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

BAKALÁŘSKÁ PRÁCE

Metody odhadů regresních parametrů

Vedoucí bakalářské práce:
doc. RNDr. Eva Fišerová, Ph.D.
Rok odevzdání: 2012

Vypracovala:
Kristýna Vaňkátová
AST, III. ročník

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D. s použitím uvedené literatury.

V Olomouci dne 20. dubna 2012

Poděkování

Na tomto místě bych chtěla poděkovat především své vedoucí bakalářské práce doc. RNDr. Evě Fišerové Ph.D. za cenné připomínky a odborné rady, kterými mi pomohla dovést tuto práci ke zdárnému konci. Také bych ráda poděkovala své rodině a přátelům, že mě po celou dobu studia podporovali.

Obsah

Úvod

Cílem této práce je popsat různé metody pro určení odhadů regresních parametrů a následné použití těchto metod při řešení konkrétního příkladu. Práce je založena na regresní analýze, která je snad nejpoužívanější statistickou metodou, protože je jí potřeba v kterémkoli odvětví, kde nás zajímá vyjádření vztahu dvou a více závislých proměnných.

První kapitola bude obsahovat základní informace o regresním modelu. Budou uvedeny předpoklady regresního modelu, jež se především v následujících dvou kapitolách stanou stěžejními. Bude představen lineární regresní model a speciálně klasický lineární model, jež bude uvažován až do konce této práce.

Druhá kapitola bude věnována metodě nejmenších čtverců, jakožto nejznámější metodě pro výpočet regresních koeficientů. Tato metoda je vyučovaná ve většině statistiských kurzů a je též hojně užívaná pro své odhady s výhodnými statistiskými vlastnostmi. Objasněna bude nejen podstata metody nejmenších čtverců, ale budou též uvedeny důvody k jejímu použití a situace, kde je její využití vhodné. Pro lepší pochopení principu metody nejmenších čtverců bude text obohacen o geometrickou interpretaci celé metody.

Další stručná kapitola bude věnována úvodu do problematiky metody maximální věrohodnosti. Obě již zmíněné metody odhadu regresních koeficientů jsou výhodné v případě, že jsou rezidua normálně rozdělena. Pokud tato základní podmínka není dodržena, metoda nejmenších čtverců i metoda maximální věrohodnosti podávají velmi zkreslené odhady parametrů, můžeme tedy říci, že odhady takto pořízené jsou velmi citlivé na porušení předpokladu normality.

Výše uvedený problém je řešitelný použitím robustní regrese, která narozdíl od metody nejmenších čtverců nepředpokládá normální rozdělení reziduí a není tedy náchylná na výskyt odlehlých pozorování či na rezidua, jež mají rozdělení s dlouhými chvosty. Tomuto tématu bude věnována další část práce, kde bude stručně představeno několik základních robustních metod pro odhad regresních parametrů.

V závěrečné části práce budou všechny metody demonstrovány na nasimulovaných příkladech, které by měly dostatečně vystihnout všechny uvedené případy kontaminace modelu a zároveň by z nich mělo být patrné, jaké odlišnosti jednotlivé regresní metody vykazují.

1. Regresní analýza

Slovo regrese by se dalo přeložit jako úpadek či zpětný vývoj. Na první pohled zde není žádná souvislost mezi významem tohoto slova a pricipem regresní analýzy. Pro lepší pochopení je třeba se podívat do historie, přesněji do období mezi roky 1877 a 1885, kdy anglický vědec Francis Galton předložil závěry své práce veřejnosti. Galton v nich regresí nazval tendenci návratu následující generace směrem k průměru, k čemuž došel na základě předchozího zkoumání výšky otce a syna. Pozorováním a analýzou údajů totiž dospěl k názoru, že malí otcové mají v průměru vyšší syny, než jsou oni sami, a naopak vysocí otcové mají v průměru menší syny, než jsou oni samotní. Odtud tedy regrese = krok zpět. [9]

Ve skutečnosti je regresní analýza souborem statistických metod schopných odhalit a kvantifikovat funkční vztahy mezi proměnnými. Regresní analýza patří k základním a nejčastějším metodám statistické analýzy vztahů. Je to metoda hojně užívaná v praxi a je obsažena ve většině standardních statistických programovacích systémů. Navíc ji lze použív v kterékoli vědní oblasti, kde sledujeme závislost mezi veličinami.[5, 8, 16]

Konkrétně při regresním zkoumání sledujeme závislost výstupní (též vysvětlované, náhodné) veličiny Y na vstupních nezávislých (vysvětlujících, nenáhodných) veličinách $x_1, \ldots, x_k$, které jsou v regresních úvahách nastavované. Dále proto mluvíme o klasických regresních modelech, jež jsou odvozeny z praktických výzkumů, kde máme možnost při plánovaném experimentu nastavit vstupní proměnné a následně zaznamenáváme změny u proměnné Y. Regresní modely popisují námi zkoumané závislosti a jejich součástí je regresní funkce, u níž se budeme zabývat odhadem optimálních regresních parametrů. [9]

