou hodnotu $Y_i = y_i + \varepsilon_i$. Abychom mohli porovnat, jaký vliv budou mít opakovaná měření, provedeme simulaci měření pro všechny hodnoty x_i ještě dvakrát a třikrát, tzn.

$Y_{ij} = \exp(\beta_0 + \beta_1 x_i) + \varepsilon_{ij}$, $x_i = 5*i - 4$, i = 1, 2, …, 15, j = 1, 2

$Y_{ij} = \exp(\beta_0 + \beta_1 x_i) + \varepsilon_{ij}$, $x_i = 5*i - 4$, i=1, 2, …, 15, j = 1, 2, 3.

Každá z těchto 3000 simulací se opakuje 10 000 krát a vždy se vypočte průměrná hodnota reziduálního součtu čtverců z 10 000 opakovaných regresí.

Zkoumání vlivu vybraných faktorů na relativní zlepšení pomocí mnohonásobné regrese

Vliv jednotlivých faktorů na relativní zlepšení můžeme zjistit pomocí mnohonásobné regrese:

$$RI = \alpha_0 + \alpha_1 \cdot \min + \alpha_2 \cdot \max + \alpha_3 \cdot krivost + \alpha_4 \cdot sigma + \alpha_5 \cdot mer,$$

kde min, resp. max je minimální, resp. maximální hodnota výchozí exponenciální funkce,

krivost je střední hodnota křivosti,

sigma je směrodatná odchylka náhodné složky ε,

mer je počet opakovaných simulovaných pozorování pro každou úroveň.

Výsledek regresní analýzy se závisle proměnnou RI:

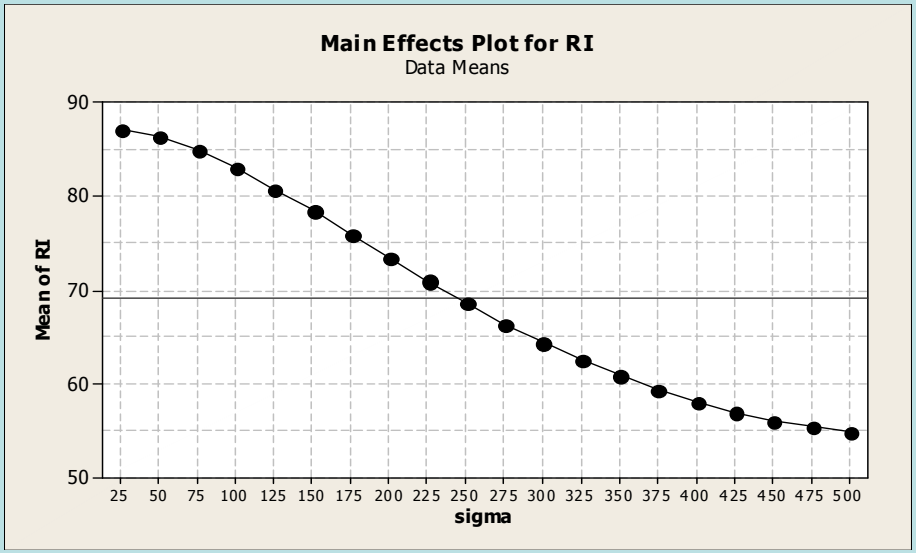
Regression Analysis: procento versus max; min; krivost; sigma; mer

```
The regression equation is
RI = 92,5 - 0,0146 max + 0,0759 min - 56,4 krivost - 0,0755 sigma + 4,38 mer


Predictor        Coef    SE Coef       T      P    VIF
Constant       92,453      1,421   65,05  0,000
max        -0,0145916  0,0004710  -30,98  0,000  1,257
min          0,075940   0,001680   45,21  0,000  3,875
krivost       -56,364      1,892  -29,80  0,000  4,132
sigma      -0,0754817  0,0008372  -90,16  0,000  1,000
mer            4,3838     0,1478   29,66  0,000  1,000


S = 6,60999   R-Sq = 90,9%   R-Sq(adj) = 90,8%


Analysis of Variance

Source            DF       SS      MS        F      P
Regression         5  1299676  259935  5949,27  0,000
Residual Error  2994   130814      44
Total           2999  1430490
```

Z tohoto výstupu je vidět, že relativní zlepšení RI významně ovlivňují všechny vybrané faktory, ale nejvíce je ovlivněno směrodatnou odchylkou *sigma*, pak hodnotou *min* a faktory *max*, *křivost* a *mer* mají přibližně stejně velký vliv. Hodnoty VIF (Variance inflation factors) ukazují, že mezi zvolenými faktory není vysoká multikolinearita.

