UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Regrese podruhé a v nesnázích

Vedoucí diplomové práce:
Mgr. Jaroslav Marek, Ph.D.
Rok odevzdání: 2011

Vypracovala:
Lucie Hudcová
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně za vedení Mgr. Jaroslava Marka, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 30. března 2011

Lucie Hudcová

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce Mgr. Jaroslavu Markovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Také bych ráda poděkovala rodině a příteli, kteří mě po celou dobu studia podporovali. Dále si zaslouží poděkování můj počítač, že vydržel moje pracovní tempo, a typografický systém TEX, kterým je práce vysázena.

Obsah

Úvod a cíle práce

V klasických kurzech matematické statistiky jsou studenti na všech vysokých školách ekonomického charakteru seznámeni s metodou nejmenších čtverců a aproximací dat přímkou. Tam, kde je kurzu statistiky věnován větší prostor, jsou obeznámeni i s aproximací obecným polynomem. Název práce nám říká, že následující text je určen těm čtenářům, kteří už základy matematické statistiky a lineární algebry mají, ale dostanou se se svými znalostmi do úzkých.

Obvykle studenti vůbec netuší, že existují metody jiné, a ani neznají důvody, které vedou k preferování metody nejmenších čtverců. Část práce proto budu věnovat historii regresní analýzy. Na historické úloze měření délky jednoho zeměpisného stupně, související s určením tvaru Země, seznámíme čtenáře s Mayerovou metodou průměrů, Boškovičovou přímkou, Lambertovou metodou a Laplaceovou metodou.

Užití metody nejmenších čtverců je možné jen při splnění nutných předpokladů. Pokud x není dáno deterministicky, lze užít komplikovanější ortogonální metodu. V případě problémů s invertabilitou matice $\mathbf{X}\mathbf{X}'$ je možné problémy odstranit pomocí hřebenové regrese. Dále je třeba řešit heteroskedasticitu či se vypořádat s odlehlými pozorováními. Kromě odhadů parametrů získáváme i odhad jejich přesnosti. Přesnost odhadů můžeme ovlivnit vhodnou volbou bodů x, ve kterých je měření provedeno. Toto je předmětem teorie designu experimentu.

Dále se budeme věnovat aproximaci dat nelineární funkcí. Opět lze využít metodu nejmenších čtverců, ale funkci je třeba nahradit jedním, respektive dvěma členy Taylorova rozvoje. Dostaneme tak lineární, resp. kvadratický odhad v linearizovaném modelu. Linearizace ale nemůže být provedena vždy a je třeba, aby počáteční řešení splnilo jistá kritéria a leželo v tzv. linearizační oblasti.

1 Historie

Koncem 18. století se ve světě vyskytovala mnohá astronomická pozorování planet a také četná měření Země, Měsíce a Slunce. Dříve používané metody pro zpracování měření nebyly dostačující, a tak se hledaly metody nové, lépe vysvětlující zákonitosti vesmíru. Velkým impulsem se pro vědce staly soutěže, vypsané v různých zemích, s cílem získat metodu pro určování zeměpisné délky. Jedno z řešení bylo založeno na pohybových rovnicích Měsíce a na určení lunárních tabulek. V podkapitolách lze nalézt metody od Tobiase Mayera, Rogera Josipa Boškoviče, Johanna Heinricha Lamberta a Pierra Simona Laplace. Podrobněji popsané je lze nalézt v literatuře [10, 11]. V tomto období také vzniká nejlepší metoda – *metoda nejmenších čtverců* – MNČ. Zdrojem následující části byly [14, 26].

1.1 Mayerova metoda průměrů

Nejvýznamnější dílo německého samouka matematiky a kresliče map Tobiase Mayera (*17. 2. 1723) bylo *Kosmographische Nahrichten*, kde popsal tehdejší astronomické práce. Více se však ale proslavil tím, že sepsal lunární tabulky, podle kterých se mohla určit zeměpisná délka. Ty byly přesné k určení pozice Měsíce na 5 minut a zeměpisné délky na moři na půl stupně, což byl velký úspěch.

Mayer dospěl k tomu, že nejideálnější je do zkoumání zahrnout všechna pozorování. Dále je pak dobré použít všechny kombinace sestavených soustav rovnic a nakonec zprůměrovat výsledky – kvůli zprůměrování metodu nazýváme *Mayerova metoda průměru*.

Metodu budeme demonstrovat na historické úloze měření délky jednoho zeměpisného stupně. Cílem je odhadnout parametry přímky $y = \alpha + \beta x$, kde y je délka jednostupňového oblouku v jednotkách zvaných toise (1 toise $\approx$ 6,39 stopy tj. 1,947 metru) a $x = sin^2(L)$, přičemž L je zeměpisná šířka středu oblouku. Naměřená Mayerova data nám ukazuje tabulka.

i	zeměpisná poloha	L	x	y
1	Quito	$0°0'$	0	56751
2	Mys Dobré Naděje	$33°18'$	0,2987	57037
3	Řím	$42°59'$	0,4648	56979
4	Paříž	$49°23'$	0,5762	57074
5	Laponsko	$66°19'$	0,8386	57422

Tabulka 2.1: Naměřené hodnoty

Podle metody je nejlepší užít všechny možné kombinace dvojic naměřených hodnot. Získáváme tak 10 různých odhadů parametrů pro zadanou přímku. Odhady jsme dostali pomocí vztahů

$$\beta_{ij} = \frac{y_j - y_i}{x_j - x_i}, \qquad \alpha_{ij} = y_i - \frac{y_j - y_i}{x_j - x_i} x_i, \qquad i \neq j, \qquad i, j = 1, \ldots, n. \tag{1}$$

Výsledky výpočtu po dosazení do vzorců (1) za různá i a j vidíme v tabulce.

i	j	β_{ij}	α_{ij}	rezidua (pro ostatní místa)
1	2	957	56751	-217, -229, -132
1	3	491	56751	139, 40, 260
1	4	561	56751	119, -33, 201
1	5	800	56751	47, -144, -138
2	3	-349	57141	-390, 134, 574
2	4	133	56997	-246, -80, 313
2	5	713	56824	-73, -176, -161
3	4	853	56583	168, 200, 124
3	5	1185	56428	323, 255, -37
4	5	1326	56310	441, 331, 53

Tabulka 2.2: Získané výsledky

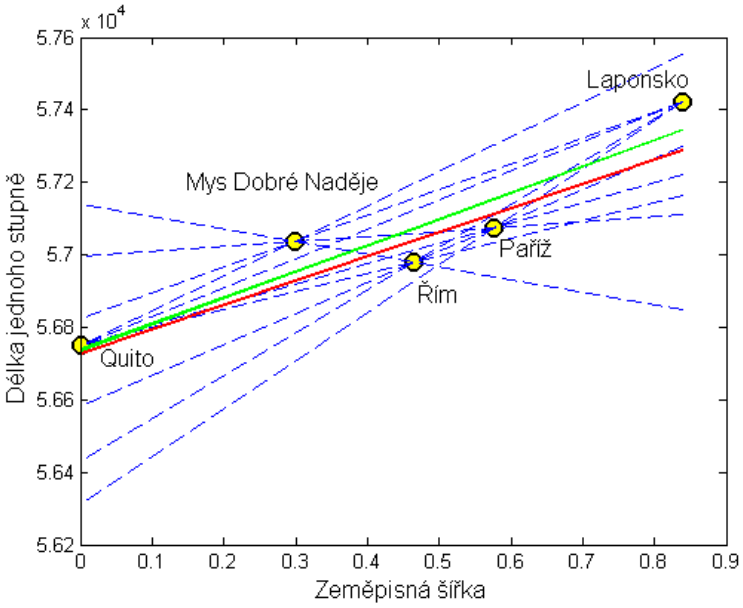
Nyní vypočítáme odhady parametrů. Určíme je jako průměry odhadů parametrů z předchozí tabulky. Získáváme $\widehat{\alpha} = 56729$ a $\widehat{\beta} = 667$ s průměrnými rezidui pro ostatní místa $22,\ 109,\ -60,\ -39,\ 134$. Regresní přímka má tedy tvar $y = 56729 + 667x$.

V dnešní době se užívají hlavně dvě kritéria pro rezidua, která by měla být co nejmenší. Prvním z nich je, že absolutní součet reziduí by měl být co nejmenší, a druhé požaduje, aby součet druhých mocnin chyb odhadnutého řešení byl co nejmenší.

Reziduální součet čtverců RSČ $= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ pro Mayerovu metodu je 35421 a absolutní součet reziduí $L = \sum_{i=1}^{n} \mid Y_i - \widehat{Y}_i \mid = 364$. Tyto kritéria budeme sledovat i u dalších metod.

Při užití klasické metody nejmenších čtverců bychom získali regresní přímku $y = 56737 + 723x$. RSČ pro MNČ je 28630 a $L = 350$. Vidíme, že MNČ je určitě lepší, protože má menší reziduální součet čtverců i menší absolutní součet reziduí než Mayerova metoda.

Na obrázku (1) jsou znázorněny všechny přímky určené řešeními dané soustavy rovnic (modře čárkovaně), průměrná Mayerova přímka (červeně), regresní přímka odhadnutá MNČ (zeleně) a body na Zemi, ve kterých bylo měřeno (žlutě).



Obrázek 1: Mayerova metoda průměru

1.2 Boškovičova přímka

Roger Josef Boškovič (*13. 2. 1783) byl velmi nadaný matematik, fyzik a filozof, který se účastnil expedicí, při nichž bylo prováděno měření délky oblouku Země. Šlo o snahu potvrdit nebo vyvrátit hypotézu, že Země má tvar rotačního elipsoidu. Je to první vědec, který navrhuje způsob určení regresní přímky.

Boškovič zkoušel více metod na určení aproximující přímky $y = \alpha + \beta x$, ale jen jedna z nich se stala nejslavnější. Tato metoda se nazývá *Boškovičova metoda nejmenších absolutních odchylek*.

Metoda byla formulována následovně. Uvažujme určitý počet pozorování. Ke každému pozorování musí být určeno reziduum, které splňuje jisté vlastnosti.

1. Součet všech kladných reziduí musí být stejný jako součet všech záporných reziduí, tj.
$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0. \tag{2}$$
Odůvodnění vlastnosti je, že kladná i záporná rezidua (odchylky od skutečné hodnoty) jsou stejně možná.

2. Součet všech reziduí musí být co nejmenší. Matematicky chceme minimalizovat výraz
$$\sum_{i=1}^{n}|y_i - \alpha - \beta x_i|. \tag{3}$$

Z první podmínky plyne $\overline{y} = \alpha + \beta\overline{x}$, tzn. že přímka prochází těžištěm bodů. Při dosazení tohoto výrazu do (3) získáme

$$K(\beta) = \sum_{i=1}^{n} \mid (y_i - \overline{y} - \beta(x_i - \overline{x})) \mid . \tag{4}$$

Výraz je pak nutné minimalizovat vzhledem k parametru β.

Uvažujme stejná naměřená data jako u Mayera. Boškovič uspořádal pozorování postupně v pořadí 5, 1, 4, 2, 3 a dále upravil měření jako $X_i = x_i - \overline{x}$, $Y_i = y_i - \overline{y}$ a

$$\beta_i = \frac{y_i - \overline{y}}{x_i - \overline{x}} \tag{5}$$

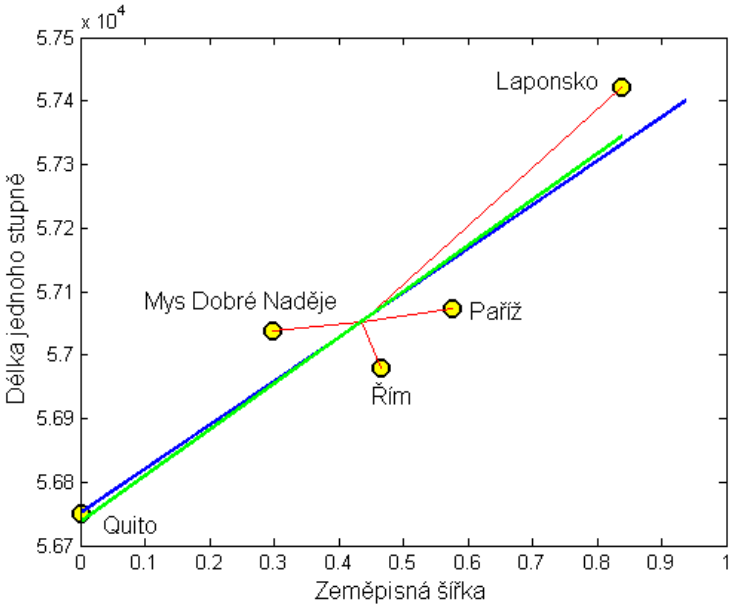
je směrnice přímky procházející těžištěm a daným bodem měření. Vznikly následující tabulky, kde k je označení toho, ve kterém bodě se nacházíme.

i	X_i	Y_i
1	0,40294	369,4
2	-0,43566	-301,6
3	0,14054	21,4
4	-0,13696	-15,6
5	0,02914	-73,6

k	β_k	$K(\beta_k)$
5	917	416
1	692	339
4	152	627
2	114	658
3	-2526	3527

Tabulky 2.3: Uspořádané hodnoty

Z druhé tabulky vidíme, že nejmenší hodnota $K(\beta_k)$ je u přímky, která prochází těžištěm a bodem 1. Vybírali jsme tedy přímku, která má nejmenší součet absolutních odchylek. Její rovnice je $y = 56751 + 692x$. Reziduální součet čtverců je pro metodu 29017 a absolutní součet reziduí $L = 339$. Pro připomenutí je hodnota RSČ pro MNČ 28630, což je blízko hodnotě pro Boškovičovu metodu. V porovnání s hodnotou $L = 350$ pro MNČ je hodnota z Boškovičovy přímky lepší. Na obrázku (2) vidíme zelenou přímku sestrojenou klasickou metodou nejmenších čtverců, modrá přímka je Boškovičova přímka a červené jsou všechny ostatní možné Boškovičovy přímky. Viz Příloha 1.



Obrázek 2: Boškovičova metoda nejmenších absolutních odchylek

1.3 Lambertova metoda

Johann Heinrich Lambert (*26. 8. 1728) byl německý matematik, astronom a fyzik, který, kromě jiného, vymyslel jednu z metod na aproximaci dat. Předpokládal, že vztahy jsou lineární, nebo jsou zlinearizované.

Jeho požadavkem bylo, aby regresní přímka $y = \alpha + \beta x$ procházela těžištěm všech pozorování $(\overline{x}, \overline{y})$. Tedy u našeho příkladu, aby procházela bodem $(\overline{x}, \overline{y}) = (0{,}43566; 57052{,}6)$. Dále rozdělil všechna pozorování na dvě přibližně stejné skupiny (měly stejný počet pozorování). Bylo to ale za předpokladu, že se v první skupině nacházely pozorování s menšími hodnotami x a ve druhé ty s většími. Následně početně určil těžiště pro skupiny ($\overline{x_1}$ a $\overline{y_1}$), $(\overline{x_2}, \overline{y_2})$. V našem případě vezměme do první skupiny 2 pozorování a do druhé skupiny 3. Naše skupiny mají těžiště o souřadnicích $(\overline{x_1}, \overline{y_1}) = (0{,}14935; 56894)$ a $(\overline{x_1}, \overline{y_2}) = (0{,}62653; 57158, 3)$.

Lambert následně vyjádřil

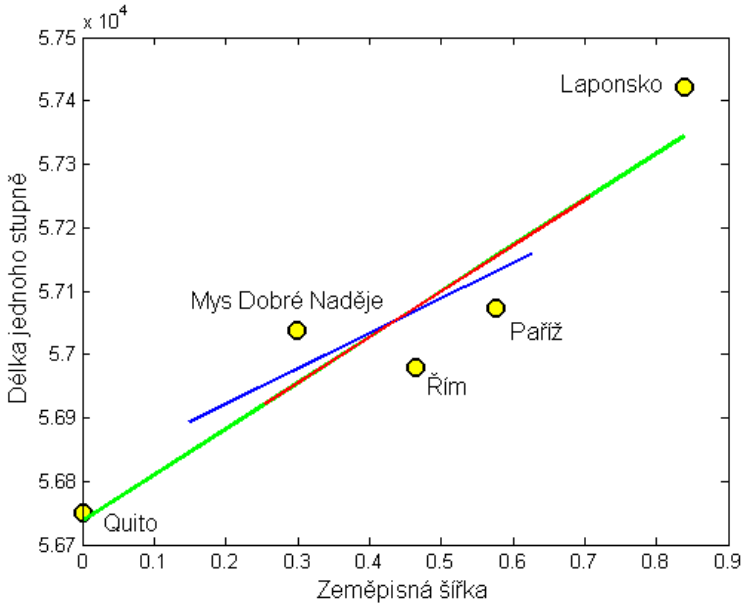
$$\beta = \frac{\overline{y_2} - \overline{y_1}}{\overline{x_2} - \overline{x_1}} = 553{,}88. \tag{6}$$

Aproximující přímka pak podle něj vypadá

$$y = \overline{y_1} - \beta\overline{x_1} + \beta x. \tag{7}$$

Po dosazení obdržíme rovnici regresní přímky $y = 56811{,}27 + 553{,}95x$ (modře).

Pokud bychom vzali do první skupiny 3 pozorování a do druhé skupiny 2, získali bychom přímku $y = 56739{,}33 + 719{,}07x$ (červeně). Na obrázku (3) jsou znázorněny obě přímky a zeleně je opět vyznačena regresní přímka získaná MNČ. RSČ pro první přímku je 39877 a $L = 413$. Pro druhou je RSČ 28637 a $L = 348$. Druhá červená Lambertova přímka je téměř totožná s přímkou získanou MNČ. RSČ se u ní od klasického RSČ u MNČ liší pouze o 7, což je asi o 0,024% a i L je velmi blízko hodnotě $L = 350$ pro MNČ. Program z Matlabu je zpracovaný v Příloze 2.



Obrázek 3: Lambertova metoda

1.4 Laplaceova metoda nejmenších absolutních odchylek (LAD)

Pierre Simon Laplace (*23. 3. 1749) ve své studii tvaru Země navázal na Boškoviče. Snažil se ale rovnice různě kombinovat. Dostal se tak blíže k obecnému řešení. Jeho metoda – *metoda nejmenších absolutních odchylek (Least Absolute Deviations)* minimalizuje funkci

$$K(\alpha, \beta) = \sum_{i=1}^{n} \mid y_i - \alpha - \beta x_i \mid . \tag{8}$$

Podle Laplace musí být absolutní součet reziduí co nejmenší. Hledáme opět odhady parametrů přímky $y = \alpha + \beta x$ tak, aby funkce (8) byla minimální. Předpokládáme, že mezi $x_1, \ldots x_n$ jsou alespoň dvě čísla různá.

Platí, že existuje optimální přímka, procházející nejméně dvěma body (x_i, y_i) a (x_j, y_j), kde $i \neq j$. Kdybychom ale měli zkoumat všechny možné kombinace, tak by nám to zabralo velké množství času a bylo by to složité. Proto se může přejít např. k metodě založené na lineárním programování.

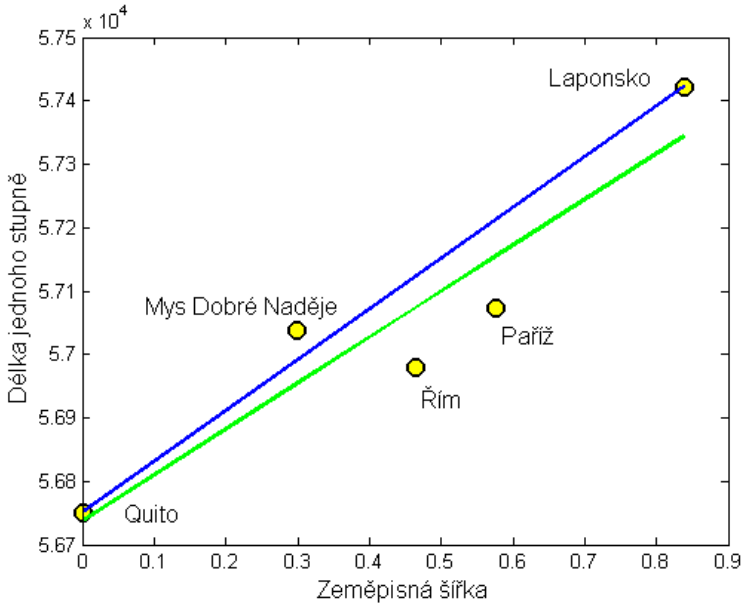
U lineárního programování platí, že minimalizace funkce $K(\alpha, \beta)$ se dá ekvivalentně přepsat do úlohy minimalizace $\sum_{i=1}^{n} r_i$ při splnění podmínek

$$r_i - \alpha + \beta x_i \geq y_i, \qquad \text{kde } i = 1, \ldots, n,$$

$$r_i - \alpha + \beta x_i \geq -y_i, \qquad \text{kde } i = 1, \ldots, n. \tag{9}$$

V programu Matlab jsme určili řešení úlohy lineárního programování: $r_1 = 0$, $r_2 = 47{,}51$, $r_3 = 143{,}55$, $r_4 = 137{,}79$, $r_5 = 0$. Hodnota zjištěných r_i nám určuje reziduum – délku úsečky vedené ve svislém směru od bodu (x_i, y_i) k optimální přímce. Protože r_1 a r_5 jsou rovny 0, optimální regresní přímka prochází body 1 a 5. Tedy má tvar $y = 56750{,}2 + 801{,}1x$. RSČ je u této přímky 42084 a $L = 330$. Dostáváme tedy nejmenší hodnotu absolutního součtu reziduí ze všech metod, což je zřejmé, protože metoda je na této minimalizaci založena.