V následujícím textu se budeme setkávat s regresním modelem ve formě

$$\text{závislá proměnná} = \text{regresní funkce} + \text{náhodná chyba} ,$$

kde regresní funkce bude známá až na parametry. Náhodná chyba je logický důsledek toho, že při experimentu máme k dispozici pouze napozorované hodnoty

veličiny Y, nikoli hodnoty bezchybné. Rozdíl mezi těmito hodnotami nazýváme chybami měření. [7]

1.1. Regresní model

Regresní analýza umožňuje jejímu uživateli popsat závislost výstupní veličiny Y na vstupních nezávislých veličinách pomocí matematického vzorce, který popisuje Y jako funkci proměnných $x_1, \ldots, x_k$. Tato regresní funkce, kterou budeme značit $f(x, \beta_1, \beta_2, \ldots, \beta_k)$, je definovaná jako podmíněná střední hodnota výstupní náhodné veličiny vzhledem k různým kombinacím hodnot vstupních nenáhodných veličin. Při dané hodnotě $\mathbf{x}$ tento vztah zapíšeme jako

$$Y = f(\mathbf{x}, \beta_1, \beta_2, \ldots, \beta_k) = E(Y|\mathbf{x}),$$

přičemž $\beta_1, \beta_2, \ldots, \beta_k$, $k \geq 1$, jsou neznáme konstanty, na kterých funkce f závisí. Konstanty $\beta_1, \beta_2, \ldots, \beta_k$ se nazývají regresní parametry. Regresí tedy rozumíme závislost mezi střední hodnotou náhodné veličiny Y a proměnnou $\mathbf{x}$. [9]

V této práci se omezíme na regresní analýzu opřenou o lineární regresní model, což znamená, že funkce f je sestavená jako lineární kombinace k funkcí

$$Y = \beta_1 f_1(\mathbf{x}) + \ldots + \beta_k f_k(\mathbf{x}),$$

přičemž jednotlivé funkce $f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ nemusí být lineární. Linearitou v tomto kontextu myslíme závislost funkce $f(\mathbf{x})$ na koeficientech $\beta_1, \ldots, \beta_k$. Tyto koeficienty navíc rozšíříme o absolutní člen β_0, jehož zařazení do rovnice je matematický výhodné a vyplývá z existence nezařazených a neuvažovaných vlivů.[9, 8]

Pokud navíc mluvíme o zcela lineárním modelu, který je oprávněný v případě vícerozměrného normálního rozdělení uvažovaných náhodných veličin, můžeme vyjádřit regresní funkci jako rovnici nadroviny

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon.$$

Ve zcela lineárním modelu se předpokládá součtový vliv všech činitelů. β_0 je absolutní člen a $\beta_1, \beta_2, \ldots, \beta_k$ jsou dílčí regresní koeficienty. Například parametr β_1 je interpretován jako očekávaná změna veličiny Y při jednotkovém růstu veličiny x_1 a při konstantním vlivu ostatních proměnných $x_2, x_3, \ldots, x_k$.

Speciálním případem pro jednu vysvětlující proměnnou je model regresní přímky $Y = \beta_0 + \beta_1 x + \varepsilon$, stejně tak model regresní roviny $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ je speciálním případem pro dvě vysvětlující proměnné. Obecně potom můžeme jako regresní funkci ve zcela lineárním modelu označit nadrovinu s k vysvětlujícími proměnnými.[9]

Podívejme se na tento model z více praktického hlediska a představme si, že experiment provedeme při n různých nastavených proměnných $x_1, x_2, \ldots, x_k$. Potom dostaneme soustavu n lineárních rovnic

$$Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n$$

Tyto rovnice lze též vyjádřit maticově, pokud uvažujme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ a nestochastickou matici čísel $\mathbf{X}_{n\times(k+1)}$. Označme $k+1 = p$ a navíc uvažujme plnou hodnost matice $\mathbf{X}$, $h(\mathbf{X}) = p$, a $n \geq p$. Tomu odpovídá formulace, že žádné dva sloupce této matice x_j, x_k pro $j \neq k$ nejsou kolineární, tj. rovnoběžné vektory. Tím je zaručeno, že matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula.[9]

Předpokládejme, že se $\mathbf{Y}$ řídí lineárním modelem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^{\mathrm{T}}$ je vektor neznámých parametrů a $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ je vektor náhodných chyb. Pro názornost ještě celý vztah přepíšeme následujícím způsobem

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad .$$

Následující dva odstavce obsahují další důležité základní předpoklady lineárního modelu, na které se budeme odkazovat v dalším textu, jakožto na důležité podmínky pro výpočet regresních koeficientů metodou nejmenších čtverců či metodou maximální věrohodnosti. Na začátek uveďme, že vektor náhodných chyb splňuje podmínky

$$E(\boldsymbol{\varepsilon}) = \mathbf{o}, \qquad var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

$\mathbf{I}$ značí jednotkovou matici, $\mathbf{o}$ je vektor nul a σ je neznámý parametr, pro který platí $\sigma \geq 0$. Předpoklad $E(\boldsymbol{\varepsilon}) = \mathbf{o}$ odpovídá skutečnosti, že pozorování vektoru $\mathbf{Y}$ nejsou zatížena systematickými chybami a jejich rozdělení pravděpodobnosti je symetrické kolem nuly. Druhý vztah $var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ ukazuje, že rozptyl náhodné chyby je konstantní pro každé měření, což nazýváme též homoskedasticita rozptylu. Na základě této informace můžeme říci, že měření jednotlivých složek vektoru $\mathbf{Y}$ jsou prováděna se stejnou přesností. Jelikož je matice rozptylu diagonální, chyby měření různých složek vektoru $\mathbf{Y}$ jsou nekorelované, $\text{cov}(\varepsilon_i; \varepsilon_j) = 0$, $i \neq j$. [1, 9, 29]