Vliv faktoru sigma

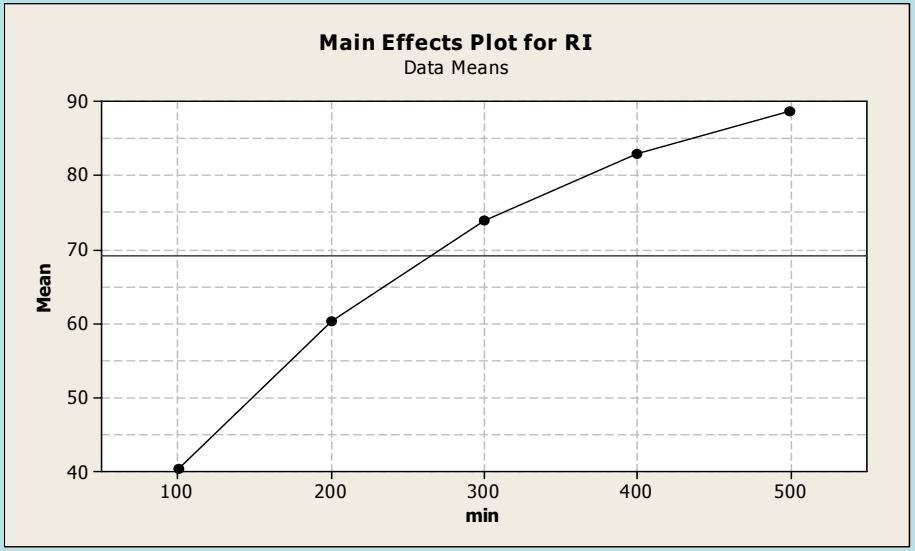
Na obrázku je znázorněn vliv směrodatné odchylky na průměrnou hodnotu relativního zlepšení reziduálního součtu čtverců pro pevně zvolené hodnoty všech ostatních faktorů.



S rostoucím sigma při konstantních hodnotách všech ostatních faktorů klesá průměrná hodnota RI (roste význam váhové funkce).

Vliv faktoru min

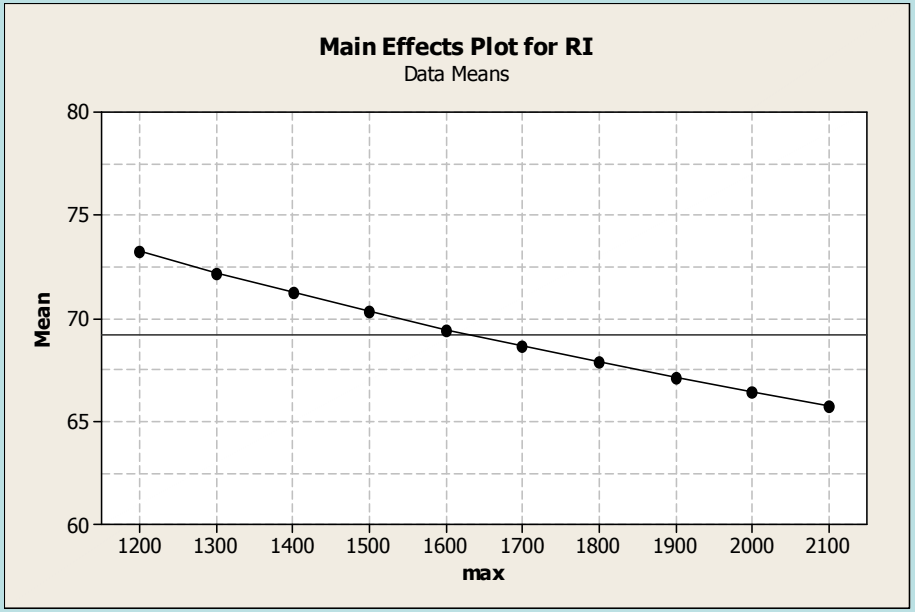
Druhým dominantním faktorem z mnohonásobné regrese se ukázal faktor min.



S rostoucím min při konstantních hodnotách všech ostatních faktorů roste průměrná hodnota RI (klesá význam váhové funkce).