Na Obrázku (4) vidíme Laplaceovu přímku (modře) a aproximující přímku získanou MNČ (zeleně).



Obrázek 4: Laplaceova metoda nejmenších absolutních odchylek

1.5 MNČ – metoda nejmenších čtverců (Least Squares Method)

Základy metody nejmenších čtverců byly poprvé publikovány v roce 1805 A. M. Legendrem v díle *Nouvelles méthodes pour la détermination des orbites des cométes*. Priorita v užití je ale přisuzována C. F. Gaussovi, který metodu objevil již dříve (před rokem 1799). Bohužel se mu zdála příliš jasná a zřejmá, a proto ji nepublikoval. Jde o matematicko – statistickou metodu, která se užívá při aproximaci dat. Cílem je určit odhady parametrů aproximační funkce, která by co nejlépe popisovala zjištěné hodnoty. MNČ poskytuje takové řešení, které splňuje podmínku, aby součet druhých mocnin chyb odhadnutého řešení byl co nejmenší. Poznatky o metodě jsem čerpala z [12, 15, 26].

Jestliže chyby $e \sim N(0, \sigma^2)$, potom MNČ má stejné výsledky jako metoda maximální věrohodnosti.

Pokud odhady parametrů $\boldsymbol{\beta}$ nejsou nejlepšími nestrannými odhady nebo např. náhodné chyby jsou závislé, či mají různý rozptyl, je třeba užít tzv. metodu zobecněných nejmenších čtverců – MZNČ. Speciálním případem je metoda vážených nejmenších čtverců. Odhady získané MZNČ jsou známy jako Aitkenovy odhady, viz [27].

MNČ se užívá hlavně při regresní analýze. Z výše zmíněných metod je MNČ tou nejlepší v souvislosti se zvoleným kritériem RSČ. Užití MNČ je také založeno na důležitém faktu. Máme totiž k dispozici nestranný odhad parametru σ^2 (rozptyl náhodné chyby), což u jiných metod nemáme. Někdy je ale vhodné použít jiné kritérium. Pak užíváme např. metodu nejmenších absolutních odchylek, váženou metodu nejmenších čtverců, Hubertovu metodu [26] atd.

Definice 1.1. *Náhodný vektor $\boldsymbol{\beta}$, který pro dané $Y_1, \dots Y_n$ minimalizuje výraz*

$$\Phi(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - g(x_j, \boldsymbol{\beta}))^2, \tag{10}$$

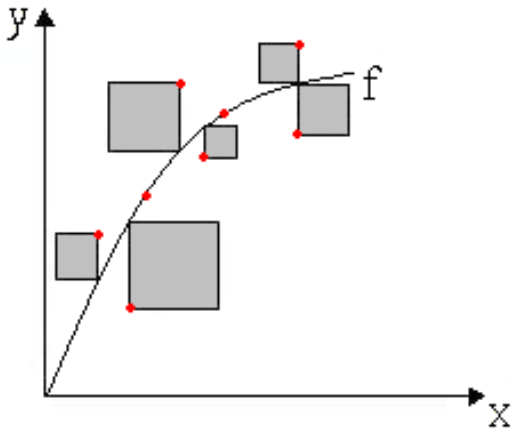
nazveme odhady parametrů určené metodou nejmenších čtverců.

Pokud je funkce lineární z hlediska parametrů, tak minimalizace vede k soustavě lineárních rovnic, kterých je stejný počet jako neznámých parametrů. Parciální derivace $\Phi(\boldsymbol{\beta})$ podle parametrů stačí položit rovny nule a dále řešit soustavu normálních rovnic. Získáme tak odhady parametrů.

Maticově můžeme odhad vektorového parametru $\boldsymbol{\beta}$ zapsat jako

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{11}$$

Na Obrázku (5) vidíme, jak metoda funguje.



Obrázek 5: Metoda nejmenších čtverců

2 Základní poznatky z regrese

Regrese se snaží vysvětlit kolísání náhodné veličiny Y, kterou nazýváme endogenní proměnná nebo vysvětlovaná proměnná, v závislosti na několika nenáhodných nezávisle proměnných (regresorech), které označujeme x.

Existuje jak jednoduchá regrese, kdy máme pouze jedinou nezávisle proměnnou, tak i mnohorozměrná regrese, kdy bychom měli místo skalární vysvětlované veličiny Y_i vektorovou veličinu $\mathbf{Y}_i$. Nejčastěji se užívá lineární model, který je také nejjednodušší. Ne vždy ale lze tento model použít. Často se totiž stává, že model lineární není. Těmito složitějšími modely se budeme zabývat v jiné kapitole.

Základní poznatky, které známe z klasického kurzu statistiky zopakujeme v této kapitole a budeme je aplikovat na konkrétní reálný příklad, který znázorňuje chemický pokus. Snažili jsme se názorně vždy uvést vzorce, podle kterých budeme počítat, a dále jsme vypsali i výsledky. Vzorce užité v kapitole pocházejí hlavně z literatury [5, 15]. Všechny výpočty byly provedeny v statistickém programu Matlab 7 (Příloha 3).

2.1 Zadání příkladu

Při zkoumání určitého chemického procesu bylo sledováno množství látky H (v gramech) po chemické reakci v závislosti na koncentraci daného roztoku k (v procentech). Určete parametry křivky popisující závislost hmotnosti vzniklé látky H na koncentraci k, máme-li k dispozici tato naměřená data:

koncentrace k	0	1	2	3	4	5	6	7	8	9	10
hmotnost H	1,1	4,8	13,2	25,1	41,5	61,4	85,1	112,7	144,5	181,3	221,5
	0,8	-	-	24,8	-	-	-	113,1	-	-	221,4
	1,2	-	-	24,8	-	-	-	112,9	-	-	220,1

Tabulka 3.1: Zadání chemického příkladu

2.2 Stochastický model

Úkol 1: Vytvořte příslušný stochastický model (chyby měření mají normální rozdělení s nulovou střední hodnotou a disperzí σ^2, jednotlivá měření jsou vzájemně nezávislá).

Pokud si naneseme hodnoty zadané v tabulce do grafu (např. v programu Maple nebo Matlab), tak zjistíme, že data mají přibližně kvadratický trend. Z toho budeme tedy při dalších výpočtech vycházet. Nyní musíme určit koeficienty funkce, která pro kvadratický trend vypadá následovně

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{12}$$

měříme – li v našem konkrétním případě hodnoty $y(x_i)$ pro $i = 0, \ldots, 10$.

V příkladu jde o model nepřímého měření vektorového parametru bez podmínek.
Obecný tvar stochastického modelu je následující

$$\mathbf{Y} = \mathbf{JA}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{13}$$

V našem případě $\mathbf{Y}$ je vektor typu 19×1, $\mathbf{JA} = \mathbf{X}$ je matice plánu typu 19×3, $\boldsymbol{\beta}$ je vektor parametrů typu 3×1 a $\boldsymbol{\epsilon}$ je vektor reziduí typu 19×1. A dále platí

$$\mathrm{h}(\mathbf{X}) = 3, \;\; \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_{19}), \;\; \mathrm{var}(\mathbf{Y}) = \Sigma \text{ p. d.}$$

Model můžeme tedy obecně zapsat

$$\begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \\ Y_{17} \\ Y_{18} \end{pmatrix} = \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_0 & x_0^2 \\ 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \\ 1 & x_6 & x_6^2 \\ 1 & x_7 & x_7^2 \\ 1 & x_7 & x_7^2 \\ 1 & x_7 & x_7^2 \\ 1 & x_8 & x_8^2 \\ 1 & x_9 & x_9^2 \\ 1 & x_{10} & x_{10}^2 \\ 1 & x_{10} & x_{10}^2 \\ 1 & x_{10} & x_{10}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \end{pmatrix}.$$

Pro naše měření matice vypadají následovně

$$\begin{pmatrix} 1{,}1 \\ 0{,}8 \\ 1{,}2 \\ 4{,}8 \\ 13{,}2 \\ 25{,}1 \\ 24{,}8 \\ 24{,}8 \\ 41{,}5 \\ 61{,}4 \\ 85{,}1 \\ 112{,}7 \\ 113{,}1 \\ 112{,}9 \\ 144{,}5 \\ 181{,}3 \\ 221{,}5 \\ 221{,}4 \\ 220{,}10 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \\ 1 & 9 & 81 \\ 1 & 10 & 100 \\ 1 & 10 & 100 \\ 1 & 10 & 100 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \end{pmatrix}.$$

2.3 Odhady parametrů

Úkol 2: Určete odhady parametrů 1. a 2. řádu a jejich varianční matice.

Pomocí programu Matlab jsme spočítali **odhady 1. řádu** – odhad parametru $\boldsymbol{\beta}$ ze vztahu

$$\widehat{\boldsymbol{\beta}}(\mathbf{Y}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{14}$$

V našem případě je V jednotková matice rozměrů 19×19. Po dosazení všech potřebných matic do vzorce (14) jsme získali výsledky pro vektorový parametr $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1{,}0096 \\ 2{,}0172 \\ 1{,}9978 \end{pmatrix}.$$

Pro určení **odhadu parametru 2. řádu** $\widehat{\sigma}^2$ musíme užít následující vzorec

$$\widehat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n - k + q}, \tag{15}$$

kde n je počet měření (tedy u nás 19), k je počet neznámých parametrů $\boldsymbol{\beta}$ (v našem případě máme 3 parametry) a q je počet podmínek (u nás 0). Po dosazení do vzorce (15) získáme

$$\widehat{\sigma}^2 = 0{,}1458.$$

Varianční matici odhadu parametrů 1. řádu $\widehat{\boldsymbol{\beta}}$ určíme ze vzorce

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{16}$$

Ze vzorečku (16) tak získáme varianční matici

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} 0{,}0374 & -0{,}0136 & 0{,}0010 \\ -0{,}0136 & 0{,}0086 & -0{,}0008 \\ 0{,}0010 & -0{,}0008 & 0{,}0001 \end{pmatrix}.$$

Dále budeme počítat **odhad varianční matice 2. řádu**

$$\widehat{\mathrm{var}}(\widehat{\sigma}^2) = \frac{2(\widehat{\sigma}^2)^2}{n - k}. \tag{17}$$

Matlab nám pak po užití vzorce (17) dá výsledek

$$\widehat{\text{var}}(\widehat{\sigma}^2) = 0{,}0027.$$

2.4 Oblasti spolehlivosti

Úkol 3: Určete oblasti spolehlivosti parametrů 1. a 2. řádu a sdružené intervaly spolehlivosti parametrů 1. řádu na hladině spolehlivosti $(1-\alpha) = 0{,}95$.

Pro určení **oblasti spolehlivosti 1. řádu** použijeme následující vztah pro k – rozměrný elipsoid, kde σ^2 je neznámé

$$C_{1-\alpha}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} : \boldsymbol{\beta} \in R^k, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \widehat{\sigma}^2 \; k \; F_{k,n-k}(1-\alpha)\right\}, \quad (18)$$

Po dosazení daných hodnot do vzorce (18) vypadá oblast spolehlivosti následovně

$$C_{0{,}95}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} : \boldsymbol{\beta} \in R^3, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq 0{,}1458 \cdot 3 \cdot F_{3,19-3}(0{,}95)\right\},$$

po upravení a dosazení za $F_{3,16} = 0{,}1458$ získáme

$$C_{0{,}95}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} : \boldsymbol{\beta} \in R^3, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq 1{,}4172\right\},$$

kde

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{X} = \begin{pmatrix} 19 & 95 & 701 \\ 95 & 701 & 5765 \\ 701 & 5765 & 50297 \end{pmatrix}.$$

Pro určení **oblasti spolehlivosti 2. řádu** použijeme vztah

$$C_{1-\alpha}(\sigma^2) = \left\langle \frac{\widehat{\sigma}^2(n-k)}{\chi^2_{n-k}(1-\frac{\alpha}{2})}, \frac{\widehat{\sigma}^2(n-k)}{\chi^2_{n-k}(\frac{\alpha}{2})} \right\rangle. \quad (19)$$

Po dosazení hodnot $\widehat{\sigma^2} = 0{,}1458$, $n = 19$, $k = 3$, $\chi^2_{16}(0{,}975) = 28{,}8$ a $\chi^2_{16}(0{,}025) = 6{,}91$ obdržíme oblast spolehlivosti pro σ^2

$$C_{0{,}95}(\sigma^2) = \langle 0{,}081; 0{,}3376 \rangle.$$

Nyní určíme **sdružené intervaly spolehlivosti parametrů 1. řádu**. Musíme užít Scheffého přístup, jelikož máme 19 pozorování. Pro jiný počet pozorování bychom mohli užít i tzv. Bonferoniho přístup. V tom případě ale musíme použít jiný vzorec pro výpočet než ten, který užijeme pro Sheffého přistup [15]

$$I_{1-\alpha}^{(j)}(\boldsymbol{\beta}_j) = \left\{ \boldsymbol{\beta}_j :\mid \widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_j \mid \leq \widehat{\sigma}\sqrt{kF_{k,n\text{-}k}(1-\alpha)}\sqrt{\mathbf{c_j}'(\mathbf{X}'\mathbf{V^{-1}X})^{-1}\mathbf{c_j}} \right\}, \quad (20)$$

$j = 1, 2, 3.$

V našem případě po dosazení do vzorce (20) za $k = 3$, $F_{3,17}(0{,}95) = 3{,}2$, $\widehat{\sigma} = \sqrt{0{,}1458}$,

$$\mathbf{c}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{c}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{c}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, (\mathbf{X}'\mathbf{V^{-1}X})^{-1} = \begin{pmatrix} 0{,}2569 & -0{,}0936 & 0{,}0071 \\ -0{,}0936 & 0{,}0590 & -0{,}0055 \\ 0{,}0071 & -0{,}0055 & 0{,}0005 \end{pmatrix}$$

získáme

$$I_{0{,}95}^{(j)}(\boldsymbol{\beta}_j) = \left\{ \boldsymbol{\beta}_j :\mid \widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_j \mid \leq \sqrt{0{,}1458}\sqrt{3 \cdot 3{,}2}\sqrt{\mathbf{c_j}'(\mathbf{X}'\mathbf{V^{-1}X})^{-1}\mathbf{c_j}} \right\}, \quad j = 1, 2, 3.$$

Tedy

$$I_{0{,}95}^{(J)}(\boldsymbol{\beta}_0) = \left\{ \boldsymbol{\beta}_0 :\mid \widehat{\boldsymbol{\beta}_0} - \boldsymbol{\beta}_0 \mid \leq 0{,}5997 \right\} = \langle 0{,}4099; 1{,}6093 \rangle ,$$

$$I_{0{,}95}^{(J)}(\boldsymbol{\beta}_1) = \left\{ \boldsymbol{\beta}_3 :\mid \widehat{\boldsymbol{\beta}_1} - \boldsymbol{\beta}_1 \mid \leq 0{,}2873 \right\} = \langle 1{,}7300; 2{,}3045 \rangle ,$$

$$I_{0{,}95}^{(J)}(\boldsymbol{\beta}_2) = \left\{ \boldsymbol{\beta}_2 :\mid \widehat{\boldsymbol{\beta}_2} - \boldsymbol{\beta}_2 \mid \leq 0{,}0276 \right\} = \langle 1{,}9702; 2{,}0255 \rangle .$$

2.5 Testování hypotézy

Úkol 4: Ověřte nulovou hypotézu H_0: „závislost hmotnosti na koncentraci je lineární“ vzhledem k alternativě H_a: „závislost hmotnosti na koncentraci není lineární“ na hladině $\alpha = 0{,}05$.

Maticově můžeme obecně hypotézu zapsat

$$H_0 : \mathbf{H}\boldsymbol{\beta} + \mathbf{h} = \mathbf{0}. \tag{21}$$

V našem konkrétním případě testujeme hypotézu H_0, kterou můžeme přepsat jako

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \quad \text{kde} \quad \beta_2 = 0$$

oproti alternativě

$$H_a : y = \beta_0 + \beta_1 x + \beta_2 x^2, \quad \text{kde} \quad \beta_2 \neq 0.$$

Testovací statistika v obecném případě vypadá

$$(\mathbf{H}\widehat{\boldsymbol{\beta}} + \mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{H}'](\mathbf{H}\widehat{\boldsymbol{\beta}} + \mathbf{h}). \tag{22}$$

Hypotézu H_0 zamítáme, jestliže tato testovací statistika je $\geq g\widehat{\sigma}^2 F_{h,n-k}(1-\alpha)$, kde g je hodnost matice H.

Naše hypotéza bude maticově vypadat

$$H_0 : (0,0,1) \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Tedy po dosazení: $H_0 : (0,0,1) \begin{pmatrix} 1{,}0096 \\ 2{,}0172 \\ 1{,}9978 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Výpočtem zjistíme, že testovací statistika má hodnotu 16480,3. Nyní dopočítáme pravou stranu. Ta je rovna $1 \cdot 0{,}1458 \cdot F_{1,19-3}(0{,}95) = 0{,}70942$. Hypotézu H_0 „závislost hmotnosti na koncentraci je lineární“ tedy **zamítáme**.

2. postup, který můžeme použít na testování hypotézy H_0: $\mathbf{H}\boldsymbol{\beta} + \mathbf{h} = \mathbf{0}$ je, že použijeme testovací statistiku

$$T = \frac{\mathbf{c}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\boldsymbol{\beta}}{\widehat{\sigma}\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}}. \tag{23}$$

Po dosazení hodnot $\mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, $\widehat{\beta_2} = 1{,}9978$, $\beta_2 = 0$, $\widehat{\sigma} = \sqrt{0{,}1458}$,

$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \begin{pmatrix} 0{,}2569 & -0{,}0936 & 0{,}0071 \\ -0{,}0936 & 0{,}0590 & -0{,}0055 \\ 0{,}0071 & -0{,}0055 & 0{,}0005 \end{pmatrix}$ do vzorce (23) nám testovací statistika vyjde $T = 9594{,}5$.

Testovací statistika má Studentovo rozdělení $t_{17}(0{,}975) = 2{,}11$. Hodnota testovací statistiky leží v kritickém oboru $W = (-\infty; -2{,}11) \bigcup (2{,}11; \infty)$, proto tedy hypotézu H_0 **zamítáme** (stejně jako v prvním postupu).

2.6 Silofunkce

Úkol 5: Určete silofunkci testu.

Pro výpočet silofunkce testu počítáme pravděpodobnosti zamítnutí H_0. Pokud H_0 není správná, znamená to, že platí

$$H_a : \mathbf{H}\boldsymbol{\beta} + \mathbf{h} = \xi \neq \mathbf{0}. \tag{24}$$

V našem případě $H_a : \beta_2 = \xi \neq 0$.

Síla testu je dána jako

$$P[F_{g,n-k}(\delta) \geq F_{g,n-k}(0, 1-\alpha)], \tag{25}$$

kde g je hodnost matice $\mathbf{H}$ a pro parametr necentrality platí vztah

$$\delta = (\xi'[\mathbf{H}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{H}']^{-1}\xi)\frac{1}{\widehat{\sigma}^2} \tag{26}$$

Některé výsledné hodnoty silofunkce ukazuje následující tabulka.

ξ	δ	síla testu
0	0	0,05
0,001	0,0126	0,0512
0,01	1,2580	0,3310
0,012	1,8116	0,5373
0,015	2,8306	0,8577
0,017	3,6357	0,9671
0,02	5,0322	0,9993
0,021	5,5480	0,9999
0,022	6,0889	1,0000

Tabulka 3.2: Hodnoty silofunkce

3 Potíže s invertabilitou

3.1 Multikolinearita

V regresi se většinou předpokládá, že regresní matice $\mathbf{X}$ má lineárně nezávislé sloupce. Multikolinearita není problém statistický, ale jde o problém datový. Přesná definice multikolinearity neexistuje, ale hovoříme o ní tehdy, pokud má tato matice lineárně závislé sloupce. Např. nechtěně užijeme stejný regresor dvakrát, špatně zvolíme kombinaci hodnot vysvětlujících proměnných atd. V tomto případě hovoříme o perfektní multikolinearitě. V praxi je ale mnohem častější případ tzv. téměř multikolinearity, kdy jsou sloupce matice $\mathbf{X}$ téměř závislé. Konkrétněji můžeme říci, že matice $\mathbf{X}'\mathbf{X}$ má determinant skoro nula. Z toho tedy plyne, že matici $\mathbf{X}'\mathbf{X}$ nelze, nebo jen velmi obtížně, invertovat. Nejde tedy prakticky určit klasický MNČ odhad, jelikož odhady jsou nestabilní, nebo odhadnuté parametry mají velmi vysoké rozptyly. My tak odhady vůbec nemůžeme použít, protože by to snížilo přesnost. Tento problém nazýváme špatně podmíněná matice. Kapitola byla zpracována pomocí literatury [1, 3].