Tyto vlastnosti náhodných chyb bývají v anglicky psané literatuře zabývající se matematickými problémy označovány jako White Noise (Bílý šum), lze tedy psát $\boldsymbol{\varepsilon} \sim \text{WN}(\mathbf{0}, \sigma^2\mathbf{I})$, což shrnuje všechny výše uvedené informace o střední hodnotě, rozptylu a kovarianci náhodných chyb. [28]

2. Metoda nejmenších čtverců

2.1. Historie

Metoda nejmenších čtverců je jednou z nejstarších metod moderní statistiky, a ačkoliv lze kořeny této metody nalézt i v řecké matematice, za prvního předchůdce moderní MNČ je většinou považován Galileo Galilei. Moderní přístup byl poprvé uveřejněn v roce 1805 francouzským matematikem Adrien-Marie Legendrem v jeho nyní již klasickém pojednání, avšak v moderní době se má za to, že metoda byla doopravdy vynalezena ještě o několik let dříve. Několik let po zveřejnění Legendreova pojednání totiž známý německý fyzik a matematik Carl Friedrich Gauss zpochybnil Legendrovo prvenství.

Carl Friedrich Gauss často nezveřejňoval své myšlenky, pokud měl dojem, že by mohly být kontroverzní nebo ještě nezralé. Stejně tomu bylo i s metodou nejmenších čtverců, kterou zmínil až v roce 1809 ve dvou ze svých prací o vesmírné mechanice. Podle těchto zmínek objevil metodu nejmenších čtverců již v roce 1795, kdy ji používal k odhadu dráhy asteroidu Ceres.

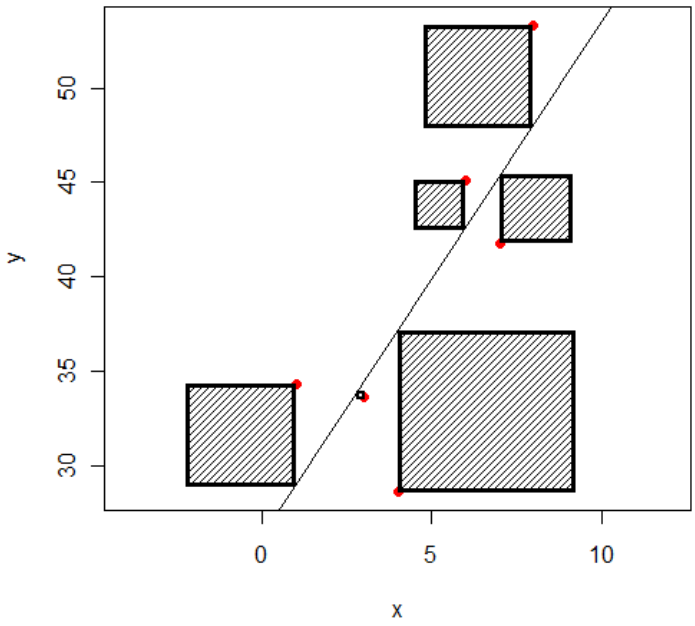
V roce 1822 byl Gauss schopen konstatovat, že v lineárním modelu, kde chyby mají nulovou střední hodnotu, jsou nekorelované a mají stejné rozptyly, je nejlepším lineárním nestranným odhadem koeficient pořízený metodou nejmenších čtverců. Tento výsledek je známý jako Gauss-Markovův teorém.

Následovaly poněkud nepříjemné spory o tom, kdo doopravdy přišel s myšlenkou metody nejmenších čtverců první, ty však nijak nesnížily popularitu této metody. Dalším sledováním metody nejmenších čtverců v rámci statistiky bychom došli až ke Galtonovi, který díky ní definoval korelaci a regresní analýzu, či Pearsonovi či Fisherovi.[9, 26]

2.2. Odhad regresních parametrů

Předpoklady o chybové složce uvedené v předchozí kapitole dovolují převést obtížný teoretický problém modelování závislosti na propracovanou statistickou

úlohu odhadu parametrů známého či předpokládaného pravděpodobnostního rozdělení. Touto úlohou je metoda nejmenších čtverců, která nám umožňuje nalézt vhodnou aproximační funkci pro dané empiricky zjištěné hodnoty tak, že je schopna vypočíst u lineární kombinace předem známých funkcí koeficienty těchto dílčích funkcí. Metoda nejmenších čtverců pracuje na zdánlivě jednoduchém principu, který spočívá v minimalizaci druhých mocnin odchylek. Graficky je tato metoda znázorněna na obrázku 1, kde jsou napozorované hodnoty $\mathbf{y}$ aproximovány přímkou.