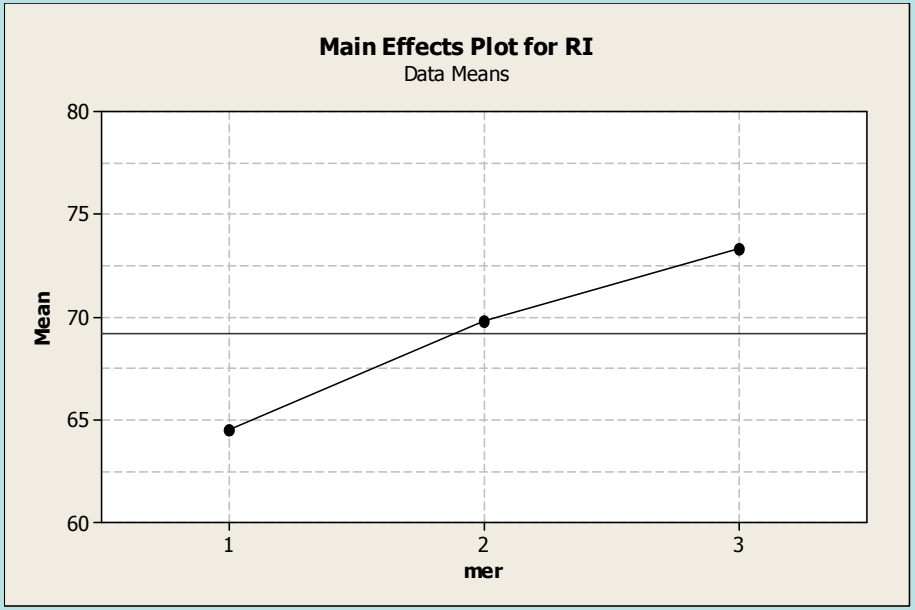
Vliv faktoru max

Na tomto obrázku můžeme sledovat vliv faktoru max na průměrnou hodnotu RI.



S rostoucí maximální teoretickou hodnotou závisle proměnné veličiny Y (faktor max) při konstantních hodnotách všech ostatních faktorů klesá průměrná hodnota RI (roste význam váhové funkce).

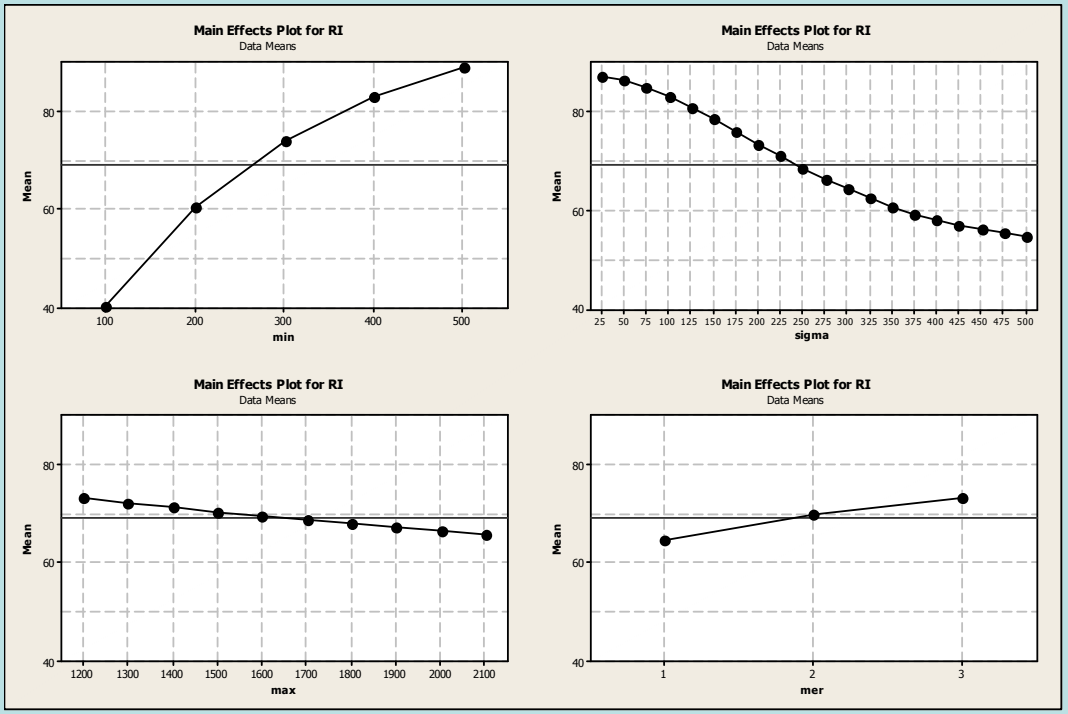
Vliv faktoru mer



S rostoucím počtem opakovaných měření (faktor mer) roste průměrná hodnota RI (klesá význam váhové funkce).