3.1.1 Kritéria pro identifikaci multikolinearity

- *Determinant korelační matice* R – při silné lineární závislosti proměnných se determinant jen velmi málo liší od nuly. Malé hodnoty determinantu totiž způsobují velké rozptyly odhadů.

- *Nejmenší charakteristické číslo* λ – určíme λ_{min}, což je nejmenší vlastní číslo korelační matice R. Malé hodnoty vlastního čísla indikují silnou lineární závislost mezi proměnnými.

- *Index podmíněnosti* – index podmíněnosti vypočteme jako odmocninu z poměru největšího a nejmenšího vlastního čísla matice $\mathbf{X}'\mathbf{X}$. Čím větší je index, tím větší je závislost mezi proměnnými. Hodnoty indexu, které leží mezi 20 a 100 ukazují na existenci mírné multikolinearity. V tomto případě

můžeme použít metody na potlačení kolinearity. Pokud hodnoty přesahují 100, pak jde o velmi silnou multikolinearitu a ne všechny metody lze použít.

- *Korelační koeficienty* – hodnoty korelačních koeficientů mezi dvojicemi vysvětlujících proměnných, které jsou blízké 1 nebo -1, poukazují na možnost multikolinearity. Hůře se ale rozpoznává multikolinearita způsobená korelovaností mezi více regresory.

- *Hodnoty VIF$_j$* – jsou diagonální prvky matice R^{-1}. Čím vyšší jsou hodnoty VIF$_j$, tím je multikolinearita silnější.

- *Kritérium M* – Scottův test [23] – čím vyšší jsou hodnoty kritéria M, tím silnější je multikolinearita. Obecně se pro $M > 0,8$ předpokládá silná závislost. Kritérium vypočítáme pomocí vzorce

$$M = \frac{\frac{F}{\sum_{j=1}^{k} t_j^2} - 1}{\frac{F}{\sum_{j=1}^{k} t_j^2} + 1}, \tag{27}$$

  kde t_j jsou testová kritéria pro individuální t – testy, F je testové kritérium pro celkový F – test.

3.1.2 Postupy pro modely s multikolinearitou

1. *Ignorování multikolinearity* – v některých případech může být model vhodný i při zjištění multikolinearity. Ta totiž neovlivní některé vlastnosti MNČ odhadů ani předpovědi. Někdy tedy skutečně můžeme multikolinearitu zanedbat.

2. *Vynechání vysvětlujících proměnných, které způsobují multikolinearitu* – postup může ale velmi narušit interpretaci modelu, což může být někdy zásadní problém. Lze jej ale užít v případě identifikace zbytečných vysvětlujících proměnných. Ke správné identifikaci nám pomohou metody, které hledají nejlepší podmnožinu vysvětlujících proměnných, regresní grafy atd.

3. *Transformace některých vysvětlujících proměnných* – jedná se o různé úpravy proměnných, pomocí kterých můžeme multikolinearitu omezit – normování, centrování odečtením průměru, nahrazení dvojice silně korelovaných regresorů jejich poměrem.

4. *Rozšíření datového souboru* – použijeme větší soubor dat (pokud je to tedy možné – např. zvýšíme frekvenci pozorování). Můžeme si také pořídit úplně nová data, což ale nezaručí, že multikolinearitu odstraníme.

5. *Použití apriorní fce* – někdy zjistíme dodatečné informace o modelu, což může působit proti multikolinearitě. Určení maximálního množství všech empirických a věcných informací o modelu a parametrech většinou vede ke zvýšení kvality modelu i ke zlepšení vlastností regresních odhadů.

6. *Použití metody hlavních komponent* – nejobjektivnější způsob jak multikolinearitě čelit. Tato metoda umožní přejít k malému počtu lineárních kombinací původních regresorů, které jsou navzájem nezávislé.

3.2 Hřebenová regrese (Ridge regression)

Další možností řešení regresních modelů vykazujících multikolinearitu je metoda hřebenové regrese. Byla navržena A. Hoerlem v roce 1962. Při užití hřebenové regrese se zlepší podmíněnost matice $\mathbf{X}'\mathbf{X}$. Princip hřebenové regrese byl zpracován pomocí [7, 8].

Metoda hřebenové regrese je založena na nahrazení klasické minimalizační úlohy

$$\Phi(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{28}$$

následující minimalizační úlohou

$$\Phi(\boldsymbol{\beta}, \delta) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \delta(\boldsymbol{\beta}'\boldsymbol{\beta} - c), \qquad \delta,\ c \in R^{+}. \tag{29}$$

Minimum (29) dostaneme standardním postupem

$$\frac{\partial \Phi}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\delta\boldsymbol{\beta} = 0, \tag{30}$$

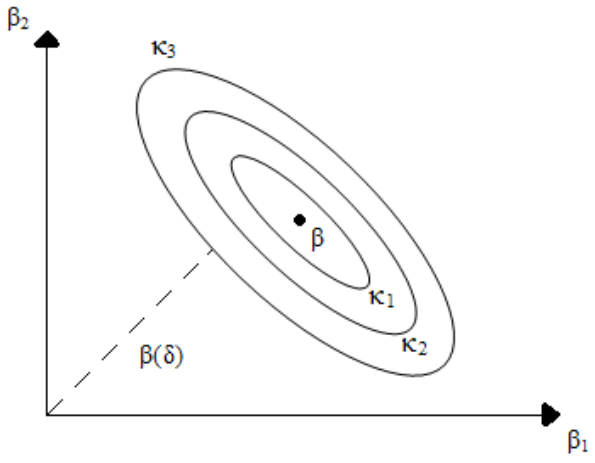
odtud

$$(\mathbf{X}'\mathbf{X} + \delta\mathbf{I})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y},$$

$$\widehat{\boldsymbol{\beta}}(\delta) = (\mathbf{X}'\mathbf{X} + \delta\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}. \tag{31}$$

Dá se ukázat, že přidání parametru δ do diagonály matice $\mathbf{X}'\mathbf{X}$ má vliv na zlepšení její podmíněnosti.

Označme $\kappa = \{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = konst\}$, tato množina je elipsa se středem v bodě $\boldsymbol{\beta}$. Pro různá řešení $\boldsymbol{\beta}$ splňující podmínku $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = konst$, má hřebenový odhad nejmenší vzdálenost $\boldsymbol{\beta}'(\delta)\boldsymbol{\beta}(\delta)$. Elipsy a hřebenový odhad vidíme na Obrázku (6).



Obrázek 6: Hřebenový odhad

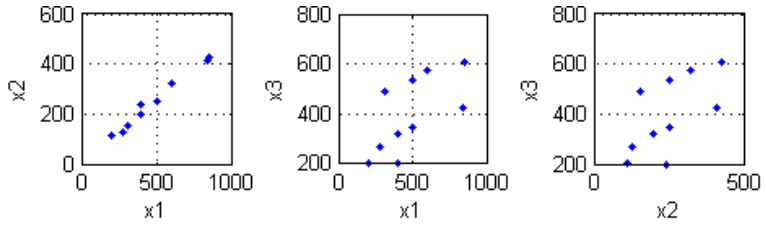
Otázkou je, jaký existuje vztah mezi $\widehat{\boldsymbol{\beta}}$ a $\widehat{\boldsymbol{\beta}}(\delta)$. Odvodíme si, že

$$\widehat{\boldsymbol{\beta}}(\delta) = (\mathbf{X}'\mathbf{X} + \delta\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X} + \delta\mathbf{I})^{-1}[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] =$$

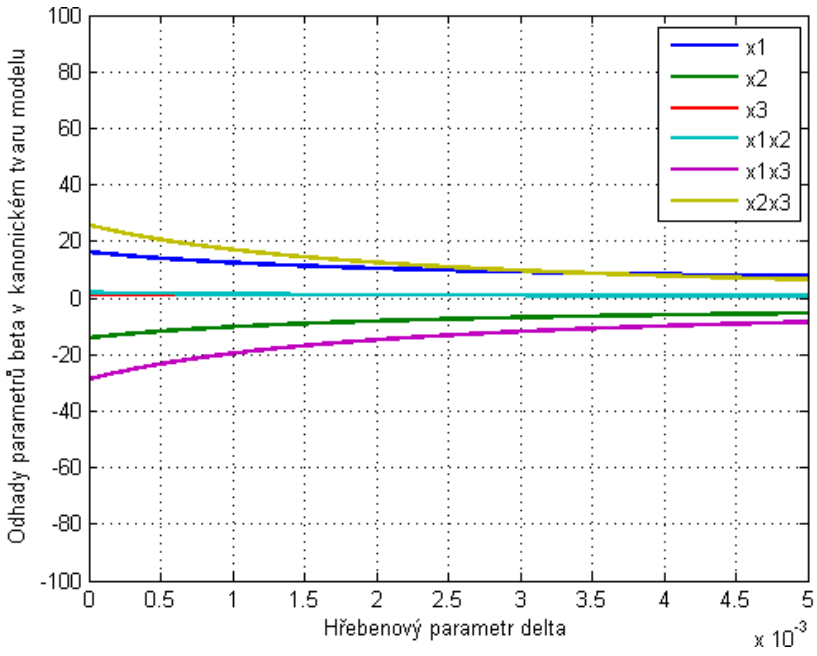
$$= (\mathbf{X}'\mathbf{X} + \delta\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I} + \delta(\mathbf{X}'\mathbf{X})^{-1})^{-1}\widehat{\boldsymbol{\beta}} = \mathbf{Z}\widehat{\boldsymbol{\beta}}. \tag{32}$$

Při užívání hřebenové regrese často sledujeme závislost mezi odhady parametrů v kanonickém tvaru $\mathbf{c} = \mathbf{V}'\boldsymbol{\beta}$ a hřebenovým parametrem δ, přičemž $\mathbf{V}$ je matice charakteristických vektorů $\mathbf{X}'\mathbf{X}$ a $\boldsymbol{\beta}$ je klasický odhad parametrů z původního modelu. Tuto závislost nazýváme hřebenová stopa – v anglicky psané literatuře se setkáme s výrazem ridge trace viz Obrázek (8).

Příklad 3.1. *Na Obrázku* (7) *vidíme, že data* x_1 *a* x_2 *vykazují mezi sebou lineární závislost. Vztahy mezi různými* δ *a odhady parametrů jsou znázorněny na Obrázku* (8).



Obrázek 7: Data



Obrázek 8: Hřebenová stopa

3.2.1 Zobecněná hřebenová regrese

Nechceme-li navyšovat všechny prvky na diagonále matice $\mathbf{X}'\mathbf{X}$ o stejnou konstantu δ, užijeme zobecněnou hřebenovou regresi. Metoda se liší od klasické hřebenové regrese pouze tím, že diagonální prvky navyšujeme o kladné konstanty $\delta_1, \ldots, \delta_k$.

Definice 3.1. *Střední čtvercovou chybou nazveme matici*

$$MSE(\widehat{\boldsymbol{\beta}}) = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \tag{33}$$

pro nevychýlené odhady a

$$MSE(\widehat{\boldsymbol{\beta}}(\delta)) = var(\widehat{\boldsymbol{\beta}}(\delta) + [E(\widehat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta})][E(\widehat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta})]' \tag{34}$$

pro vychýlené odhady.

Definice 3.2. *Čtvercovou ztrátou nazveme*

$$L(\widehat{\boldsymbol{\beta}}) = E[(\widehat{\boldsymbol{\beta}} - E\widehat{\boldsymbol{\beta}})'(\widehat{\boldsymbol{\beta}} - E\widehat{\boldsymbol{\beta}})] \tag{35}$$

pro nevychýlené odhady a

$$L(\widehat{\boldsymbol{\beta}}(\delta)) = E[(\widehat{\boldsymbol{\beta}} - E\widehat{\boldsymbol{\beta}})'(\widehat{\boldsymbol{\beta}} - E\widehat{\boldsymbol{\beta}})] + [E(\widehat{\boldsymbol{\beta}}(\delta)) - E\widehat{\boldsymbol{\beta}}]'[E(\widehat{\boldsymbol{\beta}}(\delta)) - E\widehat{\boldsymbol{\beta}}] \tag{36}$$

pro vychýlené odhady. Rozdíl $E(\widehat{\boldsymbol{\beta}}(\delta)) - E\widehat{\boldsymbol{\beta}}$ nazýváme vychýlení odhadu.

Věta 3.1. *Nechť $\boldsymbol{\beta}$ je MNČ odhad a $\widehat{\boldsymbol{\beta}}(\delta)$ hřebenový odhad, pak matice*

$$var(\widehat{\boldsymbol{\beta}}) - MSE(\widehat{\boldsymbol{\beta}}(\delta)) = -(E(\widehat{\boldsymbol{\beta}}(\delta)) - E\widehat{\boldsymbol{\beta}})(E(\widehat{\boldsymbol{\beta}}(\delta)) - E\widehat{\boldsymbol{\beta}})' \tag{37}$$

je pozitivně definitní pro $0 < \delta < \frac{2\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}$.

Důkaz: Viz [9].

Výše zmíněná věta je důležitá, protože ukazuje, že hřebenový odhad má menší čtvercovou chybu než klasický odhad nestranného parametru $\boldsymbol{\beta}$. Říká, že existuje odhad ve smyslu střední čtvercové chyby oproti MNČ odhadu.

3.2.2 Odhad parametru δ

Existuje celá řada metod pro výpočet konstanty δ. Každá úloha má své optimální δ. Vzorce můžeme nalézt v [8]. Nyní uvedu alespoň některé z nich

1. $$\delta = \frac{kS_e^2}{\sum_{j=1}^{p} \beta_j^2}, \tag{38}$$

   přičemž β_j jsou odhady z původního modelu získané MNČ, $S_e^2 = \frac{\text{RSČ}}{n-k-1}$, kde n je počet pozorování, k je počet parametrů, RSČ je reziduální součet čtverců z původního modelu,

2. optimální δ vypočteme z rovnice

   $$S_e^2 \sum_{j=1}^{k} \frac{\lambda_j}{\lambda_j + \delta} = \delta \sum_{j=1}^{k} \frac{\lambda_j^2 c_j^2}{(\lambda_j + \delta)^2}, \tag{39}$$

   λ_j jsou vlastní čísla matice $\mathbf{X}'\mathbf{X}$, $\mathbf{c} = \mathbf{V}'\boldsymbol{\beta}$, kde $\mathbf{V}$ je již dříve zmiňovaná matice charakteristických vektorů $\mathbf{X}'\mathbf{X}$,

3. modifikací odhadu (38) bychom získali

   $$\delta = \frac{kS_e^2}{\sum_{j=1}^{p} \lambda_j c_j^2}, \tag{40}$$

4. $$\delta = \sigma^2 \frac{\sum_{j=1}^{k} (\frac{1}{\lambda_j^2})}{\sum_{j=1}^{k} \frac{(3\sigma^2 + c_j^2 \lambda_j)}{\lambda_j^3}}. \tag{41}$$

3.3 Příklad

Příklad 3.2. *Mějme data, která obsahují pozorování chemického složení portlandského cementu a množství tepla Y, které je vydáno při tuhnutí betonu v kaloriích na gram cementu. Vysvětlující proměnné X_1, X_2, X_3 a X_4 jsou 4 různé složky cementu (údaje v tabulce jsou v procentech). Data jsou známá jako tzv. Haldova*

data, publikovaná už v roce 1932. *Program na výpočet příkladu v Matlabu nalezneme v Příloze 4.*

teplota	*složka 1*	*složka 2*	*složka 3*	*složka 4*
78,5	7	26	6	60
74,3	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,3	11	66	9	12
109,4	10	68	8	12

Tabulka 4.1: Haldova data

Model je ve tvaru

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon, \ i = 1, 2, \ldots, n. \tag{42}$$

Výpočtem indexu podmíněnosti $\kappa = 6056{,}344$ *zjistíme, že se v datech vyskytuje velmi silná multikolinearita. I pomocí korelačních koeficientů určíme, že jsou silně kolineární složky* 1 *a* 3 *a také* 2 *a* 4, *protože* $r_{x_1x_3} = -0{,}824$ *a* $r_{x_2x_4} = -0{,}973$.

Nejprve vypočteme odhady parametrů klasickou MNČ podle vzorce (14) *následovně*

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 62{,}4054 \\ 1{,}5511 \\ 0{,}5102 \\ 0{,}1019 \\ -0{,}1441 \end{pmatrix}, \qquad \widehat{\sigma}^2 = 6{,}83766276.$$

Vypočteme hřebenový odhad parametrů podle vzorce (31), *kde za* δ *náhodně zvolíme* 5

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}0616 \\ 2{,}1261 \\ 1{,}1680 \\ 0{,}7104 \\ 0{,}4957 \end{pmatrix}, \qquad \widehat{\sigma_\delta}^2 = 7{,}62268996.$$

Nyní zkusíme vypočítat optimální parametr δ, pomocí kterého bychom měli dosáhnout menšího $\widehat{\sigma}^2$. Optimální δ vypočteme např. podle vzorců (38) a (40). Nejprve určíme optimální δ podle (38). Získáme tak $\delta_{opt1} = 0{,}00614$ a po použití δ_{opt1} dostaneme nové odhady parametrů jako

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 10{,}3686 \\ 2{,}0863 \\ 1{,}0465 \\ 0{,}6494 \\ 0{,}3816 \end{pmatrix}, \qquad \widehat{\sigma_{opt1}}^2 = 6{,}39540385.$$

Dále získáme δ_{opt} podle jiného vzorce (40). Dostaneme $\delta_{opt2} = 0{,}000198$ a po použití nového δ_{opt2} máme odhady parametrů

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 53{,}6966 \\ 1{,}6407 \\ 0{,}5999 \\ 0{,}1935 \\ -0{,}0561 \end{pmatrix}, \qquad \widehat{\sigma_{opt2}}^2 = 5{,}99450698.$$

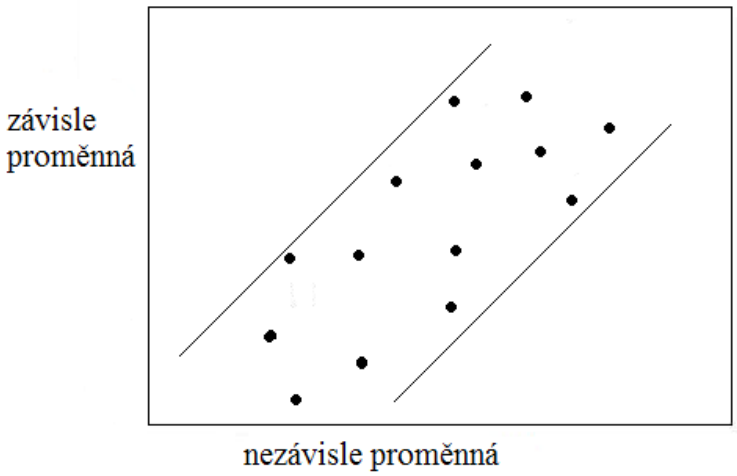
Vidíme, že u obou případů při optimálním δ máme menší rozptyl než u odhadů získaných MNČ. Při použití δ_{opt2} získáváme lepší odhady, protože $\widehat{\sigma}^2$ je menší než při použití δ_{opt1}.

Takto bychom mohli použít i ostatní vzorce pro výpočet optimálního δ a určili bychom, který odhad má nejmenší rozptyl. Pak bychom nadále pracovali s tímto nejlepším δ_{opt}.

4 Nesplnění homoskedasticity

4.1 Klasický lineární model

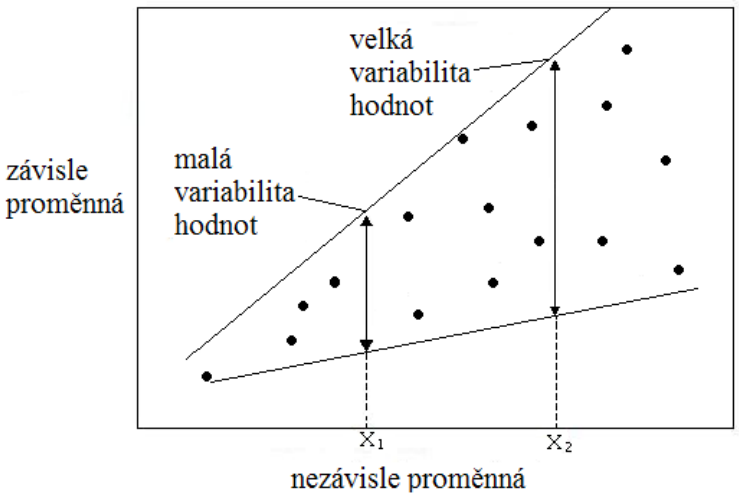
V námi známém klasickém lineárním modelu předpokládáme, že rozptyly $var(\varepsilon_i)$ náhodných veličin (reziduálních složek) $\varepsilon_1,\ \varepsilon_2, ..., \varepsilon_n$ jsou shodné a všechny se rovnají kladné neznámé konstantě σ^2, viz Obrázek (9). Model tohoto typu označujeme jako homoskedastický.