Obr. 1: Přímková regrese

Označme $\mathbf{X}\widehat{\boldsymbol{\beta}}$ lineární aproximační funkci, kde $\mathbf{X}$ je známá matice rozměrů $n \times k$ a $\boldsymbol{\beta}$ je vektor $k \times 1$ neznámých parametrů. Vyrovnanými hodnotami nazveme vektor $\widehat{\mathbf{y}}$ zjištěný pomocí lineární aproximační funkce, tedy $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. Vektor $\widehat{\boldsymbol{\beta}}$ je odhad vektoru koeficientů $\boldsymbol{\beta}$, jenž vypočteme minimalizací součtu čtverců odchylek, což je vyjádřeno v následujícím vztahu

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta})^2$$

Tato kvadratická forma, nazývaná reziduální součet čtverců, může být upravena následujícím způsobem:

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \min_{\beta} (\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}). \tag{1}$$

Minimum tohoto výrazu najdeme tak, že první parciální derivace výrazu (1) podle složek vektoru $\boldsymbol{\beta}$ položíme rovny nule, čímž dostaneme soustavu p normálních rovnic o p neznámých ve tvaru

$$-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}.$$

Tento vztah upravíme:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y},$$

čímž se dostáváme k výslednému odhadu vektoru parametrů $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$. Tato hodnota je realizací odhadu (náhodného vektoru)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

Podle Gauss-Markovovy věty je odhad regresních koeficientů pořízený metodou nejmenších čtverců nejlepším nestranným lineárním odhadem (BLUE = Best Linear Unbiased Estimator) vektoru parametrů $\boldsymbol{\beta}$ [7, 9, 27]. Nejlepší ve smyslu, že má nejmenší rozptyl ze všech nestranných odhadů parametrů. Že je odhad nestranný, poznáme dle jeho střední hodnoty. Pokud uvažujeme odhad $\widehat{\theta}$ neznámého parametru θ, pro střední hodnotu takového odhadu platí [9]

$$E(\widehat{\theta}) = \theta.$$

Z informace o rozdělení vektoru náhodných chyb snadno odvodíme rozdělení vektoru $\mathbf{Y}$. Střední hodnotu a rozptyl jednoduše určíme jako

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta},$$

$$var(\mathbf{Y}) = var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = var(\mathbf{X}\boldsymbol{\beta}) + var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}.$$

Stejné odvození rozdělení lze provést pro odhad vektoru koeficientů $\widehat{\boldsymbol{\beta}}$

$$E(\widehat{\boldsymbol{\beta}}) = E((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

$$var(\widehat{\boldsymbol{\beta}}) = var((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\, var(\mathbf{Y})\, \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} =$$

$$= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

a vektor vyrovnaných hodnot $\widehat{\mathbf{Y}}$

$$E(\widehat{\mathbf{Y}}) = E(\mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}\, E(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta},$$

$$var(\widehat{\mathbf{Y}}) = var(\mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}\, var(\widehat{\boldsymbol{\beta}})\, \mathbf{X}^{\mathrm{T}} = \sigma^2\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}.$$

Pokud jsou navíc chyby normálně rozděleny, má normální rozdělení i vektor $\mathbf{Y}$, a tedy i odhad $\widehat{\boldsymbol{\beta}}$ a vyrovnané hodnoty $\widehat{\mathbf{Y}}$ a přitom platí $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$ a $\widehat{\mathbf{Y}} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})$. Střední hodnota rozdělení vektoru $\widehat{\boldsymbol{\beta}}$ potvrzuje, že jde o nestranný odhad.[1, 9]

Rozptyl σ^2 vetšinou neznáme a je proto nutné ho odhadnout. K tomu využijeme reziduální součet čtverců, který je dán součtem čtvercových reziduí

$$\mathrm{S}_R = \sum_{i=1}^{n}(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}})^2 = \sum_{i=1}^{n}(\mathbf{e}_i)^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{e}^{\mathrm{T}}\mathbf{e}.$$

Odhad σ^2 bude vypadat následujícím způsobem

$$\hat{\sigma}^2 = \frac{\mathrm{S}_R}{n-p}. \tag{2}$$

Náhodná veličina $\hat{\sigma}^2$ je nestranným odhadem parametru σ^2 a nazveme ji reziduální rozptyl. Nestrannost dokážeme na následujících řádcích pomocí projekční matice $\mathbf{H}_{n\times n}$, která je rovna

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

a díky níž můžeme definovat vektor vyrovnaných hodnot $\widehat{\mathbf{Y}}$ jako

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{H}\,\mathbf{Y}.$$

Dále můžeme pomocí matice $\mathbf{H}$ přepsat vztah pro výpočet reziduí

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y},$$

kde výraz v závorkách označíme $\mathbf{M}$ pro jednodušší práci v pozdějších výpočtech. Jak matice $\mathbf{M}$, tak matice $\mathbf{H}$ jsou symetrické a idempotentní a právě díky jejich idempotenci platí, že jejich hodnost je rovna jejich stopě. Proto je možné psát

$$h(\mathbf{H}) = Tr(\mathbf{H}) = Tr(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}) = Tr((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}) = Tr(\mathbf{I}_p) = p,$$

a také

$$h(\mathbf{M}) = Tr(\mathbf{M}) = Tr(\mathbf{I}_n - \mathbf{H}) = n - p.$$

Těchto poznatků můžeme využít k další změně zápisu vektoru reziduí

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{M}\mathbf{Y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} =$$

$$= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\boldsymbol{\varepsilon}$$