Porovnání vlivu faktorů sigma, min, max, mer

Každý z těchto čtyř faktorů má různou sílu působení na zlepšení reziduálního součtu čtverců při použití váhové funkce, jak je vidět na obrázku.



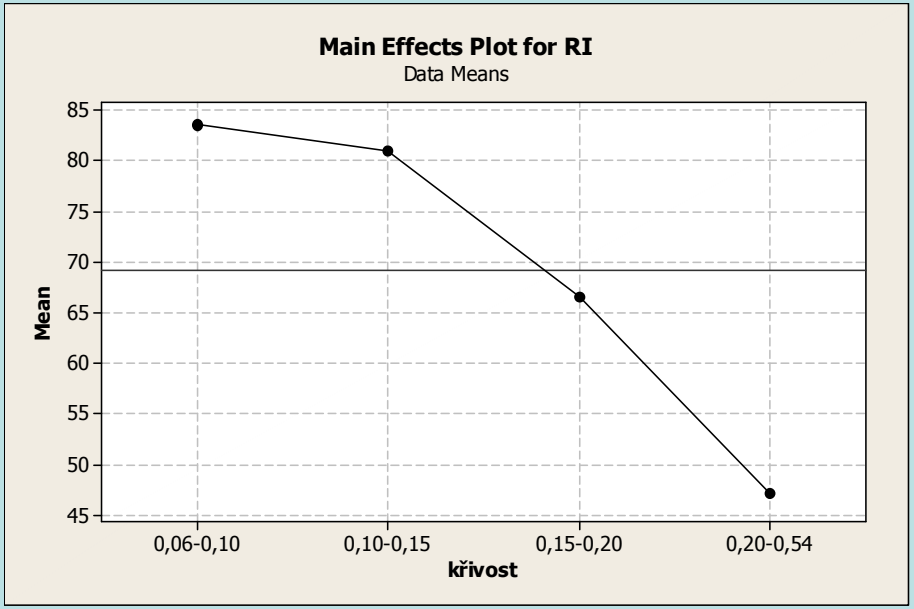
Obr. 14

Vidíme, že

- s rostoucí hodnotou min a s rostoucím počtem opakovaných měření klesá význam použití váhové funkce posuzovaný pomocí relativního zlepšení reziduálního součtu čtverců, přičemž vliv minima je výraznější než vliv počtu opakovaných měření;
- s rostoucí hodnotou max a rostoucím sigma naopak roste význam použití váhové funkce, přičemž vliv sigma je mnohem výraznější než vliv maxima.

Vliv faktoru křivost

Tento faktor je nejvíce problematický, neboť zde existuje interakce s faktory min, max a sigma. Jestliže tyto faktory zprůměrujeme, pak můžeme obdržet náhled na vliv křivosti na RI z následujícího obrázku.



Je vidět, že s rostoucí křivostí klesá průměrná hodnota RI, tudíž roste význam použití váhové funkce.

Shrnutí vlivu jednotlivých faktorů na relativní zlepšení RI

Z provedených simulací vyplývá, že při konstantních hodnotách všech ostatních faktorů:

- S rostoucím sigma klesá hodnota RI (roste význam váhové funkce).
- S rostoucím min roste průměrná hodnota RI (klesá význam váhové funkce).
- S rostoucím max faktorů klesá průměrná hodnota RI (roste význam váhové funkce).
- S rostoucím mer roste průměrná hodnota RI (klesá význam váhové funkce).
- S rostoucí křivostí klesá průměrná hodnota RI, tudíž roste význam použití váhové funkce.

Význam váhové funkce tedy roste s rostoucím sigma, max a křivost, přičemž vliv sigma je výraznější než vliv max či vliv křivosti.
Naopak, význam váhové funkce klesá s rostoucím min a mer. Faktor min má na růst hodnot RI větší vliv než faktor mer.

Literatura

Maroš, B.: Empirické modely I. CERM, Brno 2001

Mathcad software. Dostupné na http://www.dtn.mathsoft.cz/

MINITAB software. Dostupné na http://www.minitab.com/

Zvára, K.: Regresní analýza. Academia, Praha 1989.

Děkuji za pozornost.