Obrázek 9: Homoskedasticita

4.2 Heteroskedasticita

Problém nastává u modelu, který nesplňuje podmínku, že všechny rozptyly jsou rovny konstantě σ^2. Tedy jinak řečeno, že náhodnost obsažená ve výstupech je pro každé pozorování jiná. Toto pak nazýváme heteroskedasticita. Příslušným modelům i vysvětlovaným proměnným se říká heteroskedastické – Obrázek (10). Část práce týkající se heteroskedasticity byla zpracována hlavně pomocí literatury [25]. Testy heteroskedasticity jsou uváděny i v dále citovaných pramenech.



Obrázek 10: Heteroskedasticita

Mezi nejčastější důvody nesplnění podmínky homoskedasticity patří:

1. *chybná specifikace modelu* – např. vynecháme některý podstatný regresor,

2. *údaje jsou agregované (seskupené)* – např. získaná data (původně homoskedastická) zprůměrujeme, a tím se nám může stát, že vzniknou data heteroskedastická,

3. *model má náhodné parametry* – náhodné veličiny jsou nekorelované s nulovými středními hodnotami a rozptyly $var(x)$, pak rozptyly $var(\varepsilon_i)$ nesplňují podmínku homoskedasticity,

4. *ekonomická data* – makro i mikro ekonomická data jsou i v jednom výběru pozorování velmi rozličná,

5. *rozptyly Y jsou skedastickou funkcí vysvětlujících proměnných* – např. když sledujeme příjmy a výdaje v náhodném výběru domácností. Je jasné, že s růstem příjmu domácností rostou průměrné výdaje ale i variabilita těchto výdajů. Tzn., že řádky matice X nemají stejné střední hodnoty ani rozptyly.

V další části oddílu zjednodušeně předpokládejme, že neexistují korelace mezi náhodnými veličinami $\varepsilon_1,\ \varepsilon_2, ..., \varepsilon_n$, ani mezi veličinami $y_1,\ y_2, ..., y_n$.

Minimalizace z hlediska parametru $\boldsymbol{\beta}$ reziduálního součtu čtverců je ve tvaru

$$Q(\varepsilon_Z) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}(\frac{\varepsilon_{\mathrm{i}}}{\sigma_{\mathrm{i}}}), \tag{43}$$

kde Ω je kovarianční matice v následujícím tvaru, přičemž w_i jsou váhy

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix} = \boldsymbol{\sigma}^2 \begin{pmatrix} w_1 & 0 & ... & 0 \\ 0 & w_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & w_n \end{pmatrix} = \boldsymbol{\sigma}^2\mathbf{W}. \tag{44}$$

Je dobré zavést podmínku $st(\mathbf{W}) = \sum_{i=1}^{n} w_i = n$. Je jedno, zda pracujeme s maticí $\boldsymbol{\Omega}$ nebo $\mathbf{W}$, protože konstanta $\boldsymbol{\sigma}^2$ na odhadu $\boldsymbol{\beta}$ nic nezmění. Při znalosti $\boldsymbol{\Omega}$, popřípadě $\mathbf{W}$ je jednoduché získat kvalitní odhady parametrů. Minimalizace vede k zobecněnému odhadu $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}. \tag{45}$$

Většinou ale matici $\boldsymbol{\Omega}$ ani matici $\mathbf{W}$ neznáme, a proto pracujeme s maticemi $\widehat{\boldsymbol{\Omega}}$, $\widehat{\mathbf{W}}$ viz [1]. Výše zmíněný postup nazýváme vážená metoda nejmenších čtverců.

4.2.1 Testování heteroskedasticity

Otázkou je, jak ověřit předpoklad stejných nebo nestejných rozptylů. Může se stát, že heteroskedasticitu poznáme z grafu, do kterého zaneseme zkoumané hodnoty. To se ale v praxi příliš nestává. Většinou tedy musíme užít testy na testování heteroskedasticity. Snažíme se provést rozhodnutí, zda existuje velké porušení homogenity rozptylů, potom lze odchylky od homoskedasticity označit za statisticky významné. Těmto testům říkáme *nekonstruktivní (nonconstructive)*. Jinými testy jsou testy *konstruktivní (constructive)*, kde je testování spojeno s odhadem parametrů v modelu.

Goldfeld – Quandtův test (parametrický)

Parametrický test předpokládá, že dokážeme rozpoznat příčinu heteroskedasticity. Konkrétně je možné uspořádat pozorování podle velikosti regresoru, který ovlivňuje směrodatnou odchylku. Spadá mezi nekonstruktivní testy a používá se na soubor, který má menší počet pozorování nebo nedostatečný počet vysvětlujících proměnných. Test vychází z předpokladu, že rozptyly σ_i^2 jsou monotónní funkce některé vysvětlující proměnné X_j. Můžeme ho najít v literatuře [19]. Existuje i neparametrický Goldfeld – Quandtův test, ale tím se v naší práci zabývat nebudeme.

Postupujeme následovně:

- testujeme hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$,
- předpokládejme, že chyby mají normální n – rozměrné rozdělení,
- pozorování setřídíme dle rostoucího rozptylu $\sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_n^2$,
- zvolíme $r \doteq \frac{n}{4}$ tak, aby $n-r$ bylo sudé a $\frac{n-r}{2} > p$, kde p je počet parametrů,
- vynecháme r prostředních pozorování, zůstane nám tedy $\frac{n-2}{2}$ pozorování na začátku a $\frac{n-2}{2}$ pozorování na konci,
- pro horní i dolní části dat odhadneme regresní funkce a spočteme reziduální součty čtverců – RSČ1 a RSČ2,
- při normálním rozdělení náhodné složky a homoskedasticitě platí, že testovací statistika je

  $$F = \frac{\text{RSČ2}}{\text{RSČ1}} \sim F_{\frac{n-r}{2},\frac{n-r}{2}},$$

  v případě $F \geq F_{1-\alpha}(\frac{n-r}{2}, \frac{n-r}{2})$ hypotézu H_0 zamítneme (tedy vysoké hodnoty F indikují, že proměnná X_j způsobuje heteroskedasticitu).

V Příloze 5 nalezneme program vytvořený v Matlabu, který testuje data pomocí Goldfeld – Quandtova testu. Program určí, zda jsou vložená data homoskedastická, nebo heteroskedastická.

Whiteův test

Jedním z nejpoužívanějších testů v ekonomii je právě Whiteův test. Ten nám ale vůbec nenaznačí, co dělat v případě zamítnutí homoskedasticity. Je obecnější a nepředpokládá konkrétní závislost. Opět spadá mezi nekonstruktivní testy. Můžeme ho užít i pro detekci odlehlých pozorování. Whiteův test nalezneme v literatuře [3].

Postup:

- opět testujeme hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$,
- mějme např. model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{46}$$

- pro Whiteův test musíme vytvořit pomocný model, který je ve tvaru

$$\widehat{\varepsilon_i} = \alpha_0 + \textit{proměnné + mocniny proměnných + součiny proměnných} + v_i, \tag{47}$$

  kde v_i jsou normálně rozdělená rezidua.
  Pro náš konkrétní příklad by pomocný model vypadal

$$\widehat{\varepsilon_i} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 {x_{i1}}^2 + \alpha_4 x_{i2}^2 + \alpha_5 x_{i1} x_{i2} + v_i, \qquad i = 1, \ldots, n. \tag{48}$$

  Postup volíme proto, protože chceme zjistit, zda se rozptyl původních chyb (levá strana v modelu (47)), mění v závislosti na všech regresorech modelu (46),

- v pomocném modelu (47) provedeme souhrnný F – test lineárních omezení. Testujeme $H_0 : \alpha_k = 0$, kde $k = 0, \ldots, j$, neboť za platnosti homoskedasticity by se tato hypotéza neměla zamítnout.
  Testovací statistika je:

$$F = \frac{n-j}{j-1} \frac{\text{RSČ1} - \text{RSČ2}}{\text{RSČ1}},$$

kde RSČ1 je rezidualní součet čtverců modelu (47) omezeného na pouhý intercept a RSČ2 je reziduální součet čtverců získaný z pomocného modelu (47).
H_0 zamítáme, jestliže $F \geq F_{1-\alpha}(j-1, n-j)$,

- alternativně můžeme užít i χ^2 – test. Musíme jen najít koeficient determinace R^2 v modelu (47). Kritický obor $H_0 : \alpha_k = 0$ na hladině významnosti α je

$$F = (n-j)R^2 \geq \chi^2_{1-\alpha}(j-1).$$

Spearmanův test

Jednoduchý test, který se opět řadí mezi nekonstruktivní testy. Nemáme zde ale předpoklad na rozdělení chyb. Předpokládá se, že $var(y)$ závisí na X_j (tedy na j – tém sloupci matice X). Najdeme ho v literatuře [25]. Je vhodný jak pro velké, tak i malé výběry. C. Spearman ho poprvé uveřejnil ve svém článku *The Proof and Measurement of Association between Two Things.*

Postup:

- mějme opět nulovou hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$,
- spočteme absolutní hodnoty reziduí $|\varepsilon_i|$, vzestupně je seřadíme a očíslujeme (určíme tak pořadí – i_ε),
- příslušné pořadí přiřadíme k původním neseřazeným reziduím,
- spočteme absolutní hodnoty pozorování $|X_i|$, vzestupně je seřadíme a očíslujeme (určíme tak pořadí – i_x),
- příslušné pořadí opět přiřadíme k původním neseřazeným pozorováním,
- spočteme Spearmanův koeficient korelace pořadí mezi reziduem ε a danou vysvětlovanou proměnnou X

$$r_s = 1 - \frac{6\sum(i_x - i_\varepsilon)^2}{n(n^2-1)},$$

kde n je počet pozorování. Pokud se r_s pohybuje okolo 1, pak očekáváme homoskedasticitu. Pokud se pohybuje okolo 0, pak dostáváme heteroskedasticitu. Musí ale platit, že $r_s \in \langle -1, 1 \rangle$. Když zjistíme, že r_s se v tomto intervalu nevyskytuje, tak je nejspíše někde chyba.

- H_0 zamítáme, jestliže

$$T = \sqrt{r_s \frac{n-2}{1-r_s^2}} \geq t_{1-\alpha}(n-2).$$

Konstruktivní testy předpokládají určitý typ heteroskedastického modelu, tedy že už jsme dříve zjistili, že model je heteroskedastický. Obecně jde o snahu redukovat počet neznámých parametrů. Oblíbený heteroskedastický model předpokládá, že neznámé rozptyly $var(\varepsilon_i) = \sigma_i^2$ pro $i = 1,\ 2, \ldots, n$, jsou lineární funkcí $f(\mathbf{z_i'}, \boldsymbol{\alpha})$ podmnožiny L vysvětlujících proměnných $Z_1,\ Z_2, ..., Z_L$, kde $\mathbf{z_i'}$ je transponovaný i – tý řádek matice $\mathbf{Z}$. $\mathbf{Z}$ je většinou proti $\mathbf{X}$ zmenšené, protože ne všechny vysvětlující proměnné způsobují heteroskedasticitu. A $\boldsymbol{\alpha}$ je vektor neznámých heteroskedastických parametrů. Heteroskedastický model má tvar

$$\sigma_i^2 = f(\mathbf{z_i'}, \boldsymbol{\alpha}) + \textit{chyba modelu.} \tag{49}$$

Glejserův test

Patří mezi nejstarší testy. Je založen na odhadu heteroskedastických parametrů lineární rovnice

$$\mid \varepsilon_i \mid = \alpha_0 + \sum_{j=1}^{L} \alpha_j z_{ij} \quad \text{pro } i = 1, \ldots, n,$$

kde ε_i jsou běžná rezidua. Hypotézu $H_0 : \alpha_0 = \alpha_1 = ... = \alpha_L$ ověřujeme obvyklým postupem – test kvality modelu. Její zamítnutí se považuje za důkaz heteroskedastického modelu. Dále můžeme pomocí t – testů posuzovat, jestli se skutečně všechny uvažované proměnné $Z_1,\ Z_2, ..., Z_L$ podílejí na heteroskedasticitě. Test je popsán v literatuře [1].

4.2.2 Důsledky heteroskedasticity

Důsledky heteroskedasticity rozumíme důsledky projevující se při ignorování heteroskedastického modelu a při současném užití klasického MNČ odhadu. Platí:

- odhad parametrů $\boldsymbol{\beta}$ zůstane nestranným a konzistentním odhadem,
- nebude ale obecně nejlepším odhadem (nebude mít nejmenší rozptyl mezi všemi rozptyly ostatních odhadů),
- odhad parametru σ^2 nezůstane obecně nestranným,
- nemůžeme užít standardní vzorce, protože to může vést k chybným výsledkům.

4.2.3 Řešení heteroskedasticity

Pokud zjistíme heteroskedasticitu, tak je řešení jednoduché, ale jen pouze pokud známe její příčiny. V praxi většinou příčiny heteroskedasticity neznáme. Existují procedury, které je dokáží zjistit, ale ty jsou často příliš složité. Jednou z dalších možností je např. aplikace logaritmické či jiné transformace tak, aby došlo k redukci extremálních hodnot, které mohou heteroskedasticitu způsobit.

5 Odlehlá pozorování

5.1 Co to jsou odlehlá pozorování a jak je identifikovat

Mezi základní diagnostické nástroje při hodnocení kvality regresní funkce a dat patří analýza reziduí. Obecně jde o nástroj pro posuzování platnosti předpokladů zvoleného modelu. Můžeme říci, že jakákoliv nenáhodnost zjištěná u reziduí poukazuje na nějaký nedostatek odhadnutého regresního modelu (např. nevhodně jsme zvolili typ regresní funkce, použili nenáhodný výběr nebo se nám vyskytly extrémní či vybočující pozorování – odlehlá pozorování atd). Kapitola byla zpracována pomocí [1].

V praxi se někdy objeví pozorování, která vybočují z řady. Odlehlá pozorování můžeme rozlišovat na extrémní (*extremes* nebo také *hight leverages*) a vybočující (*outliers*). Vybočující pozorování jsou takové nízké či vysoké hodnoty y, zásadně se odlišující od ostatních hodnot vysvětlované proměnné Y. Extrémní pozorování jsou body $\mathbf{x_i'}$ (což je vlastně řádek matice $\mathbf{X}$), které se značně odlišují od ostatních. Jsou to body hodně vzdálené od centroidu (což je vektor průměru všech bodů).

U odhadů, získaných metodou nejmenších čtverců, je velký problém s jejich citlivostí na odlehlá pozorování. Pokud totiž v některých případech vypustíme odlehlé body, projeví se to zásadní změnou v odhadnutých parametrech. Musíme tedy správně posoudit vliv jejich případného vyloučení ze souboru na regresní charakteristiky.

Platí, že čím vzdálenější je bod x_i od průměru $\overline{x}$ (centroidu), tím větší váhu má odpovídající hodnota y_i na odhadnutou $\widehat{y_i}$. Každá hodnota vysvětlované proměnné má vliv na všechny vyrovnané hodnoty, takže vybočující hodnota y_i ovlivňuje všechny hodnoty $\widehat{y_i}$.

Mějme symetrickou čtvercovou idempotentní matici H řádu n

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'. \tag{50}$$

Zmíněné matici se říká projekční matice. Velmi užitečné jsou její diagonální prvky

$$h_{ii} = \mathbf{x_i'}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{x_i}, \quad \text{pro } i\text{=}1, \ldots, n, \tag{51}$$

kde $\mathbf{x_i'}$ představuje i – tý řádek matice $\mathbf{X}$. Tyto prvky vyjadřují váhu hodnoty y_i na odhadnutou hodnotu $\widehat{y_i}$. Pro vektor odhadnutých hodnot platí $\widehat{y_i} = \mathbf{Hy}$.

Diagonální prvky h_{ii} se nazývají efekty nebo také projekční h – prvky. Neznamená to ale, že vysoké h – prvky naznačují velký vliv na odhady parametrů a my jednoduše příslušné pozorování ze souboru vyškrtneme – tzv. zlaté pravidlo. To tedy říká, že se z dat nesmí vyškrtnout jakákoliv hodnota jen ze statistických důvodů. Hodnotu můžeme ze souboru vyškrtnout pouze na základě dostatečných odborných zkušeností nebo jestliže zjistíme příčinu jejího extrémního chování.

Pro lineární model s absolutním členem platí, že součet diagonálních prvků matice $\mathbf{H}$ je

$$\sum_{i=1}^{n} h_{ii} = st(H) = p, \tag{52}$$

kde p je počet regresních parametrů včetně absolutního členu. U lineárního regresního modelu bez absolutního členu je dolní mez diagonálního prvku nula. Pro součet prvků v každém řádku i sloupci platí, že se rovná jedné. Z předpokladu idenpotentnosti matice $\mathbf{H}$ vyplývá

$$h_{ii} = \sum_{i=1}^{n}\sum_{i=1}^{n} h_{ii}^2 = h_{jj}^2 + \sum_{i\neq j=1}^{n} h_{ii}, \tag{53}$$

což znamená, že se každý diagonální prvek rovná součtu čtverců příslušného sloupce i řádku. Prvky, které neleží na hlavní diagonále nabývají hodnot od -1 do 1. Ze vztahu (53) plyne, že všechny mimodiagonální prvky jsou blízké nule, jestliže hodnota h_{ii} je blízko nule nebo jedné.

Průměrná hodnota diagonálního prvku je rovna $\frac{p}{n}$, kde n je počet pozorování, což vidíme ze vztahu (52). Pro pozorování, jehož diagonální prvek matice $\mathbf{H}$ je

větší než $\frac{2p}{n}$ platí, že je považováno za dosti vzdálené od ostatních bodů. Pro počet regresních parametrů větší jak 6 a pro $n-p$ větší jak 12 se doporučuje užít jako kritérium $\frac{3p}{n}$.

Problém odlehlých pozorování je známý už dlouho. V roce 1863 se americký astronom Chauvenet zabýval postupem pro vylučování odlehlých pozorování. Jeho postup ale nebyl nejlepší, protože pro velké počty dat vyřazoval správná pozorování s 40 – ti procentní pravděpodobností. Známý je také Mendělejův postup, kdy Mendělejev vždy odebral 1/3 největších a 1/3 nejmenších pozorování.

5.2 Příklad

Příklad 5.1. *U deseti náhodně vybraných studentů jsme zjišťovali dobu přípravy na kontrolní test v hodinách a dále výsledky testu. Hodnoty jsme si zapisovali do tabulky. Při zápisu do tabulky byly ale chybně zapsány hodnoty x_{10} a y_{10}, které zapisovatel přehodil.*

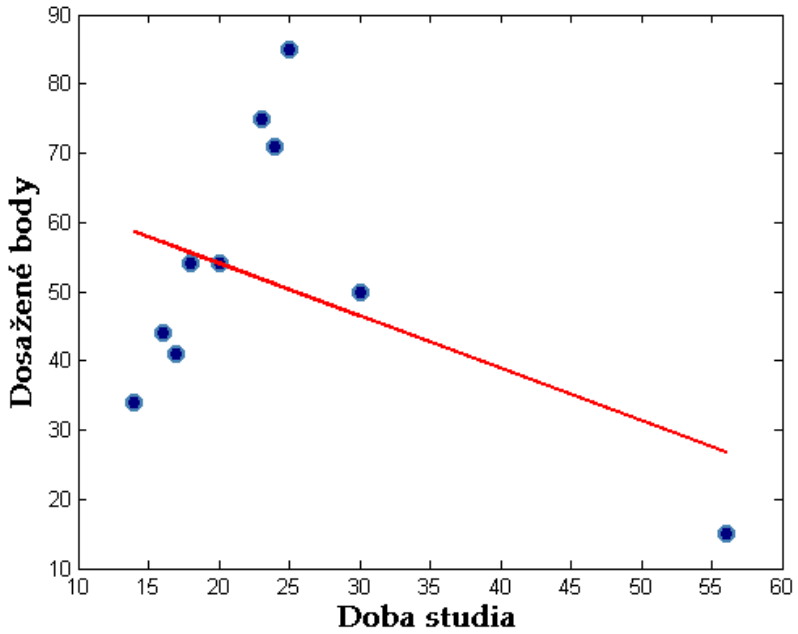
i	1	2	3	4	5	6	7	8	9	10
x_i – doba studia	18	16	23	24	20	17	16	25	14	56
y_i – dosažené body	54	44	75	71	54	41	50	85	34	15

Tabulka 4.2: Získaná data

Užitím metody nejmenších čtverců odhadneme regresní přímku $y = 66{,}86 - 0{,}63x$. *Korelační koeficient je* $-0{,}39$, *my ale očekáváme větší závislost mezi dobou studia a dosaženými body, takže se nám zdá hodnota nízká. Nakreslíme si Obrázek* (11) *a z něj již vidíme, že asi něco není v pořádku, protože nám křivka neprokládá dobře data. Prověříme tedy, zda nějaké pozorování nepatří mezi odlehlé.*

Zajímá nás, která pozorování naše výsledky takto ovlivňují. Vypočteme si matici $\mathbf{H}$ *podle vztahu* (50) *a vypíšeme si její diagonální prvky:*

h_{ii} : 0,1179; 0,1354; 0,1; 0,1009; 0,1063; 0,1259; 0,1355; 0,1033; 0,159; 0,9159.