Následně můžeme vyjádřit reziduální součet čtverců jako kvadratickou formu

$$\mathrm{S}_R = \mathbf{e}^{\mathrm{T}}\mathbf{e} = \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}\,\boldsymbol{\varepsilon}.$$

Střední hodnota reziduálního součtu čtverců po drobných úpravách a s pomocí věty o střední hodnotě kvadratické formy [2] je rovna

$$E(\mathrm{S}_R) = E(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}\,\boldsymbol{\varepsilon}) = Tr(\mathbf{M}\,var(\boldsymbol{\varepsilon})) + E(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M})E(\boldsymbol{\varepsilon}) = Tr(\mathbf{M}\,\sigma^2\mathbf{I}) = \sigma^2(n-p),$$

díky čemuž pro reziduální rozptyl platí

$$E(\hat{\sigma}^2) = E\left(\frac{\mathrm{S}_R}{n-p}\right) = \frac{1}{n-p}\,E(\mathrm{S}_R) = \sigma^2.$$

Tím jsme dokázali nestrannost odhadu $\hat{\sigma}^2$. [10]

2.3. Přímková regrese

Klasickým a dobře známým příkladem lineárního modelu je jednoduchá lineární regrese, jinak řečeno obecná přímka. Mějme model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde matice plánu vypadá následovně

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Pokud navíc zapíšeme i podobu

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

a dále

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix},$$

není již problémem sestavit soustavu normálních rovnic

$$n\widehat{\beta}_0 + \sum_{i=1}^n x_i \widehat{\beta}_1 = \sum_{i=1}^n Y_i$$

$$\sum_{i=1}^n x_i \widehat{\beta}_0 + \sum_{i=1}^n x_i^2 \widehat{\beta}_1 = \sum_{i=1}^n x_i Y_i.$$

Jejím řešením dostaneme odhad vektoru regresních koeficientů $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^{\mathrm{T}}$, kde

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i x_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^n Y_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}.$$

Regresní koeficient $\widehat{\beta}_1$ je směrnicí přímky a koeficient $\widehat{\beta}_0$ udává posunutí přímky ve směru osy y. Odhad $\widehat{\beta}_1$ navíc udává průměrnou změnu závisle proměnné Y při jednotkové změně nezávisle proměnné x. Nabývá kladných hodnot, pokud je závislost mezi proměnnými přímá, a záporných hodnot, je-li nepřímá.

Tento model je zde zmíněn především proto, že je jednoduchý a dají se na něm lehce ukázat zákonitosti regresní analýzy. Stejně tak bude pro názornost aplikovaný v příkladové části. Dalším důvodem, proč zde model přímkové regrese představujeme, je grafická ukázka proložení dat funkcí a zvýraznění čtvercových reziduí. [1, 12, 22]

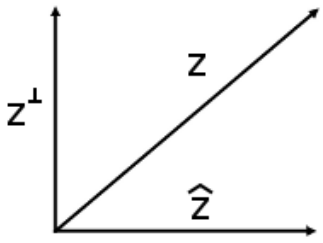
2.4. Geometrický přístup k metodě nejmenších čtverců

Metoda nejmenších čtverců může být vysvětlena také z pohledu geometrického jako kolmý průmět (ortogonální projekce) vektoru dat do prostoru vymezeného nezávislou proměnnou. [26]

2.4.1. Teoretický základ

Mějme dán vektorový prostor $\mathbf{U}$ a jeho podprostor $\mathbf{W}$. Podprostor $\mathbf{V}$ je ortogonálním doplňkem množiny $\mathbf{W}$ v $\mathbf{U}$, tvoří-li ho množina vektorů z $\mathbf{U}$ kolmých na $\mathbf{W}$, což vyjádříme jako $\mathbf{V} = \{\mathbf{z} \in \mathbf{U}; \forall \mathbf{x} \in \mathbf{W} : \mathbf{x}^{\mathrm{T}}\mathbf{z} = 0\}$.

Základní myšlenkou kolmého průmětu do podprostoru je fakt, že každý vektor $\mathbf{z} \in \mathbf{U}$ lze vyjádřit ve tvaru $\mathbf{z} = \widehat{\mathbf{z}} + \mathbf{z}^{\perp}$, kde $\widehat{\mathbf{z}} \in \mathbf{W}$, $\mathbf{z}^{\perp} \in \mathbf{V}$.



Obr. 2: Rozklad vektoru $\mathbf{z}$

Vektor $\widehat{\mathbf{z}}$ se nazývá kolmý průmět vektoru $\mathbf{z}$ do podprostoru $\mathbf{W}$ a vektor $\mathbf{z}^{\perp}$ nazýváme kolmice vektoru $\mathbf{z}$ na podprostor $\mathbf{W}$. Vzdálenost mezi podprostorem $\mathbf{W}$ a vektorem $\mathbf{z}$ je obecně dána nejmenší vzdáleností vektoru $\mathbf{z}$ od nějakého vektoru náležícího podprostoru $\mathbf{W}$. Ukazuje se, že tímto vektorem je právě kolmý průmět $\widehat{\mathbf{z}}$, délka kolmice $\mathbf{z}^{\perp}$ potom určuje vzdálenost vektoru $\mathbf{z}$ od podprostoru $\mathbf{W}$. Tuto vzdálenost označujeme $\rho(\mathbf{z}, \mathbf{W})$ a definujeme ji vztahem