Obrázek 11: Špatně zapsaná data

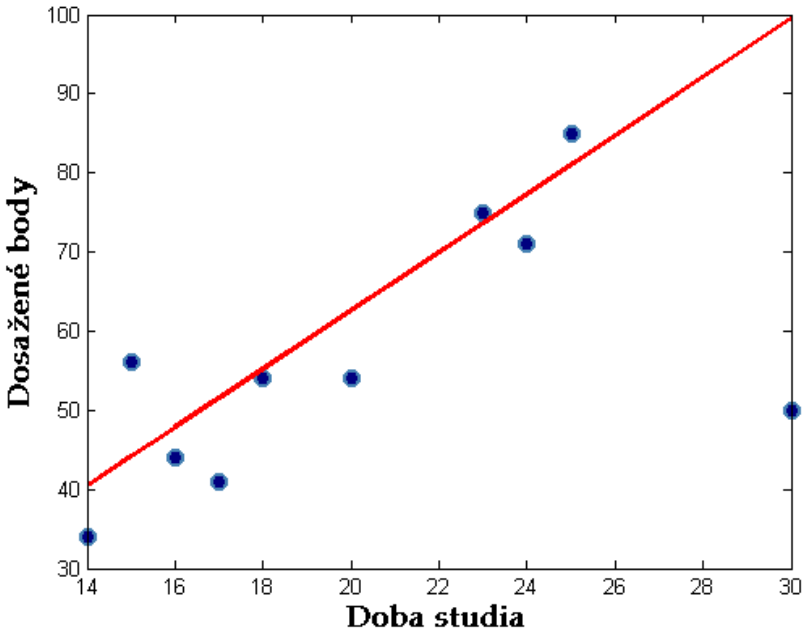
Po porovnání diagonálních prvků s hodnotou kritéria $\frac{2p}{n}$*, tedy s hodnotou* 0,4 *zjistíme, že hodnota posledního diagonálního prvku je mnohem větší. Poslední pozorování nám tedy nejvíce ovlivňuje celý model.*

Při správném zadání (tzn. nemáme přehozeny hodnoty x_{10} *a* y_{10}*) získáme graf hodnot* (12)*, který už vypadá mnohem lépe, protože prokládá data křivkou, která se zdá být v pořádku.*

Po použití MNČ získáme regresní přímku $y = -12{,}61 + 3{,}67x$*. Korelační koeficient je* 0,576*, což už poukazuje na větší závislost mezi proměnnými. Diagonální prvky matice* $\mathbf{H}$ *tentokrát vypadají:*

h_{ii} : 0,1045; 0,1554; 0,2246; 0,291; 0,1102; 0,1229; 0,1554, 0,3715; 0,2627; 0,202.

Po porovnání s hodnotou 0,4 *(zůstane stejná jako minule) vidíme, že ji žádný diagonální prvek nepřekračuje, a tudíž žádné pozorování nemá na odhadnuté parametry nijak extrémní vliv.*



Obrázek 12: Dobré zadání dat

6 Design experimentu

Nezastupitelnou součástí výzkumné činnosti je právě experiment. Design experimentu se zabývá sběrem dat v situaci, kdy je požadovaná informace zatížena nahodilostí (z experimentálních dat totiž náhodnost nelze vypustit). Jedná se o účinný nástroj, který nám pomůže vyhledat optimální strategii nebo postup za pomoci experimentů. Jeho úkolem je nalézt vhodnou volbu bodů x, ve kterých provádíme měření a dále stanovit hodnoty předem daných parametrů, určit oblasti spolehlivosti nebo testovat různé hypotézy. Design experimentu se užívá hlavně v průmyslových a inženýrských oborech, ale např. i v geovědních oborech. Kapitola design experimentu byla zpracovala pomocí literatury [5].

6.1 Fáze experimentu

1. rozpoznání problému,
2. ustanovení dostačujícího množství parametrů (přímo měřitelných),
3. určení cílových parametrů ovlivňujících proces,
4. volba optimálního designu experimentu,
5. provedení experimentu,
6. analýza dat pomocí statistických metod,
7. stanovení závěru a poskytnutí doporučení.

6.2 Základní pojmy

Definice 6.1. *Množinou experimentálních bodů nazveme množinu* $E = \{e_1,\ e_2, \ldots, e_n\}$ *přímo měřitelných parametrů.*

Definice 6.2. *Funkci* $\delta : E \rightarrow < 0, 1 >,\ \delta(e_i) \geq 0,\ i = 1,\ 2, \ldots, N_0,\ \sum_{i=1}^{N_0} \delta(e_i) = 1$ *nazvěme plánem experimentu, kde* N_0 *jsou všechny možné experimentální body.*

Definice 6.3. *Spektrum plánu, označované také jako nosič, je množina* $S_p = \{e_i : \delta(e_i) > 0,\ e_i \in E\}$.

Definice 6.4. *Matici* $\mathbf{M}(\delta) = \sum_{i \in Sp(\delta)} \delta(e_i)\lambda_i \mathbf{f}_i \mathbf{f}_i^{'}$ *nazýváme informační maticí experimentu při plánu* δ. *Přičemž* $\mathbf{f}_i^{'}$ *bereme jako* i *– tý řádek matice plánu* $\mathbf{F}$ *a* λ_i *jsou váhy* i *– tého měření.*

Matice plánu $\mathbf{F}$ vypadá následovně

$$\begin{pmatrix} \frac{\partial f_1(x,\beta_0)}{\partial \beta_1} & \cdots & \frac{\partial f_1(x,\beta_0)}{\partial \beta_k} \\ \vdots & & \vdots \\ \frac{\partial f_{N_0}(x,\beta_0)}{\partial \beta_1} & \cdots & \frac{\partial f_{N_0}(x,\beta_0)}{\partial \beta_k} \end{pmatrix}. \tag{54}$$

Máme několik kritérií optimalit (D – optimalita, A – optimalita, L – optimalita, restringovaná D – optimalita, restringovaná A – optimalita, Σ – optimalita, G – optimalita a I – optimalita). Zmínila bych definice jen dvou z nich, a to D – optimality, která je nejčastěji užívaná a A – optimality.

Definice 6.5. *Nechť* $\triangle_{reg}$ *je třída všech designů takových, že* $\triangle_{reg} = \{\delta : \delta \in \triangle,\ \mathbf{M}(\delta)\}$, *kde* $\triangle$ *je třída všech známých parametrů, je pozitivně definitní. Design* δ_D^* *je D – optimální právě tehdy, když*

$$det(\mathbf{M}^{-1}(\delta_D^*)) = min\{det(\mathbf{M}^{-1}(\delta)) : \delta \in \triangle_{reg}\}. \tag{55}$$

Jde vlastně o minimalizaci determinantu inverze informační matice.

Definice 6.6. *Nechť* $\triangle_{reg}$ *je třída všech designů takových, že* $\triangle_{reg} = \{\delta : \delta \in \triangle,\ \mathbf{M}(\delta)\}$, *kde* $\triangle$ *je třída všech známých parametrů, je pozitivně definitní. Design* δ_D^* *je A – optimální právě tehdy, když*

$$T_r(\mathbf{M}^{-1}(\delta_A^*)) = min\{T_r(\mathbf{M}^{-1}(\delta)) : \delta \in \triangle_{reg}\}. \tag{56}$$

U této optimality minimalizujeme součet disperzí jednotlivých parametrů $\boldsymbol{\beta}$. D – optimalitu si předvedeme na konkrétním příkladě.

6.3 Příklad na D – optimalitu

Příklad 6.1. *Úlohou je připravit měření pro určení funkce* $y = \frac{\alpha_1 x}{\alpha_2 + x}$, *kde* $x = 1; 1{,}5; 2; 2{,}5; \ldots; 50$. *K dispozici máme* 99 *bodů, ve kterých lze provádět měření.*

Startovací odhady parametrů jsme získali výpočtem: $\alpha_{10} \doteq 9{,}9682$ *a* $\alpha_{20} \doteq 3{,}6351$. *Funkce je nelineární a proto ji musíme zlinearizovat. Rozvineme ji pomocí Taylorova rozvoje v přibližném bodě* α_{10}, α_{20}, *přičemž zanedbáme druhé a vyšší derivace. Potom*

$$y_{lin} = \frac{\alpha_{10} x}{\alpha_{20} + x} + \frac{\partial}{\partial \alpha_{10}}\left(\frac{\alpha_{10} x}{\alpha_{20} + x}\right)\triangle\alpha + \frac{\partial}{\partial \alpha_{20}}\left(\frac{\alpha_{10} x}{\alpha_{20} + x}\right)\triangle\alpha =$$

$$= \frac{\alpha_{10} x}{\alpha_{20} + x} + \frac{x}{\alpha_{20} + x}\triangle\alpha - \frac{\alpha_{10} x}{(\alpha_{20} + x)^2}\triangle\alpha,$$

kde $\triangle\alpha = \alpha - \alpha_0$.

Nejprve musíme určit startovací plán. Zvolíme $S_p = \{2, 49, 98\}$ *a k němu odpovídající* $\delta = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$. *Matice* $\mathbf{F}$, *kterou potřebujeme pro výpočet informační matice* $\mathbf{M}$*(def. (6.4)), vypadá:*

$$F = \begin{pmatrix} \frac{x_1}{\alpha_{20}+x_1} & -\frac{x_1 \alpha_{10}}{(\alpha_{20}+x_1)^2} \\ \vdots & \vdots \\ \frac{x_N}{\alpha_{20}+x_{99}} & -\frac{x_1 \alpha_{10}}{(\alpha_{20}+x_{99})^2} \end{pmatrix}.$$

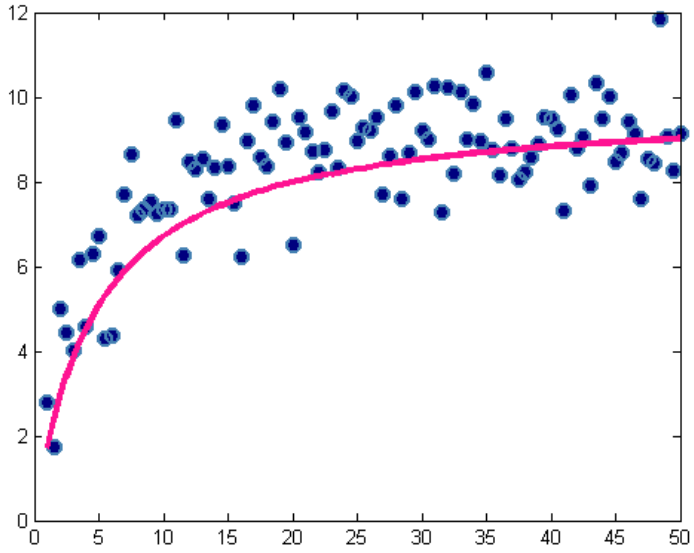
Našim úkolem je určit D – optimální plán, proto musíme maximalizovat kritérium: $max\{\lambda_i f_i' M^{-1}(\delta_D^*) f_i\}$. *Tzn., že pro všechny hodnoty* $i = 1; 1{,}5; 2; 2{,}5; \ldots; 50$ *spočteme zadané kritérium a dále určíme index i*, pro který je hodnota kritéria maximální. Bod s indexem i* pak dodáme do spektra a upravíme vektor* δ. *Výpočet ukončíme, jestliže nastane tzv. stop kritérium* $\lambda_i f_i'(M(\delta))^{-1} f_i < \varepsilon$, $\varepsilon = 0{,}001 + 2$ *(číslo* 2 *nám značí počet parametrů v modelu).*

Průběh několika iterací jsme zaznačili v tabulce níže. Požadované přesnosti 0,001 *bylo dosaženo až po* 1276 *iteracích.*

iterace	δ_1	δ_2	δ_3	δ_5	δ_6	δ_{49}	δ_{50}	δ_{97}	δ_{98}	δ_{99}
0	*0*	*1/3*	*0*	*0*	*0*	*1/3*	*0*	*0*	*1/3*	*0*
1	*0*	*1/4*	*0*	*1/4*	*0*	*1/4*	*0*	*0*	*1/4*	*0*
3	*0*	*1/6*	*0*	*1/3*	*0*	*1/6*	*0*	*0*	*1/6*	*1/6*
5	*0*	*1/8*	*0*	*3/8*	*0*	*1/8*	*0*	*0*	*1/8*	*1/4*
12	*0*	*1/14*	*0*	*3/7*	*0*	*1/14*	*0*	*0*	*1/14*	*5/14*
128	*0*	*1/130*	*0*	*32/65*	*0*	*1/130*	*0*	*0*	*1/130*	*63/130*
...	...	...	...	...	...	...	...	...	...	...
1276	*0*	*1/1278*	*0*	*319/639*	*0*	*1/1278*	*0*	*0*	*1/1278*	*637/1278*

Tabulka 4.3: Průběh několika iterací

Určili jsme tedy, že optimálním designem pro měření parametrů funkce $y = \frac{\alpha_1 x}{\alpha_2 + x}$ *je spektrum s body 2, 5, 49, 98, 99. Velikosti* δ_2, δ_{49}, δ_{98} *jsou ale zanedbatelně malé. Proto bychom mohli použít spektrum jen s body 5 a 99. Získali jsme rovněž optimální odhady parametrů, které vyšly* $\widehat{\boldsymbol{\alpha}} = (9{,}8766252; 4{,}7300569)'$. *Na Obrázku (13) vidíme původní hodnoty a jejich aproximaci získanou metodou nejmenších čtverců.*



Obrázek 13: Aproximace dat

7 Ortogonální regrese

7.1 Co to je ortogonální regrese

Ortogonální regresi poprvé uvedl v roce 1901 anglický matematik a filozof Karl Pearson. Metoda spočívá v minimalizaci ortogonálních projekcí bodů na regresní přímku. Jako ortogonální je označována proto, že neminimalizuje součet druhých mocnin projekcí kolmých na osu x jako klasická MNČ, ale minimalizuje projekce kolmé na regresní přímku.

Ortogonální regrese se dá užít tam, kde některá pozorování jsou důležitější a jiná méně a můžeme jim tak dát váhy (to lze samozřejmě i u MNČ). Dále ji lze např. použít, pokud jsou obě proměnné zatíženy šumem nebo také skrytými vztahy mezi proměnnými. Metoda souvisí s vyrovnáváním dat, kde jsou závisle i nezávisle proměnné (X i Y) zatíženy chybou. Při tvorbě kapitoly jsem čerpala hlavně z [13].

Podstata spočívá v minimalizaci výrazu

$$\Phi(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_1(X_i - \widehat{X_i})^2 + w_2(Y_i - \widehat{Y_i})^2, \tag{57}$$

kde $\widehat{Y_i} = f(\boldsymbol{\beta}, x_i)$, $w_1 = \frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}$, $w_2 = \frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2}$. Tedy platí, že $w_1 + w_2 = 1$.

Mějme běžná označení:

$$\overline{x} = \frac{1}{n}\sum x_i,\ \overline{y} = \frac{1}{n}\sum y_i,\ s_{xy} = \frac{1}{n}\sum x_i y_i - \overline{x}\ \overline{y},$$

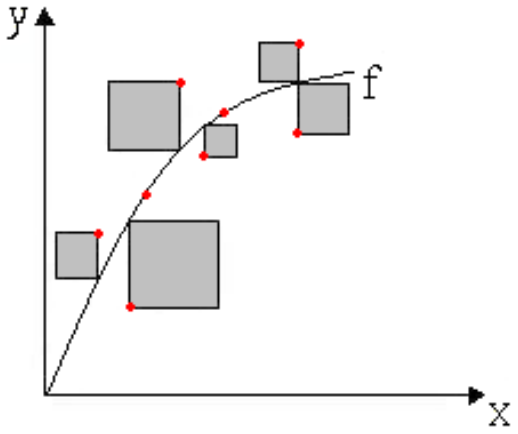
$$s_x^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2,\ s_y^2 = \frac{1}{n}\sum y_i^2 - \overline{y}^2. \tag{58}$$

Věta 7.1. *Mějme $s_{xy} \neq 0$, $s_x^2 > 0$, $s_y^2 > 0$. Parametry přímky $y = \alpha + \beta x$, které získáme metodou ortogonální regrese jsou*

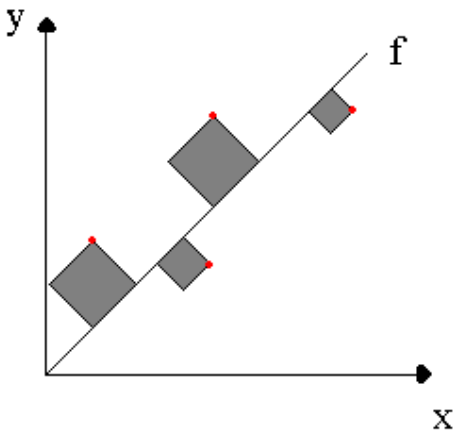
$$\widehat{\beta} = \frac{s_y^2 - s_x^2 + \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2}}{2s_{xy}}, \qquad \widehat{\alpha} = \overline{y} - \widehat{\alpha}\overline{x}. \tag{59}$$

Důkaz: Důkaz věty můžeme nalézt v [13].

Na obrázcích vidíme, jak situace ortogonální regrese (15) vypadá v porovnání s MNČ (14) graficky.



Obrázek 14: Metoda nejmenších čtverců



Obrázek 15: Ortogonální regrese

Aproximační přímka u ortogonální regrese prochází těžištěm bodů, stejně jako u metody nejmenších čtverců.

7.2 Aplikace ortogonální regrese

Nyní přistupme k aproximaci dat jinou křivkou než přímkou. Ortogonální regresi lze popsat nelineárním regresním modelem s podmínkou typu II. Model má tvar

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim_n \left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{pmatrix}, \begin{pmatrix}\sigma_x^2, & 0\\ 0, & \sigma_y^2\end{pmatrix}\right). \tag{60}$$

Věta 7.2. *Uvažujme linearizovaný regresní model přímého měření vektorového parametru se systémem podmínek. Potom BLUE (Best linear unbiased estimator – nejlepší lineární nestranný odhad) parametrů* $\widehat{\widehat{\boldsymbol{\Theta}}}$ *a* $\widehat{\widehat{\boldsymbol{\beta}}}$ *jsou dány vztahy*

$$\widehat{\widehat{\boldsymbol{\beta}}} = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'[\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{C}(\mathbf{C}'\mathbf{S}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{S}^{-1}](\mathbf{b} + \mathbf{B}\widehat{\boldsymbol{\beta}}), \tag{61}$$

$$\widehat{\widehat{\boldsymbol{\Theta}}} = -(\mathbf{C}'\mathbf{S}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{S}^{-1}(\mathbf{b} + \mathbf{B}\widehat{\boldsymbol{\beta}}), \tag{62}$$

$$\mathbf{S} = \mathbf{B}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}' + \mathbf{C}\mathbf{C}', \tag{63}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}, \tag{64}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2\mathbf{V}. \tag{65}$$

Kovarianční matice $var(\widehat{\widehat{\boldsymbol{\beta}}})$, $var(\widehat{\widehat{\boldsymbol{\Theta}}})$ *jsou dány vztahy*

$$var(\widehat{\widehat{\boldsymbol{\beta}}}) = \sigma^2\{\mathbf{I} - (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'[\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{C}(\mathbf{C}'\mathbf{S}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{S}^{-1}]\mathbf{B}\}\cdot \tag{66}$$

$$\cdot (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\{\mathbf{I} - (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}(\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{C}(\mathbf{C}'\mathbf{S}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{S}^{-1})\mathbf{B}'\},$$

$$var(\widehat{\widehat{\boldsymbol{\Theta}}}) = \sigma^2\{(\mathbf{C}'\mathbf{S}^{-1}\mathbf{C})^{-1} - \mathbf{I}\}. \tag{67}$$

Důkaz: Důkaz viz [18].