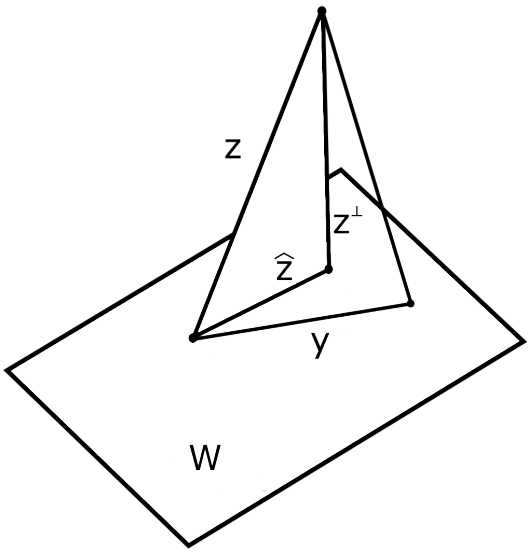
$$\rho(\mathbf{z}, \mathbf{W}) = \rho(\mathbf{z}, \widehat{\mathbf{z}}).$$

Pokud máme dán podprostor $\mathbf{W}$ a vektor $\mathbf{z} \in \mathbf{U}$, potom pro každý vektor $\mathbf{y} \in \mathbf{W}$ platí:

$$\rho(\mathbf{z}, \mathbf{y}) \geq \rho(\mathbf{z}, \widehat{\mathbf{z}}). \tag{3}$$

Rovnost ve výrazu (3) nastává pouze v případě, že $\mathbf{y} = \widehat{\mathbf{z}}$. Tím jsme pro vzdálenost vektoru od podprostoru získali vztah

$$\rho(\mathbf{z}, \mathbf{W}) = \min_{\mathbf{y} \in \mathbf{W}} \{\rho(\mathbf{z}, \mathbf{y})\}.$$



Obr. 3: Vzdálenost vektoru od podprostoru

Všechny pojmy jsou graficky znázorněny na obrázku 3. Vektor $\mathbf{y}$ byl zvolen libovolně a je jasně patrné, že jeho vzdálenost od vektoru $\mathbf{z}$ je větší, než tomu je u kolmého průmětu $\widehat{\mathbf{z}}$. Další podrobnosti k dané problematice lze nalézt např. v [13].

2.4.2. Geometrická interpretace

Po definování základních pojmů souvisejících s projekcí a vzdáleností vektoru od podprostoru je možné tyto znalosti využít pro vysvětlení metody nejmenších čtverců z geometrického hlediska. Jako v předchozích kapitolách budeme pracovat s maticí $\mathbf{X}$ známých konstant o n řádcích a p sloupcích, jejíž hodnost je p, přičemž platí $p \leq n$. Prostor vymezený nezávislou proměnnou je vlastně lineární obal p lineárně nezávislých sloupcových vektorů matice $\mathbf{X}$, značíme $\mathcal{M}(\mathbf{X})$:

$$\mathcal{M}(\mathbf{X}) = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \mathbf{X}\mathbf{t}, \mathbf{t} \in \mathbb{R}^p\}.$$

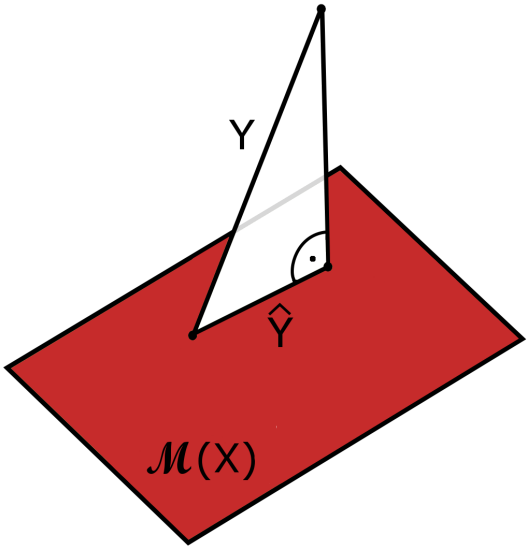
Princip kolmého průmětu spočívá v myšlence, že soustava lineárních rovnic daná maticí $\mathbf{X}$ a sloupcovým vektorem empirických hodnot $\mathbf{y}$, o neznámých $(b_1, b_2, \ldots, b_p)$, neboli

$$\begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

není kvůli náhodným chybám měření řešitelná. Jinak řečeno, pro vektor $\mathbf{y} \in \mathbb{R}^n$ platí $\mathbf{y} \notin \mathcal{M}(\mathbf{X})$. Vzhledem k tomuto faktu bude naším úkolem najít co nejpřesnější přibližné řešení. Musíme tedy v rovnici $\mathbf{X}\,\mathbf{b} = \mathbf{y}$ vektor $\mathbf{y}$ nahradit vektorem, který je mu nejbližší a zároveň náleží do lineárního obalu $\mathcal{M}(\mathbf{X})$. Takový vektor podle předchozí kapitoly dostaneme jako ortogonální průmět vektoru $\mathbf{y}$ do prostoru $\mathcal{M}(\mathbf{X})$ [13, 22]. To odpovídá situaci, kdy místo původní soustavy řešíme náhradní soustavu lineárních rovnic $\mathbf{X}\mathbf{b} = \widehat{\mathbf{y}}$. Pro námi nalezený vektor $\mathbf{b}$ a libovolný vektor $\mathbf{c} \in \mathbb{R}^p$ platí stejně jako ve vztahu (3) následující nerovnice