Příklad 7.1. *Uvažujme příklad z kapitoly design experimentu, ve kterém máme funkci* $y = \frac{\alpha_1 x}{\alpha_2 + x}$, *kde* $x = 1; 1{,}5; 2; 2{,}5; \ldots; 50$. *Přičemž je k dispozici* 99 *měření.*

Parametry α_1 *a* α_2 *musí splnit pro* $i = 1, \ldots, 99$ *podmínky*

$$g_i(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = \nu_i - \frac{\alpha_1 \mu_i}{\alpha_2 + \mu_i} = 0. \tag{68}$$

Podmínky jsou ale nelineární, a proto je musíme zlinearizovat pomocí Taylorova rozvoje. V linearizovaném tvaru mají podmínky tvar $\mathbf{B}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\alpha} + \mathbf{b} = 0$, *kde* $\mathbf{B} = \frac{\partial \mathbf{g}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0)}{\partial(\boldsymbol{\mu}', \boldsymbol{\nu}')'}$, $\mathbf{C} = \frac{\partial \mathbf{g}(\boldsymbol{\mu}^0, \boldsymbol{\nu}^0, \boldsymbol{\alpha}^0)}{\partial \boldsymbol{\alpha}'}$, *a* $\mathbf{b} = \mathbf{g}(\boldsymbol{\mu}^0, \boldsymbol{\nu}^0, \boldsymbol{\alpha}^0)$ *v přibližném řešení* $(\boldsymbol{\mu}^0, \boldsymbol{\nu}^0, \boldsymbol{\alpha}^0)$.

$$\mathbf{B} = \begin{pmatrix} \frac{\partial g_1}{\partial \mu_1} & \cdots & \frac{\partial g_1}{\partial \mu_{99}} & \frac{\partial g_1}{\partial \nu_1} & \cdots & \frac{\partial g_1}{\partial \nu_{99}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_{99}}{\partial \mu_1} & \cdots & \frac{\partial g_{99}}{\partial \mu_{99}} & \frac{\partial g_{99}}{\partial \nu_1} & \cdots & \frac{\partial g_{99}}{\partial \nu_{99}} \end{pmatrix}, \tag{69}$$

kde pro $i, j = 1, \ldots, n$ *je*

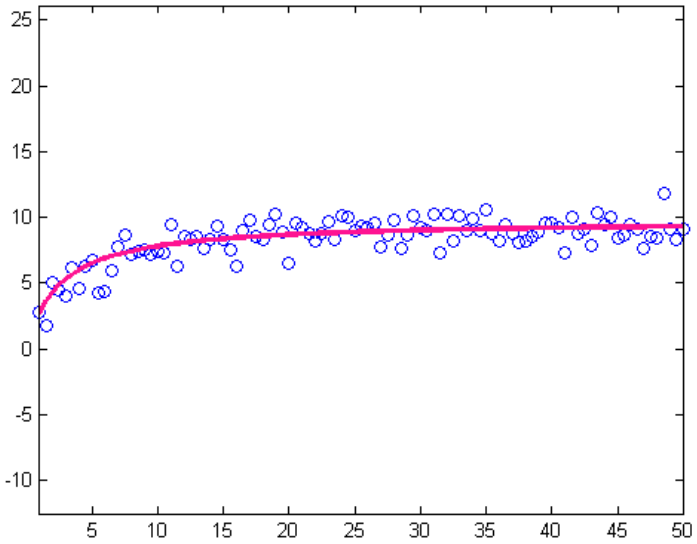
$$\mathbf{B}_{i,j} = \begin{cases} \frac{\alpha_1 \mu_i - \alpha_1 \alpha_2 - \alpha_1 \mu_i}{(\alpha_2 + \mu_i)^2} & \text{pro } i = j, \\ 1 & \text{pro } j = n + i, j > i, \\ 0 & \text{jinde,} \end{cases}$$

$$\mathbf{C} = \begin{pmatrix} \frac{\partial g_1}{\partial \alpha_1} & \frac{\partial g_1}{\partial \alpha_2} \\ \vdots & \vdots \\ \vdots & \vdots \\ \frac{\partial g_{99}}{\partial \alpha_1} & \frac{\partial g_{99}}{\partial \alpha_2} \end{pmatrix} = \begin{pmatrix} -\frac{x_1}{\alpha_2 + x_1} & \frac{\alpha_1 x_1}{(\alpha_2 + x_1)^2} \\ \vdots & \vdots \\ \vdots & \vdots \\ -\frac{x_{99}}{\alpha_2 + x_{99}} & \frac{\alpha_1 x_{99}}{(\alpha_2 + x_{99})^2} \end{pmatrix}. \tag{70}$$

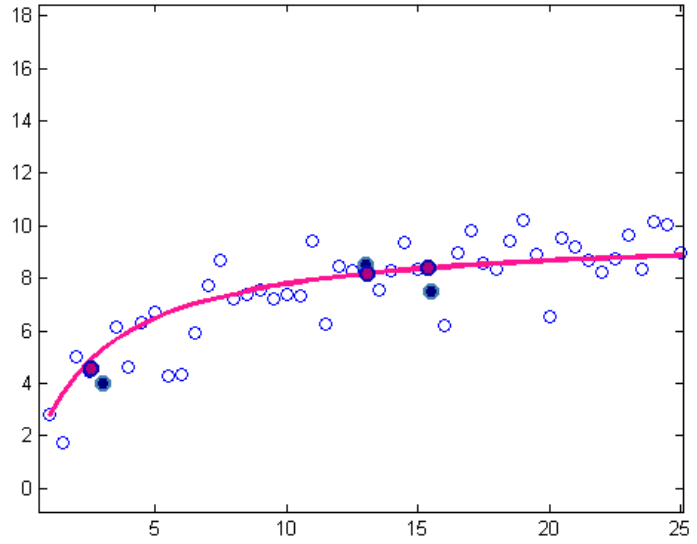
Po užití vzorců (61) *a* (62) *dostaneme odhady parametrů získané metodou ortogonální regrese aplikované na model s podmínkou typu II. Za počáteční řešení jsme brali odhady parametrů získané pomocí designu experimentu*
$\boldsymbol{\alpha}^0 = (9{,}8766252; 4{,}7300569)'$.

Vypočetli jsme $\widehat{\boldsymbol{\alpha}} = (9{,}7739458; 2{,}5480354)'$. *Diagonální prvky varianční matice odhadů jsou* $diag(var(\widehat{\boldsymbol{\alpha}})) = (0{,}26051^2; 0{,}66524^2)'$.

Na Obrázku (16) *vidíme aproximaci dat získanou metodou ortogonální regrese. Obrázek* (17) *znázorňuje přiblížení předchozího Obrázku* (16). *Jsou na něm vyznačeny tři náhodná data a jejich aproximace. Pokud spojíme bod a její aproximaci, tak spojnice je totožná s normálou k funkci (princip ortogonální regrese).*



Obrázek 16: Aproximace dat metodou ortogonální regrese



Obrázek 17: Přiblížení

8 Problém linearizace

Při popisování určité závislosti mezi veličinami není vždy vhodné užívat jen klasický lineární model. Tedy, že regresní vztah mezi veličinami x a Y nelze jednoduše popsat. Proto se pak vztah popisuje pomocí nelineárních modelů. Modely ale nemusí mít tak jasnou interpretaci parametrů. A mohou vznikat i další problémy. Někdy se ale prostě nelineárním modelům nevyhneme. Zdrojem byla hlavně [15, 21].

Modely mohou být zcela lineární, kdy je linearita zastoupena jak v parametrech, tak i v regresorech. Jsou velmi oblíbené, protože jsou velmi jednoduché. Dále existují modely, které jsou lineární z hlediska všech parametrů, ale nelineární z hlediska vysvětlujících proměnných x. Do této skupiny modelů spadá i velmi užívaný model regresní paraboly – zvláště pak kvadratický model, který jsme užili v příkladu v K apitole 2. Nelineární modely v parametrech jsou ze všech modelů nejsložitější.

Některé nelineární modely jsou převoditelné určitou transformací na lineární modely. Po provedení transformace závisí nová regresní funkce na parametrech už jen lineárně, což může být někdy velmi přínosné, protože se tak model značně zjednoduší. Musíme si ale uvědomit, že odhady, které získáme metodou linearizace, nejsou stejné jako odhady získané z původního modelu.

8.1 Regresní modely

Pro modelování chemického experimentu použijeme dva lineární regresní modely. Nejprve aplikujeme model bez podmínek a v druhém případě vyzkoušíme model se systémem podmínek (model se systémem podmínek typu II).

8.1.1 Nepřímé měření vektorového parametru bez podmínek

Představme si situaci, kdy je třeba nalézt odhady neznámých parametrů nelineární funkce f pro data z chemického experimentu, kde

$$f(x) = \Theta_1 \cdot x + \Theta_2 + \Theta_3 \mid x - \Theta_4 \mid, \qquad x \in \langle 0, \infty). \tag{71}$$

Funkce popisuje závislost mezi koncentrací (mmol/L) a potenciálem (mV) chemické látky.

Nejdříve danou nelineární funkci aproximujeme lineárním členem Taylorovy řady kolem přibližného bodu $\boldsymbol{\Theta^0}=[\Theta_1^0, \Theta_2^0, \Theta_3^0, \Theta_4^0]'$. Přibližný bod je bod, ve kterém budeme funkci rozvíjet. V tomto bodě budou výsledné aproximace nejpřesnější. Při aproximaci vezmeme přitom v potaz jen derivace prvního řádu, ostatní derivace, tedy druhého a vyšších řádů, zanedbáme. Zadanou nelineární funkci aproximujeme Taylorovým polynomem prvního stupně tj. lineární funkcí. Zanedbání vyšších derivací má samozřejmě vliv na chybu modelu, což je ale cena za zjednodušení nelineární funkce. Zjištěná chyba je však v okolí našeho přibližného bodu přijatelně malá, takže linearizace je pozitivní krok ke zjednodušení modelu.

Způsob, který jsme právě naznačili, nemusí být ale vždy možný. Není možný například, když nelineární funkce není v přibližném bodě spojitá, pokud nemá svou vlastní derivaci atd. V těchto a dalších případech není linearizace vůbec přesná. Musíme tak užít mnohem složitější postupy.

Stochastický model pro algoritmus je

$$\mathbf{Y} = \mathbf{JA}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{72}$$

dále platí, že

$$\mathrm{h}(\mathbf{JA}) = k \leq n, \quad \mathrm{var}(\mathbf{Y}) = \Sigma \quad \text{p.d.},$$

kde n je počet měření a k je počet parametrů. Model tedy můžeme zapsat

$$\mathbf{Y} \sim (f(\boldsymbol{\Theta}), \boldsymbol{\Sigma}), \tag{73}$$

přičemž $f(\boldsymbol{\Theta})$ je známá nelineární funkce s vektorem parametrů $\boldsymbol{\Theta}$. Tato funkce má tvar (71) Budeme tedy linearizovat (z nelineární funkce na lineární). Pro

provedení linearizace je ovšem nutné znát přibližné hodnoty parametrů $\boldsymbol{\Theta}$. Ty si pro lepší přehlednost označíme $\boldsymbol{\Theta}^0$. Linearizovaný model je tvaru

$$\overline{\mathbf{Y}} \sim_n (\mathbf{F}\Theta, \sigma^2\mathbf{V}), \tag{74}$$

kde $\boldsymbol{\Theta}=[\Theta_1, \Theta_2, \Theta_3, \Theta_4]^T$ je vektor neznámých parametrů, které chceme určit, $\mathbf{V}$ je známá pozitivně definitní matice, $\sigma^2 > 0$ a

$$\{\mathbf{F}\}_{i\cdot} = \frac{\partial f(x_i, \boldsymbol{\Theta}^0)}{\partial \boldsymbol{\Theta}'} = \left(\frac{\partial f_i(x_i, \boldsymbol{\Theta}^0)}{\partial \Theta_1}, \frac{\partial f_i(x_i, \boldsymbol{\Theta}^0)}{\partial \Theta_2}, \frac{\partial f_i(x_i, \boldsymbol{\Theta}^0)}{\partial \Theta_3}, \frac{\partial f_i(x_i, \boldsymbol{\Theta}^0)}{\partial \Theta_4} \right) \tag{75}$$

je matice plánu, přičemž

$$\frac{\partial f_i}{\partial \Theta_1} = x_i, \quad \frac{\partial f_i}{\partial \Theta_2} = 1, \quad \frac{\partial f_i}{\partial \Theta_3} = \mid x_i - \Theta_4 \mid, \quad \frac{\partial f_i}{\partial \Theta_3} = sgn(x_i - \Theta_4)(-\Theta_3).$$

Věta 8.1. *Mějme linearizovaný regresní model nepřímého měření vektorového parametru* (74). *Potom BLUE – odhad parametru* $\boldsymbol{\Theta}$ *je dán ve tvaru*

$$\widehat{\boldsymbol{\Theta}} = \left(\mathbf{F}'\mathbf{V}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{V}^{-1}\mathbf{Y} \right). \tag{76}$$

Varianční matice je

$$var(\widehat{\boldsymbol{\Theta}}) = \sigma^2 \left(\mathbf{F}'\mathbf{V}^{-1}\mathbf{F} \right)^{-1}. \tag{77}$$

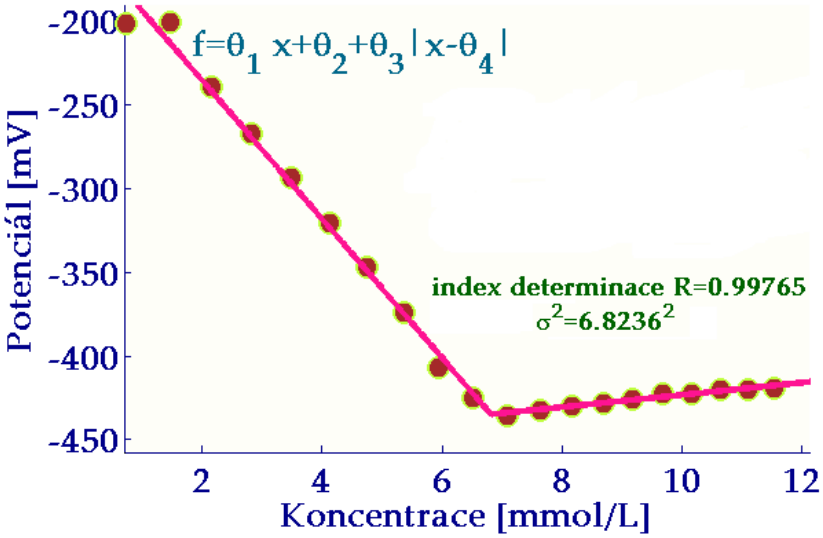
Do vzorce (76) pro nepřímé měření vektorového parametru bez podmínek jsme nyní dosadili hodnoty dat z chemického příkladu. Získali jsme tak odhady parametrů $\Theta_1,\ \Theta_2,\ \Theta_3$ a Θ_4. Pro přibližný výpočet jsme potřebovali vědět počáteční odhady parametrů. Ty jsme si vhodně zvolili $\boldsymbol{\Theta}^0=(-20; -268; 21; 6,5)$.

$$\widehat{\boldsymbol{\Theta}} = (-18{,}8855; -306{,}3788; 22{,}5396; 6{,}8106)'.$$

Po dosazení do (77) dostáváme varianční matici. Diagonální prvky varianční matice odhadů jsou $diag(var(\widehat{\boldsymbol{\Theta}})) = (0{,}28079^2; 2{,}2281^2; 0{,}28079^2; 0{,}039799^2)'$.

Index determinace $R^2 = 1 - \frac{\sum(Y-\widehat{Y}^2)}{\sum(\widehat{Y}-\overline{Y})^2} = 0{,}99765$. Čím blíže jedné je index, tím lepší je zvolená aproximace.

Aproximace dané funkce získaným algoritmem je zobrazena na Obrázku (18).



Obrázek 18: Aproximace algoritmem 1

8.1.2 Neúplné přímé měření vektorového parametru s podmínkami typu II

Nyní ukážeme, jak se vypořádat sice s lineárním modelem, ale modelem s podmínkou typu II. Nelinearita se zde vyskytuje v podobě nelineárních podmínek, které je nutno linearizovat. Obecný stochastický model pro algoritmus je

$$\mathbf{Y} = \mathbf{J}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \mathbf{B}\boldsymbol{\beta}_1 + \mathbf{C}\boldsymbol{\beta}_2 + \mathbf{b} = \mathbf{0}, \tag{78}$$

dále platí, že

$$h(\mathbf{J}) = k_1 \leq n, \quad h(\mathbf{C}) = k_2 < q, \quad h(\mathbf{B}, \mathbf{C}) = q < k_1 + k_2, \quad var(\mathbf{Y}) = \Sigma \;\; p.d.,$$

kde n je počet měření, k je počet parametrů, q je počet podmínek, k_1 je počet přímo měřitelných parametrů $\boldsymbol{\beta}_1$ a k_2 je počet nepřímo měřitelných parametrů $\boldsymbol{\beta}_2$ (umíme je určit jen z podmínky).

Nyní se budeme snažit data proložit dvěmi parabolami. V našem případě

budeme za $\boldsymbol{\beta}_1$ uvažovat $\boldsymbol{\beta}$ a za $\boldsymbol{\beta}_2$ uvažujme $\boldsymbol{\Theta}$. Nelineární model můžeme popsat

$$\mathbf{Y} \sim N_n \left[\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \boldsymbol{\Sigma} \right], \tag{79}$$

přičemž zároveň musí platit soustava $n+1$ podmínek $\mathbf{g}(\boldsymbol{\beta}, \boldsymbol{\Theta})$

$$\begin{gathered} \Theta_1 + \Theta_2 x_1 + \Theta_3 x_1^2 - \beta_1 = 0, \\ \dots \\ \Theta_1 + \Theta_2 x_s + \Theta_3 x_1^2 - \beta_s = 0, \\ \Theta_3 + \Theta_4 x_{s+1} + \Theta_5 x_{s+1}^2 - \beta_{s+1} = 0, \\ \dots \\ \Theta_3 + \Theta_4 x_n + \Theta_5 x_n^2 - \beta_n = 0, \\ \Theta_1 + \Theta_2 \Theta_7 + \Theta_3 \Theta_7^2 - \Theta_4 - \Theta_5 \Theta_7 - \Theta_6 \Theta_7^2 = 0. \end{gathered} \tag{80}$$

Našim cílem je najít odhady $\widehat{\widehat{\boldsymbol{\beta}}}$ (dvě stříšky znamenají, že se jedná o odhad v modelu s podmínkami) skutečných hodnot $\boldsymbol{\beta}$ a také nás zajímají odhady parametru $\widehat{\widehat{\boldsymbol{\Theta}}}$.

I v tomto případě musíme provést linearizaci pomocí aproximace Taylorova rozvoje. Nelineární podmínky $\mathbf{g}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ tedy zlinearizujeme Taylorovým rozvojem a opět zanedbáme derivace druhého a vyšších řádů. Linearizované podmínky můžeme zapsat jako

$$\mathbf{B}_l \boldsymbol{\beta} + \mathbf{C}_l \boldsymbol{\Theta} + \mathbf{b}_l = \mathbf{0}, \tag{81}$$

kde

$$\mathbf{B}_l = \frac{\partial \mathbf{g}(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\Theta}^{\mathbf{0}})}{\partial \boldsymbol{\beta}'}, \quad \mathbf{C}_l = \frac{\partial \mathbf{g}(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\Theta}^{\mathbf{0}})}{\partial \boldsymbol{\Theta}'}, \quad \mathbf{b}_l = \mathbf{g}(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\Theta}^{\mathbf{0}}). \tag{82}$$

Pro lepší přehlednost si naše konkrétní matice zapíšeme maticově

$$\mathbf{B}_l = \begin{pmatrix} -1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & -1 \\ 0 & 0 & \dots & 0 \end{pmatrix},$$

$$\mathbf{C}_l = \begin{pmatrix} \frac{\partial g_1}{\partial \Theta_1} & \frac{\partial g_1}{\partial \Theta_2} & \frac{\partial g_1}{\partial \Theta_3} & \frac{\partial g_1}{\partial \Theta_4} & \frac{\partial g_1}{\partial \Theta_5} & \frac{\partial g_1}{\partial \Theta_6} & \frac{\partial g_1}{\partial \Theta_7} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_{n+1}}{\partial \Theta_1} & \frac{\partial g_{n+1}}{\partial \Theta_2} & \frac{\partial g_{n+1}}{\partial \Theta_3} & \frac{\partial g_{n+1}}{\partial \Theta_4} & \frac{\partial g_{n+1}}{\partial \Theta_5} & \frac{\partial g_{n+1}}{\partial \Theta_6} & \frac{\partial g_{n+1}}{\partial \Theta_7} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & x_1 & x_1^2 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_s & x_s^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{s+1} & x_{s+1}^2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & x_n & x_n^2 & 0 \\ 1 & \Theta_7 & \Theta_7^2 & -1 & -\Theta_7 & -\Theta_7^2 & * \end{pmatrix},$$

kde $* = \Theta_2 + 2\Theta_3\Theta_7 - \Theta_5 - 2\Theta_6\Theta_7$.

Nyní dosadíme do vzorců pro model s podmínkou typu II dané hodnoty z chemického pokusu. Při výpočtu jsme používali počáteční řešení, které jsme získali tak, že jsme nejprve odhadli MNČ parametry dvou parabol. Poslední parametr jsme si zvolili jako jejich průsečík. Počítali jsme tedy s počátečním řešením

$$\boldsymbol{\Theta}^0 = (-168{,}14603; -30{,}18829; -1{,}53823; -502{,}75189; 13{,}06670; -0{,}50469; 6{,}8).$$

Za počáteční odhad $\widehat{\boldsymbol{\beta}}$ jsme brali původní naměřené hodnoty Y. Po provedení výpočtu (Příloha 6) s pomocí vzorců (61) a (62) jsme získali odhady parametrů v modelu s podmínkami typu II.

$$\widehat{\widehat{\boldsymbol{\Theta}}} = (-168{,}1460; 30{,}1883; -1{,}5382; -502{,}7518; 13{,}0667; -0{,}5047; 6{,}8000)'.$$

$$\widehat{\widehat{\boldsymbol{\beta}}} = (-201{,}54; -200{,}63; -239{,}67; -267{,}67; -29{,}33; -321; -347; -374; -407; -425{,}33;$$

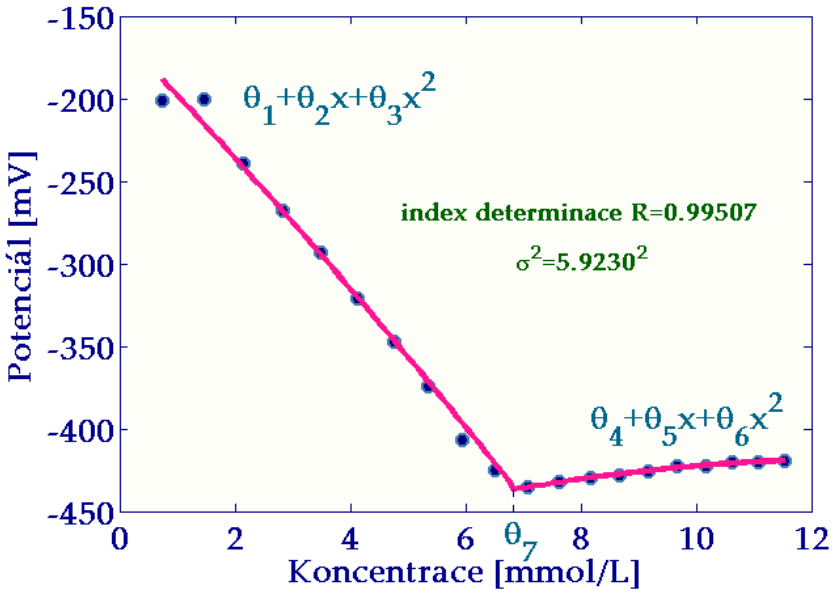
$$-435{,}67; -432; -429{,}67; -428; -425{,}67; -422{,}67; -422{,}67; -420{,}33; -420; -419{,}33)'.$$

Po dosazení do (67) získáme varianční matici odhadů. Diagonální prvky varianční matice odhadů jsou

$diag(var(\widehat{\Theta})) = (5{,}86338^2; 3{,}65037^2; 0{,}48779^2; 70{,}28072^2; 15.31809^2; 0{,}81949^2; 0{,}11031^2)'$.

Index determinace $R^2 = 0.99700$. V porovnání s indexem determinace pro první algoritmus bez podmínek vidíme, že index determinace pro model s podmínkami je o trochu menší.

Výsledky algoritmu jsou zobrazeny na Obrázku (19).



Obrázek 19: Model s podmínkami typu II

8.2 Linearizační oblasti a míry křivosti

Následující podkapitola byla zpracována pomocí [4, 5, 17]. Nelinearitu neboli křivost je potřeba měřit. Měření spočívá v porovnání lineární a kvadratické aproximace. Uvedeme dvě křivosti, a to *vnitřní křivost* (intrinsic curvature) a *parametrickou křivost* (parameter – effects curvature). Hodnota vnitřní křivosti je nezávislá na hodnotě parametrické křivosti.

Batesovu a Wattsovu vnitřní křivost z roku 1980 vypočteme podle vzorce

$$K^{(int)}(\boldsymbol{\beta}_0) = \sup\left\{ \frac{\sqrt{\boldsymbol{\kappa}'(\delta\mathbf{b})\boldsymbol{\Sigma}^{-1}\mathbf{M}_F^{\Sigma^{-1}}\boldsymbol{\kappa}(\delta\mathbf{b})}}{\delta\mathbf{b}'\mathbf{C}\delta\mathbf{b}} : \delta\mathbf{b} \in R^k \right\}. \tag{83}$$

Batesova a Wattsova parametrická křivost je

$$K^{(par)}(\boldsymbol{\beta}_0) = \sup\left\{ \frac{\sqrt{\boldsymbol{\kappa}'(\delta\mathbf{b})\boldsymbol{\Sigma}^{-1}\mathbf{P}_F^{\Sigma^{-1}}\boldsymbol{\kappa}(\delta\mathbf{b})}}{\delta\mathbf{b}'\mathbf{C}\delta\mathbf{b}} : \delta\mathbf{b} \in R^k \right\}. \tag{84}$$

Přičemž $\boldsymbol{\kappa}$ je matice druhých derivací funkce f, $\delta\mathbf{b}$ je libovolný vektor z k – rozměrného prostoru, $\boldsymbol{\Sigma}^{-1}$ je inverze známe p. d. varianční matice, $\mathbf{C} = (\mathbf{M}_F^{\Sigma^{-1}}\boldsymbol{\Sigma}\mathbf{M}_F^{\Sigma^{-1}})^+$ (znaménko plus značí Moore – Penroseovu inverzi), $\mathbf{F}$ je matice plánu tzn. matice prvních derivací funkce f, $\mathbf{M}_F^{\Sigma^{-1}}$ a $\mathbf{P}_F^{\Sigma^{-1}}$ jsou projekční matice, pro které platí

$$\mathbf{M}_F^{\Sigma^{-1}} = \mathbf{I} - \mathbf{P}_F^{\Sigma^{-1}},$$

$$\mathbf{P}_F^{\Sigma^{-1}} = \mathbf{F}(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^-\mathbf{F}'\boldsymbol{\Sigma}^{-1}$$

(znaménko minus značí pseudoinverzi).

Obecně můžeme použít následující vzorec pro výpočet $K^{(int)}(\boldsymbol{\beta}_0)$, $K^{(par)}(\boldsymbol{\beta}_0)$ a dalších měr nelinearity, které zde nejsou uvedeny

$$C(\beta_0) = sup\left\{ \frac{\sqrt{\alpha'_{\delta u}\mathbf{S}\alpha_{\delta u}}}{\delta\mathbf{u}'\mathbf{R}\delta\mathbf{u}} : \delta\mathbf{u} \in R^k \right\}, \tag{85}$$

kde $\alpha'_{\delta u} = (\delta\mathbf{u}'\mathbf{A_1}\delta\mathbf{u}, \ldots, \delta\mathbf{u}'\mathbf{A}_r\delta\mathbf{u})$, kde $\mathbf{A}_1, \ldots, \mathbf{A}_r$ jsou $k \times k$ symetrické matice, $\mathbf{S}$ je $r \times r$ p. s. d. matice a $\mathbf{R}$ je $k \times k$ p. s. d. matice, δu je libovolný vektor z k – rozměrného prostoru.

8.2.1 Algoritmus na hledání suprema

Algoritmus pro určení $C(\beta_0)$ lze najít v [5].

1. krok – zvolit libovolný vektor $\delta\mathbf{u}_1$ z k – rozměrného prostoru, pro který platí $\delta\mathbf{u}_1'\delta\mathbf{u}_1 = 1$.

2. krok – určit další vektor jako

$$\delta\mathbf{s} = \mathbf{R}^{-1}(\mathbf{A}_1\delta\mathbf{u}_1, \ldots, \mathbf{A}_r\delta\mathbf{u}_1)\mathbf{S}\begin{pmatrix}\delta\mathbf{u}_1'\mathbf{A}_1\delta\mathbf{u}_1\\ \vdots\\ \delta\mathbf{u}_1'\mathbf{A}_r\delta\mathbf{u}_1\end{pmatrix}. \tag{86}$$

Tento vektor použijeme pro určení vektoru $\delta\mathbf{u}_2 = \frac{\delta\mathbf{s}}{\sqrt{\delta\mathbf{s}'\delta\mathbf{s}}}$.

3. krok – pokud platí nerovnost $\delta\mathbf{u}_2'\delta\mathbf{u}_1 \geq 1-\varepsilon$, kde ε je malé kladné číslo, pak iterační proces ukončíme a

$$C(\beta_0) = \frac{\sqrt{\alpha_{\delta u_2}'\mathbf{S}\alpha_{\delta u_2}}}{\delta\mathbf{u}_2'\mathbf{R}\delta\mathbf{u}_2}. \tag{87}$$

Pokud platí $\delta\mathbf{u}_2'\delta\mathbf{u}_1 < 1-\varepsilon$, pak se vrátíme ke kroku 1, přičemž ale nyní užijeme jako libovolný vektor vektor $\delta\mathbf{u}_2$ (nahradíme jim $\delta\mathbf{u}_1$) .

8.2.2 Linearizační oblasti

Pokud skutečná hodnota parametru $\boldsymbol{\beta}$ leží v linearizační oblasti, můžeme nelineární model nahradit lineárním. Pro výpočet odhadu a jeho přesnosti můžeme pak použít vztahy z Kapitoly 2. Zmiňme jen dvě možné linearizační oblasti. Existuje jich sice více, ale prostor v práci je nedovoluje všechny popsat.

$\mathcal{O}_a$ linearizační oblast – měří shodu dat s linearizovaným modelem

Jestliže

$$\delta\mathbf{b} \in \mathcal{O}_a(\beta_0) = \left\{ \delta\mathbf{b} \colon (\delta\mathbf{b})'\mathbf{C}\delta\mathbf{b} \leq \frac{2\sqrt{\delta_{\max}}}{K^{(int)}(\beta_0)} \right\}, \tag{88}$$

kde $\delta_{\max}$ je dána rovnicí

$$P\{\chi^2_{n-k}(\delta_{\max}) \geq \chi^2_{n-k}(0; 1-\alpha)\} = \alpha + \varepsilon, \tag{89}$$

pak $P\{\mathbf{v}'\Sigma^{-1}\mathbf{v} \geq \chi^2_{n-k}(0; 1-\alpha)) \leq \alpha + \varepsilon$, kde $\mathbf{v} = M_F^{\Sigma^{-1}}(Y - f_0)$, f_0 je funkční hodnota funkce f v počátečním řešení.

$\mathcal{O}_b$ linearizační oblast – bias odhadu

Nechť $c = a\sqrt{\chi^2_k(0; 1-\alpha)}$. Jestliže

$$\delta\mathbf{b} \in \mathcal{O}_b(\beta_0) = \left\{ \delta\mathbf{b} \colon \delta\mathbf{b}'\mathbf{F}'\Sigma^{-1}\mathbf{F}\delta\mathbf{b} \leq \frac{2c}{K^{(par)}(\beta_0)} \right\}, \tag{90}$$

pak

$$\forall\{\mathbf{h} \in R^k\} \left| b^*_h(\delta\mathbf{b}) \right| \leq c\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}. \tag{91}$$

8.2.3 Příklad

Příklad 8.1. *Mějme opět stejný příklad z kapitoly design experimentu s funkcí* $y = \frac{\beta_1 x}{\beta_2 + x}$, *kde* $x = 1; 1{,}5; 2; 2{,}5; \ldots; 50$.

Matice prvních derivací, kterou jsme při výpočtech používali vypadá

$$\mathbf{F} = \left(\frac{x_i}{\beta_{20} + x_i}, -\frac{x_1 \beta_{10}}{(\beta_{20} + x_i)^2} \right).$$

Matice druhých derivací

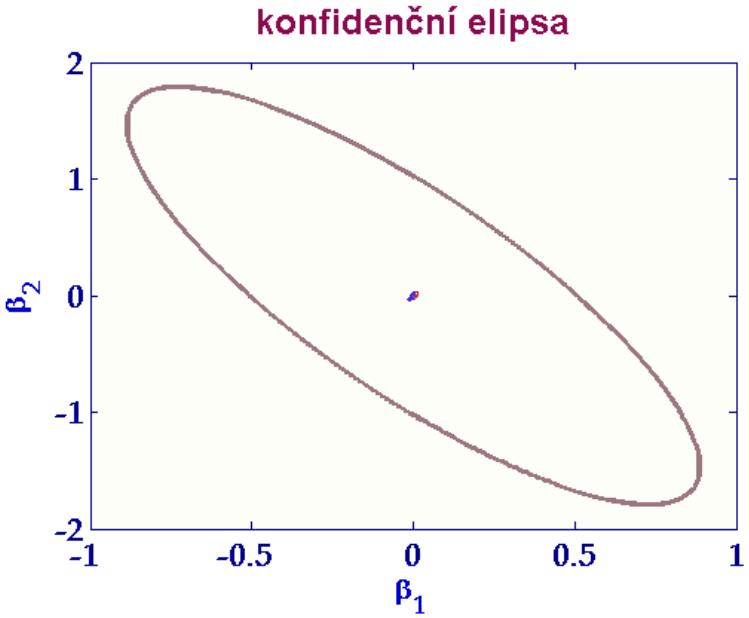
$$\boldsymbol{\kappa} = \begin{pmatrix} 0 & \frac{-x_i}{(\alpha_{20}+x_i)^2} \\ \frac{-x_i}{(\beta_{20}+x_i)^2} & \frac{2x_i\beta_{10}}{(\beta_{20}+x_i)^3} \end{pmatrix}.$$

Při využití předchozích dvou matic a vztahů pro míry křivosti (83) *a* (84) *jsou* $K^{(int)} = 0{,}001530845$ *a* $K^{(par)} = 0{,}00326417$.

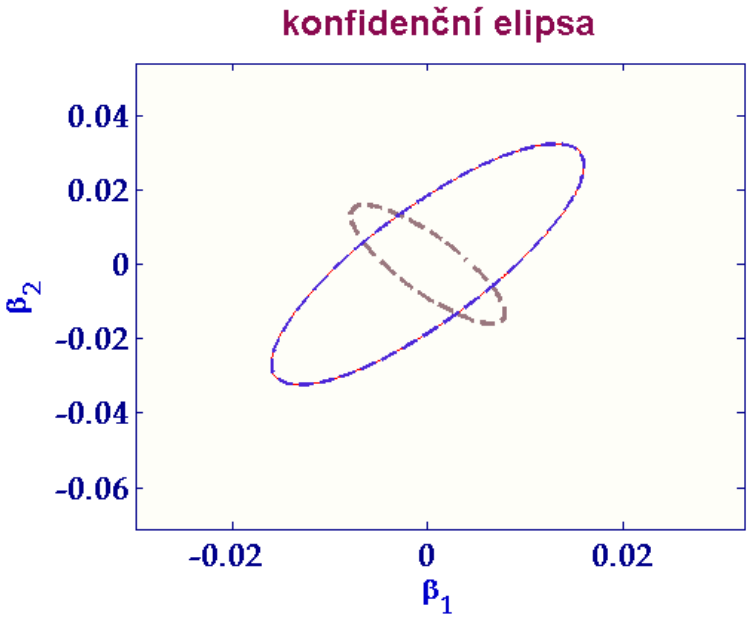
Pro linearizační oblasti $\mathcal{O}_a$ a $\mathcal{O}_b$ dostaneme Obrázky (20), (21). Obrázek (20) znázorňuje $\mathcal{O}_b$ linearizační oblast a Obrázek (21) znázorňuje $\mathcal{O}_a$ linearizační oblast (modře) a konfidenční elipsu (čárkovaně).

Konfidenční elipsa určuje množinu bodů mimo tuto elipsu, kde leží skutečná hodnota s p-stí nejvýše α % (v našem příkladě jsme zvolili 5 %).

Vidíme, že část konfidenční elipsy neleží v linearizační oblasti, je možné, že pro některá počáteční řešení metoda nebude konvergovat. Nikdy totiž neznáme skutečnou hodnotu parametru $\boldsymbol{\beta}$, ale jen odhad $\widehat{\boldsymbol{\beta}}$ a protože víme, že skutečná hodnota neleží v konfidenční elipse s pravděpodobností 5 %, může se stát, že pro získaný odhad nebudeme v linearizační oblasti.



Obrázek 20: O_b linearizační oblast



Obrázek 21: O_a linearizační oblast a konfidenční elipsa

Závěr

Hlavním cílem diplomové práce bylo popsat, analyzovat a názorně na příkladech ukázat problémy, které se mohou vyskytnout při regresní analýze. Nebylo to příliš jednoduché, protože každý problém je sám o sobě velmi složitý a také bylo nutné pracovat s velkou spoustou literatury. Nakonec se ale cíl diplomové práce povedlo naplnit a já tak mohla přinést shrnutí všech podstatných problémů.

Nejprve byli čtenáři seznámeni s historickými metodami aproximace přímkou. Dozvěděli jsme se, že dnes nejlepší a nejužívanější je metoda nejmenších čtverců (MNČ), která má nejmenší reziduální součet čtverců a známe u ní nestranný odhad rozptylu náhodné chyby. Další v současnosti využívanou metodou je Laplaceova metoda nejmenších absolutních odchylek (LAD), která minimalizuje absolutní součet reziduí.

Další kapitola byla zaměřena na pojmy z klasických kurzů statistiky. Snažili jsme se vzorce přímo aplikovat na příkladu z oblasti zkoumání chemického procesu. Zmínili jsme pojmy jako stochastický model, odhady parametrů 1. a 2. řádu, oblasti spolehlivosti, testování hypotéz a silofunkce.

Ve 3. kapitole se dostáváme k prvnímu problému, a tím jsou potíže s multikolinearitou (tedy s jistým druhem závislosti mezi proměnnými). Podrobněji jsme rozebrali metodu hřebenové regrese, která je vhodná na modely vykazující multikolinearitu. Hřebenovou regresi jsme aplikovali na již dlouho známá Haldova data. Zjistili jsme, že regresory jsou lineárně závislé, a proto jsme se rozhodli na ně metodu použít. V Matlabu jsme odhadli hřebenový parametr a pomocí něj jsme pak vypočetli odhady parametrů.

Heteroskedasticita znamená, že rozptyly nejsou shodné, což je opět problém. Existuje velké množství testů na rozpoznání heteroskedasticity. Ty dělíme na konstruktivní a nekonstruktivní. V této části není zmíněn žádný příklad, ale v příloze je uveden jeden z testů naprogramovaný v Matlabu, který nám pomůže heteroskedasticitu identifikovat.

O problém odlehlých pozorování se vědci zajímali již v historii. V kapitole, která se jimi zabývá, byly zmíněny základní pojmy a bylo poukázáno na to, jak

pomocí projekční matice **H** odlehlé hodnoty v datech nalézt.

Přesnost určených odhadů je ovlivněna volbou bodů, ve kterých je prováděno měření. Pokud tedy zvolíme špatné body, tak se může stát, že odhady nebudou vhodné na další použití. Tímto zkoumáním se zabývá design experimentu. Na příkladu je ukázána D – optimalita, která nám vhodně zvolí body, ve kterých máme měřit, aby odhady byly co nejlepší.

Kapitola 7 se zabývala ortogonální regresí. Ta je odlišná od klasické MNČ tím, že minimalizuje projekce kolmé na regresní přímku a ne součet druhých mocnin projekcí kolmých na osu x. Na příkladu jsme užili model s podmínkou typu II, který není příliš obvyklý a užívaný, ale i přesto jsme ho řešili.

Poslední kapitola byla věnována problému linearizace. Linearizace je vlastně převedení nelineární funkce na lineární. Používali jsme data z chemických měření, na která jsme aplikovali dva regresní modely – model bez podmínek a model s podmínkami typu II. U modelu bez podmínek jsme se data snažili aproximovat nelineární funkcí. Stejná data jsme v druhém modelu prokládali dvěmi parabolami, přičemž jsme měli soustavu nelineárních podmínek. V porovnání vyšly oba modely relativně stejně, a proto je složitější vybrat ten lepší.

Všechny výpočty jsem dělala pomocí programu Matlab, statistického programu, s jehož pomocí lze řešit složité matematické příklady. Celý text je napsán v typografickém programu TEX. Užití zmíněných programů mi velmi pomohlo zdokonalit si mé znalosti s jejich použitím.

Při psaní diplomové práce jsem si prohloubila své vědomosti ohledně různých statistických metod a vyzkoušela si práci s velkým množstvím literatury, což bylo pro mě hodnotným přínosem. Věřím, že moje práce bude, ať již mnou nebo někým jiným, v budoucnu využita jako vhodná pomůcka při nesnázích v regresi.