$$\left\| \mathbf{Xb} - \mathbf{y} \right\| \leq \left\| \mathbf{Xc} - \mathbf{y} \right\|,$$

přičemž rovnost nastává právě tehdy, vyhovuje-li vektor $\mathbf{c}$ soustavě lineárních rovnic $\mathbf{Xc} = \widehat{\mathbf{y}}$. Projekce vektoru $\mathbf{y}$ do prostoru $\mathcal{M}(\mathbf{X})$ je znázorněna na obrázku 4. Můžeme tedy říci, že odhad $\mathbf{b}$ pořízený metodou nejmenších čtverců minimalizuje euklidovskou vzdálenost mezi vektorem empirických hodnot $\mathbf{y}$ a prostorem $\mathcal{M}(\mathbf{X})$. [8, 13]



Obr. 4: Kolmý průmět vektoru $\mathbf{y}$

3. Metoda maximální věrohodnosti

Metoda maximální věrohodnosti je univerzální metoda pro konstrukci odhadů parametrů. Používáme ji pro odhad parametrů, které určují rozdělení, např. střední hodnota a rozptyl normálního rozdělení, pravděpodobnost nastoupení jevu alternativního rozdělení apod. Princip metody maximální věrohodnosti spočívá v tom, že za odhad neznámého parametru vezmeme tu jeho hodnotu, ve které tzv. věrohodnostní funkce nabývá svého maxima. [22]

V případě lineární regrese se snažíme touto metodou odhadnout vektor koeficientů $\boldsymbol{\beta}$ a za předpokladu, že pracujeme s lineárním modelem, kde chyby měření mají normální rozdělení a jsou nezávislé, poskytuje tato metoda stejné výsledky odhadů jako metoda nejmenších čtverců.

Podívejme se nejprve na tuto metodu z obecného hlediska. Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ pořízený z rozdělení, které je charakterizováno hustotou $f(\mathbf{x}, \boldsymbol{\theta})$. Toto rozdělení závisí na neznámém vektoru parametrů $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, přičemž $\boldsymbol{\theta}$ náleží parametrickému prostoru Θ, který je podmnožinou p-rozměrného prostoru reálných čísel $\mathbb{R}^p$. Sdružená hustota náhodného vektoru $\mathbf{X}$ je dána předpisem

$$f(x_1, x_2, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}),$$

protože složky náhodného výběru jsou nezávislé náhodné veličiny. Funkci $f(\mathbf{x}; \boldsymbol{\theta})$ proměnné $\boldsymbol{\theta}$ pro pevné $\mathbf{x}$ se říká věrohodnostní funkce a značíme ji $L(\boldsymbol{\theta})$,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}).$$

Analogicky v případě diskrétního rozdělení náhodného výběru charakterizovaného pravděpodobnostmi $p(x_i; \boldsymbol{\theta}), i = 1, 2, \ldots, n$, definujeme věrohodnostní fukci jako

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i, \boldsymbol{\theta}).$$

Pokud je $\widehat{\boldsymbol{\theta}} \in \Theta$ odhadem parametru $\boldsymbol{\theta}$ metodou maximální věrohodnosti, platí pro něj [1, 22]

$$L(\widehat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}) \quad \text{pro každé } \boldsymbol{\theta} \in \Theta.$$

Nyní využijeme metodu maximální věrohodnosti také pro odhad parametrů v lineární regresi. Připoměňme si uvažovaný model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde jsou chyby měření nezávislé a normálně rozdělené s konstantním rozptylem, tedy $\boldsymbol{\varepsilon} \sim N_n(\mathbf{o}, \sigma^2\mathbf{I})$, neboli $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$. Funkce hustoty jednotlivých chyb má v tomto případě tvar

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big(- \frac{1}{2\sigma^2}\, \varepsilon_i^2 \Big).$$

Podobně vyjádříme tvar hustoty vektoru náhodných chyb jako sdruženou hustotu $\prod_{i=1}^{n} f(\varepsilon_i)$. Věrohodnostní funkce se dá zapsat následovně

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} exp\Big(- \frac{1}{2\sigma^2} \boldsymbol{\varepsilon}^{\mathrm{T}} \boldsymbol{\varepsilon} \Big).$$

Jelikož víme, že $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, můžeme funkci dále upravit

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} exp\Big(- \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \Big).$$

V tomto případě je vhodné místo $L(\boldsymbol{\beta}, \sigma^2)$ použít logaritmickou funkci věrohodnosti

$$\ln L(\boldsymbol{\beta}, \sigma^2) = \ln\left(\frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} exp\Big(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\Big)\right) =$$

$$= \ln\big(\sigma^{-n}(2\pi)^{-\frac{n}{2}}\big) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) -$$

$$-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

Tuto funkci s proměnnými $\boldsymbol{\beta}$ a σ^2 se snažíme maximalizovat, a proto ji budeme derivovat nejprve podle vektoru $\boldsymbol{\beta}$ a poté dle proměnné σ^2. Tyto výrazy položíme rovny nule a řešíme vzniklé rovnice. Již z tvaru věrohodnostní funkce je zřejmé, že při řešení problému maximalizace funkce se vlastně dostáváme k úloze, jak minimalizovat výraz