Příloha 1

Boškovičova metoda zpracovaná v programu Matlab 7

```
x=[0,0.2987,0.4648,0.5762,0.8386]
y=[56751,57037,56979,57074,57422]
plot(x,y,'o','LineWidth',2,'MarkerEdgeColor','k',...
'MarkerFaceColor','y','MarkerSize',10)
xlabel('Zeměpisná šířka','Fontsize',12)
ylabel('Délka jednoho stupně','Fontsize',12)
hold on

xprumer=sum(x)/5
yprumer=sum(y)/5
X=x-xprumer
Y=y-yprumer
for i=1:1:5
beta(i)=(y(i)-yprumer)/(x(i)-xprumer)
end

K5=abs(Y(1)-beta(5)*(X(1)))+abs(Y(2)-beta(5)*(X(2)))+abs(Y(3)-...
beta(5)*(X(3)))+abs(Y(4)-beta(5)*(X(4)))+abs(Y(5)-beta(5)*(X(5)))
K1=abs(Y(1)-beta(1)*(X(1)))+abs(Y(2)-beta(1)*(X(2)))+abs(Y(3)-...
beta(1)*(X(3)))+abs(Y(4)-beta(1)*(X(4)))+abs(Y(5)-beta(1)*(X(5)))
K4=abs(Y(1)-beta(4)*(X(1)))+abs(Y(2)-beta(4)*(X(2)))+abs(Y(3)-...
beta(4)*(X(3)))+abs(Y(4)-beta(4)*(X(4)))+abs(Y(5)-beta(4)*(X(5)))
K2=abs(Y(1)-beta(2)*(X(1)))+abs(Y(2)-beta(2)*(X(2)))+abs(Y(3)-...
beta(2)*(X(3)))+abs(Y(4)-beta(2)*(X(4)))+abs(Y(5)-beta(2)*(X(5)))
K3=abs(Y(1)-beta(3)*(X(1)))+abs(Y(2)-beta(3)*(X(2)))+abs(Y(3)-...
beta(3)*(X(3)))+abs(Y(4)-beta(3)*(X(4)))+abs(Y(5)-beta(3)*(X(5)))
```

```
primkax=[x(1),xprumer,0.937861271676301]
primkay=[y(1),yprumer,57400]
plot(primkax,primkay,'LineWidth',2)
hold on
primkax=[x(2),xprumer]
primkay=[y(2),yprumer]
plot(primkax,primkay,'r')
hold on
primkax=[x(3),xprumer]
primkay=[y(3),yprumer]
plot(primkax,primkay,'r')
hold on
primkax=[x(4),xprumer]
primkay=[y(4),yprumer]
plot(primkax,primkay,'r')
hold on
primkax=[x(5),xprumer]
primkay=[y(5),yprumer]
plot(primkax,primkay,'r')
hold on
yodhad=[56751+x*692]
yodhad2=yodhad'
rscboskovic=sum(((y'-yodhad2)'*(y'-yodhad2)))
L=sum(abs(y'-yodhad2))

gtext('Quito','Fontsize',12)
gtext('Mys Dobré Naděje','Fontsize',12)
gtext('Řím','Fontsize',12)
gtext('Paříž','Fontsize',12)
gtext('Laponsko','Fontsize',12)
```

Příloha 2

Lambertova metoda zpracovaná v programu Matlab 7

```
x=[0,0.2987,0.4648,0.5762,0.8386]
y=[56751,57037,56979,57074,57422]
plot(x,y,'o','LineWidth',2,'MarkerEdgeColor','k',...
'MarkerFaceColor','y','MarkerSize',10)
xlabel('Zeměpisná šířka','Fontsize',12)
ylabel('Délka jednoho stupně','Fontsize',12)
hold on

xprumer=sum(x)/5
yprumer=sum(y)/5
x1prumer=(x(1)+x(2))/2
y1prumer=(y(1)+y(2))/2
x2prumer=(x(3)+x(4)+x(5))/3
y2prumer=(y(3)+y(4)+y(5))/3
beta=(y2prumer-y1prumer)/(x2prumer-x1prumer)
alfa=y1prumer-beta*x1prumer
yodhad=alfa+X(:,2)*beta
L=sum(abs(y'-yodhad))
rsclambert=sum(((y'-yodhad)'*(y'-yodhad)))

primkax=[x1prumer,xprumer,x2prumer]
primkay=[y1prumer,yprumer,y2prumer]
plot(primkax,primkay,'b','LineWidth',2)
hold on

x1aprumer=(x(1)+x(2)+x(3))/3
y1aprumer=(y(1)+y(2)+y(3))/3
```

```
x2aprumer=(x(4)+x(5))/2
y2aprumer=(y(4)+y(5))/2
beta=(y2aprumer-y1aprumer)/(x2aprumer-x1aprumer)
alfa=y1aprumer-beta*x1aprumer
yodhad=alfa+X(:,2)*beta
L2=sum(abs(y'-yodhad))
rsclambert2=sum(((y'-yodhad)'*(y'-yodhad)))

primkax=[x1aprumer,xprumer,x2aprumer]
primkay=[y1aprumer,yprumer,y2aprumer]
plot(primkax,primkay,'r','LineWidth',2)
hold on

gtext('Quito','Fontsize',12)
gtext('Mys Dobré Naděje','Fontsize',12)
gtext('Řím','Fontsize',12)
gtext('Paříž','Fontsize',12)
gtext('Laponsko','Fontsize',12)
```

Příloha 3

Příklad na základní poznatky z regrese zpracované v programu Matlab 7

```
Y=[1.1;0.8;1.2;4.8;13.2;25.1;24.8;24.8;41.5;61.4;85.1;112.70;...
113.10;112.90;144.50;181.30;221.50;221.40;220.10]
X=[1,0,0;1,0,0;1,0,0;1,1,1;1,2,4;1,3,9;1,3,9;1,3,9;1,4,16;1,5,...
25;1,6,36;1,7,49;1,7,49;1,7,49;1,8,64;1,9,81;1,10,100;1,10,100;...
1,10,100]

%odhady parametrů%
V=eye(19)
betaodhad=inv(X'*inv(V)*X)*X'*inv(V)*Y
sigmaodhad=((Y-X*betaodhad)'*V^(-1)*(Y-X*betaodhad))/(19-3)
varbetaodhad=sigmaodhad*inv(X'*inv(V)*X)
varsigmaodhad=2*sigmaodhad^2/(19-3)

%oblasti spolehlivosti%
X'*inv(V)*X
dolni=sigmaodhad*(19-3)/28.8
horni=sigmaodhad*(19-3)/6.91
inv(X'*inv(V)*X)
c1=[1;0;0]
c2=[0;1;0]
c3=[0;0;1]
schefedolbeta1=betaodhad(1)-0.1458^(1/2)*9.6^(1/2)*...
(c1'*inv(X'*inv(V)*X)*c1)^(1/2)
schefehorbeta1=betaodhad(1)+0.1458^(1/2)*9.6^(1/2)*...
(c1'*inv(X'*inv(V)*X)*c1)^(1/2)
schefedolbeta2=betaodhad(2)-0.1458^(1/2)*9.6^(1/2)*...
(c2'*inv(X'*inv(V)*X)*c2)^(1/2)
```

```
schefehorbeta2=betaodhad(2)+0.1458^(1/2)*9.6^(1/2)*...
(c2'*inv(X'*inv(V)*X)*c2)^(1/2)
schefedolbeta3=betaodhad(3)-0.1458^(1/2)*9.6^(1/2)*...
(c3'*inv(X'*inv(V)*X)*c3)^(1/2)
schefehorbeta3=betaodhad(3)+0.1458^(1/2)*9.6^(1/2)*...
(c3'*inv(X'*inv(V)*X)*c3)^(1/2)

%testování hypotéz%
H=[0,0,1]
h=1
test=(H*betaodhad+h)'*inv(H*inv(X'*inv(V)*X)*H')*(H*betaodhad+h)
T=(c3'*1.9978)/(0.1458^(1/2)*(c3'*inv(X'*inv(V)*X)*c3))

%silofunkce%
silofunkce1=1-cdf('nct',1.74,17,0)
silofunkce2=1-cdf('nct',1.74,17,0.0126)
silofunkce3=1-cdf('nct',1.74,17,1.2580)
silofunkce4=1-cdf('nct',1.74,17,1.8116)
silofunkce5=1-cdf('nct',1.74,17,2.8306)
silofunkce6=1-cdf('nct',1.74,17,3.6357)
silofunkce7=1-cdf('nct',1.74,17,5.0322)
silofunkce8=1-cdf('nct',1.74,17,5.5480)
silofunkce9=1-cdf('nct',1.74,17,6.0889)
```

Příloha 4

Hřebenová regrese užitá na Haldova data v programu Matlab 7

```
load data2.txt %načteme data%
data=data2
Y=data(:,5)
X1=data(:,1:4)
n=length(X1)
vektor1=ones(n,1)
X=[vektor1,X1]
(X'*X)
v=eig(X'*X) %určíme vlastní čísla%
%otestujeme multikolinearitu%
indexpodm=sqrt(max(v)/min(v))  %spočteme index podmíněnosti%
korkoef12=corrcoef(X(:,2),X(:,3)) %korelační koeficienty%
korkoef13=corrcoef(X(:,2),X(:,4))
korkoef14=corrcoef(X(:,2),X(:,5))
korkoef23=corrcoef(X(:,3),X(:,4))
korkoef24=corrcoef(X(:,3),X(:,5))
korkoef34=corrcoef(X(:,4),X(:,5))

betaodhad=inv(X'*X)*X'*Y %vypočteme klasický MNČ odhad%
m=length(betaodhad)
V=eye(n)
I=eye(m)
%vypočteme odhad rozptylu%
sigmaodhad=((Y-X*betaodhad)'*V^(-1)*(Y-X*betaodhad))/(n-m-1)

delta=5 %zvolíme si hřebenový parametr delta%
%vypočteme hřebenový MNČ odhad parametrů beta%
```

```
betaodhaddelta=inv(X'*X+delta*I)*X'*Y
%vypočteme odhad rozptylu s hřebenovým parametrem%
sigmaodhaddelta=((Y-X*betaodhaddelta)'*V^(-1)*(Y-X*...
betaodhaddelta))/(n-m-1)

v=eig(X'*X) %určíme vlastní čísla a vlastní vektory%
[vl_vektory vl_cisla]=eig(X'*X)
K=vl_vektory
c=K'*betaodhad
Yodhad=betaodhad(1)+betaodhad(2)*X(:,2)+betaodhad(3)*X(:,3)+...
betaodhad(4)*X(:,4)+betaodhad(5)*X(:,5)
rsc=sum(((Y-Yodhad)'*(Y-Yodhad)))
senadruhou=rsc/(n-m)

%možnost odvození delta%
deltaopt=(4*senadruhou)/sum(betaodhad.^2)
betaodhaddeltanew=(inv(X'*X+deltaopt*I)*X'*Y)
%měl by nám vyjít odhad rozptylu menší než u odhadnutého delta%
sigmaodhaddeltanew=roundn(((Y-X*betaodhaddeltanew)'*V^(-1)*...
(Y-X*betaodhaddeltanew))/(n-m) ,-8)

%další možnost odvození delta%
deltaopt1=(4*senadruhou)/sum((v(:)).*(c(:)).^2)
betaodhaddeltanew1=(inv(X'*X+deltaopt1(1,:)*I)*X'*Y)
%měl by nám vyjít odhad rozptylu menší než u odhadnutého delta%
sigmaodhaddeltanew1=roundn(((Y-X*betaodhaddeltanew1)'*V^(-1)*...
(Y-X*betaodhaddeltanew1))/(n-m) ,-8)

%hřebenová stopa%
x1=X1(:,1)
```

```
x2=X1(:,2)
x3=X1(:,3)
x4=X1(:,4)
X = [x1 x2 x3 x4];
D = x2fx(X,'interaction');
D(:,1) = [];
k~= 0:1e-5:5e-3;
b = ridge(Y,D,k);
figure
plot(k,b,'LineWidth',2)
ylim([-100 100])
grid on
xlabel('Hřebenový parametr delta')
ylabel('Odhady parametrů beta v~kanonickém tvaru modelu')
legend('x1','x2','x3','x4','x1x2','x1x3','x1x4','x2x3',...
'x2x4','x3x4')
```

Příloha 5

Goldfeld – Quandtův test naprogramovaný v Matlabu 7

```
load data2.txt %načteme data%
data=data2
y=data(:,1) %závisle proměnná%
x=data(:,2) %nezávisle proměnná%
y=[y]
x=[x]
plot(y,x,'o') %vykreslíme data%
n=length(x)
xprumer=sum(x)/n %vypočteme průměr z x%
sigma=[1/n*(x-xprumer).^2] %rozptyly%
r=ceil(n/4) %kolik prostředních hodnot budeme vypouštět%
A=[sigma,y,x]
%setřídění matice%
sortA=sortrows(A,-1)
x=sortA(:,3)
y=sortA(:,2)
jedna=ones((n-r)/2,1) %sloupcový vektor, složen z n-r/2 číslic 1%
vyberx1=x(1:(n-r)/2,:)
X1=[jedna,vyberx1]  %matice X1, vybereme prvních n-r/2 členů%
Y1=y(1:(n-r)/2,:) %odpovídajících prvních n-r/2 y%
betaodhad1=(inv(X1'*X1)*X1'*Y1)'%odhad parametru%
yodhad1=betaodhad1(:,1)+vyberx1*betaodhad1(:,2) %odhad y%
rsc1=sum(((Y1-yodhad1)'*(Y1-yodhad1))) %spočteme 1.RSČ%
vyberx2=x(((n+r)/2)+1:n,:)
X2=[jedna,vyberx2] %matice X2, vybereme posledních n-r/2 členů%
Y2=y(((n+r)/2)+1:n,:)
betaodhad2=(inv(X2'*X2)*X2'*Y2)'
```

```
yodhad2=betaodhad2(:,1)+vyberx2*betaodhad2(:,2)
rsc2=sum(((Y2-yodhad2)'*(Y2-yodhad2)))%spočteme 2.RSČ%
if rsc1>rsc2
  F=rsc1/rsc2 %testovací statistika%
else F=rsc2/rsc1
end
kvan=icdf('f',0.95,(n-r)/2,(n-r)/2)
if F>kvan
    vysledek='heteroskedasticita'
else vysledek='homoskedasticita'
end
```

Příloha 6

Model s podmínkou typu II naprogramovaný v Matlabu 7

```
load datasouckova1.txt %načteme data%
data=datasouckova1
x=data(:,1) %nezávisle proměnná%
y=data(:,2) %závisle proměnná%
n=length(x)
Y=[y]
X=[x]
plot(X,Y,'o') %vykreslíme data%
hold on
s=11
for k=1:1:s %matice plánu pro prvních s dat%
F1(k,:)=[1,x(k),x(k)*x(k)];
end
Y1=Y(1:s,:)
X1=X(1:s,:)
betaodhad1=inv(F1'*F1)*F1'*Y1
Yodhad1=betaodhad1(1,:)+F1(:,2)*betaodhad1(2,:)+...
F1(:,3)*betaodhad1(3,:)
plot(X1,Yodhad1)
hold on

Y2=Y(s+1:n,:)
X2=X(s+1:n,:)
for h=s+1:n %matice plánu pro další data%
G(h,:)=[1,x(h),x(h)*x(h)];
end
F2=G(s+1:n,:)
```

```
betaodhad2=inv(F2'*F2)*F2'*Y2
Yodhad2=betaodhad2(1,:)+F2(:,2)*betaodhad2(2,:)+...
F2(:,3)*betaodhad2(3,:)
plot(X2,Yodhad2)
hold on

theta1=betaodhad1(1,:) %počáteční odhady parametrů%
theta2=betaodhad1(2,:)
theta3=betaodhad1(3,:)
theta4=betaodhad2(1,:)
theta5=betaodhad2(2,:)
theta6=betaodhad2(3,:)
theta7=6.8
thetaodhad=[theta1,theta2,theta3,theta4,theta5,theta6,theta7]'
betaodhad=[Y]

B1=-1*eye(n)
B2=(zeros(1,n))
B=[B1;B2] %zlinearizovaná matice B%
Sigma=diag(ones(n,1)) %matice sigma%
for q=1:1:s %zlinearizovaná matice C%
Q(q,:)=[1,x(q),x(q)*x(q),0,0,0,0];
end
for w=s+1:n
W(w,:)=[0,0,0,1,x(w),x(w)*x(w),0] ;
end
E=W(s+1:n,:);
C1=[1,theta7,theta7^2,-1,-theta7,-theta7^2,theta2+...
2*theta3*theta7-theta5-2*theta6*theta7];
```

```
C=[Q;E;C1]
b=-B*betaodhad-C*thetaodhad
%odhady parametrů%
X=eye(n,n)
S=B*(X'*Sigma^(-1)*X)^(-1)*B'+C*C';
betaodhadodhad=betaodhad-((X'*Sigma^(-1)*X)^(-1)*B'*(S^(-1)-...
S^(-1)*C*(C'*S^(-1)*C)^(-1)*C'*S^(-1))*(b+B*betaodhad))
thetaodhadodhad=-(C'*S^(-1)*C)^(-1)*C'*S^(-1)*(b + B*betaodhad)

%nakreslení grafu odhadů%
x=data(:,1)
y=data(:,2)
Y=[y]
X=[x]
figure(3)
cas=[min(x):0.1:max(x)]';
for i=1:length(cas)
      if cas(i)<thetaodhadodhad(7)
    YY(i,1)=thetaodhadodhad(1)+thetaodhadodhad(2)*cas(i)+...
    thetaodhadodhad(3)*cas(i)^2;
    else
    YY(i,1)=thetaodhadodhad(4)+thetaodhadodhad(5)*cas(i)+...
    thetaodhadodhad(6)*cas(i)^2;
    end
end

plot(X,Y,'o','Color',[70/256, 130/256, 180/256],'MarkerSize',8,...
'LineWidth',1.8,'MarkerFaceColor',[0/256, 0/256, 128/256])
hold on
plot(cas,YY,'-','Color',[255 20 147]/256,'LineWidth',3.4,...
```

```
'MarkerSize',8);xlabel('Koncentrace [mmol/L]','Fontsize',...
18,'Fontweight','Demi','Fontname','Palatino Linotype')
ylabel('Potenciál [mV]','Fontsize',18,'Fontweight','Demi',...
'Fontname','Palatino Linotype')
set(gca,'Fontsize',18,'Fontweight','Demi','Fontname',...
'Palatino Linotype','Color',[255/256, 255/256, 249/256],...
'XColor',[0/256, 0/256, 139/256],'YColor',[0/256, 0/256, 139/256])
gtext('\theta_1+\theta_2x+\theta_3x^2','Fontsize',20,...
'Fontweight','Demi','Fontname','Palatino Linotype','Color',...
[0 104 139]/256)
gtext('\theta_4+\theta_5x+\theta_6x^2','Fontsize',20,...
'Fontweight','Demi','Fontname','Palatino Linotype','Color'...
,[0 104 139]/256)
gtext('\theta_7','Fontsize',20,'Fontweight','Demi','Fontname',...
'Palatino Linotype','Color',[0 104 139]/256)
cas(i)
Yodhad1=thetaodhadodhad(1)+thetaodhadodhad(2)*F1(:,2)+...
thetaodhadodhad(3)*F1(:,3)
Yodhad2=thetaodhadodhad(4)+thetaodhadodhad(5)*F2(:,2)+...
thetaodhadodhad(6)*F2(:,3)
Yodhad=[Yodhad1;Yodhad2]
mean(Y)
Se=sum((Y-Yodhad).^2)
SY=sum((Yodhad-mean(Y)).^2)
R=1-Se/SY
sigmanadruhou=Se/(n-2)
gtext('index determinace R=0.99507','Fontsize',14,'Fontweight',...
'Demi','Fontname','Palatino Linotype','Color',[0 100 0]/256)
gtext('\sigma^2=5.9230^2','Fontsize',14,'Fontweight','Demi',...
'Fontname','Palatino Linotype','Color',[0 100 0]/256)
```

Literatura

[1] Hebák, P., Svobodová, A.: Regrese – II.část. Nakladatelství Oeconomica, Praha, 2002.

[2] Anděl, J., Zvára, K.: Ekonomicko – matematický obzor. Československá akademie věd, Praha, 1985, 444 – 456.

[3] Cipra, T.: Finanční ekonometrie. Ekopress, Praha, 2008.

[4] Zvára, K.: Regrese. Matfyzpress, Praha, 2008.

[5] Kubáček, L., Kubáčková, L.: Statistika a metrologie. UP Olomouc, Olomouc, 2000.

[6] Hebák, P.: Příklady z regrese. SPN, Praha, 1984.

[7] Hoerl, A., Kennard, W., Kennard H: Ridge regression biased estimation for nonorthogonal problems. American Statistical Assotiation and American Society for quality, 2000, 80 – 86.

[8] Hebák, P., Hustopecký, J.: Vícerozměrné statistické metody s aplikacemi. SNTL/Alfa, Praha, 1987.

[9] Zvára, K.: Regresní analýza. Academia, Praha, 1989.

[10] Hald, A.: A History of Matematical Statistic (from 1750 to 1930). A Wiley interscience Publication, 2000.

[11] Stigler, S. M.: History of Statistic – The Measurement of Uncertainty before 1900, The Belknap Press of Harvard University Press, Cambridge, Messachusetts and London, 1986.

[12] Kunderová, P.: Základy pravděpodobnosti a matematické statistiky, Olomouc, 2004.

[13] Anděl, J.: Statistické modely, Matfyzpress, Praha, 2007.

[14] Spohnerová, K.: Řešení neřešitelných rovnic: Historie a současnost, Bakalářská práce, UP Olomouc, Olomouc, 2008.

[15] Montgomery, D. C., Peck, E. A., Vining, G. G.: Introduction to linear regression analysis, A John Wiley & sons, inc. publication, Hoboken, New Jersey, 2006.

[16] Hušek, R.: Základy ekonometrie, VŠE Praha, Praha, 1992.

[17] Kubáček, L., Tesaříková, L.: Weally nonlinear regression models, UP Olomouc, Olomouc, 2008.

[18] Marek, J., Tuček, P.: Statistické algoritmy pro aproximaci Lorentzovy funkce, KMAaAM PřF UP v Olomouci, Olomouc, Preprint 9/2009. http://mant.upol.cz/en/preprinty.asp

[19] http://www.snc.cz/old/honza/data/zek/heteroskedasticita.ppt#1636,1,ZÁKLADY EKONOMETRIE Heteroskedasticita [online 19. 10. 2010]

[20] http://blog.i-page.net/wp-content/uploads/matematicka-statistika-30-10-2006-pred-5.pdf [online 19. 10. 2010]

[21] http://www.mti.tul.cz/files/zsr/Linearizace.pdf [online 21. 10. 2010]

[22] http://www.wikipedia.cz [online 16. 2. 2011]

[23] http://meloun.upce.cz/docs/books/ucebnice-sken.pdf [online 7. 3. 2011]

[24] http://www1.lf1.cuni.cz/ ldohna/publik/Kap_8_chybodlhodnoty.pdf [online 9. 3. 2011]

[25] http://petrsoukal.profitux.cz/diplomka_heteroskedasticita1999.pdf [online 12. 3. 2011]

[26] http://is.muni.cz/th/44303/prif_d/dizertace.doc [online 21. 3. 2011]

[27] http://is.muni.cz/th/78391/prif_m/Aplikace_GLM_modelu_v_provozni_praxi.doc [online 21. 3. 2011]