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

stejně, jako tomu bylo u metody nejmenších čtverců [9, 11]. Přesto si dané derivace pro názornost rozepišme [24]

$$\frac{\partial \ln L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\left(-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0}$$

$$\frac{\partial \ln L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^3} = 0$$

Řešením první rovnice dojdeme ke stejnému odhadu (náhodného vektoru) regresních koeficientů

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

jako při použití metody nejmenších čtverců. Z druhé rovnice potom vyplývá hodnota odhadu σ^2 získaného metodou maximální věrohodnosti

$$\tilde{\sigma}^{\,2} = \frac{(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{n}.$$

Tento odhad rozptylu náhodných chyb je zkreslený, narozdíl od nestranného odhadu získaného metodou nejmenších čtverců (2). Mezi těmito dvěma odhady existuje očividný vztah

$$\tilde{\sigma}^{\,2} = \frac{n-p}{n}\ \hat{\sigma}^{\,2}.$$

Zkreslení odhadu $\tilde{\sigma}$ je malé. Jak se n zvětšuje, podíl $(n-p)/n$ se blíží jedné. Odhad $\tilde{\sigma}$ je tedy asymptoticky nestranný, což můžeme vyjádřit jako

$$\lim_{n\to\infty} E(\tilde{\sigma}^{\,2}) = \sigma^2.$$

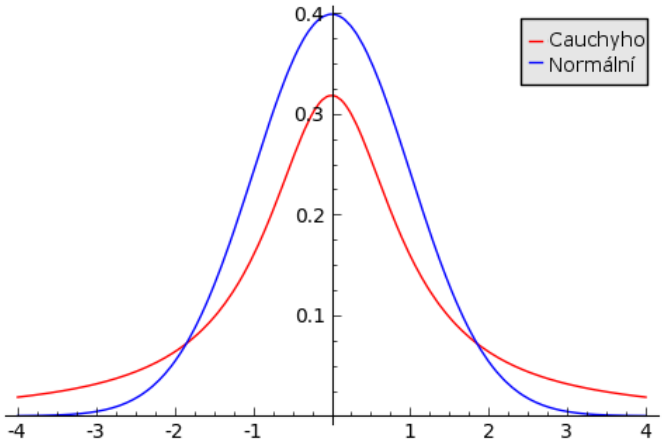
[15]

4. Robustní přístup v lineární regresi

4.1. Důvody k použití robustní regrese

V předchozích kapitolách jsou uvedeny metody výpočtu odhadů regresních koeficientů v lineárním modelu, které jsou optimální v případě, že chyby měření jsou normálně rozděleny. Nicméně v řadě praktických situací je předpoklad normality porušen, čímž použití metody nejmenších čtverců či metody maximální věrohodnosti ztrácí na efektivitě, jelikož odhady regresních koeficientů jsou zkreslené a nevystihují pravou závislost.

Náhodné chyby mohou pocházet ze zcela jiného rozdělení, v praktických situacích se často setkáváme s tím, že náhodné chyby podléhají rozdělení s těžkými chvosty (viz obrázek 5). Taková rozdělení mají tendenci generovat odlehlá pozorování, která mohou silně ovlivňovat metodu nejmenších čtverců v tom smyslu, že jsou schopná změnit směr závislosti. Někdy jsou odlehlá pozorování důsledkem kontaminace souboru (část dat pochází ze souboru s jiným rozdělením či s rozdělením stejným, lišícím se jen střední hodnotou a rozptylem). V nejednom případě je i jediné odlehlé pozorování schopné zcela znehodnotit kvalitu regresních odhadů získaných metodou nejmenších čtverců a může vyvolat nesprávnou představu o kvalitě odhadnuté regresní funkce. [5]



Obr. 5: Příklad rozdělení s těžkými chvosty - Cauchyho rozdělení

Jednou cestou, jak se s odlehlými daty vyrovnat, je jejich vynechání, které by mělo za důsledek regresní funkci procházející hladce zbytkem dat. To však nemusí být jednoduché ani vhodné řešení už jen proto, že odlehlé hodnoty může být složité identifikovat. Mnohem výhodnější postup je přidělení menších vah podezřelým pozorováním. Vyloučení odlehlých hodnot ze souboru také zmenšuje velikost datového vzorku, čímž může zkreslit představu o rozdělení chyb. Dalším způsobem, jak odhadnou regresní koeficienty pro data obsahující odlehlá pozorování, jsou robustní metody, kterým bude věnována tato kapitola a značná pozornost v příkladové části.

Nejprve uvedeme základní klasifikaci odlehlých pozorování.

1. **Odlehlá pozorovnání ve směru osy y**

   Též nazývaná vybočující pozorování. Jsou jimi takové vysoké či nízké hodnoty, které se zásadně liší od ostatních hodnot náhodné proměnné Y.

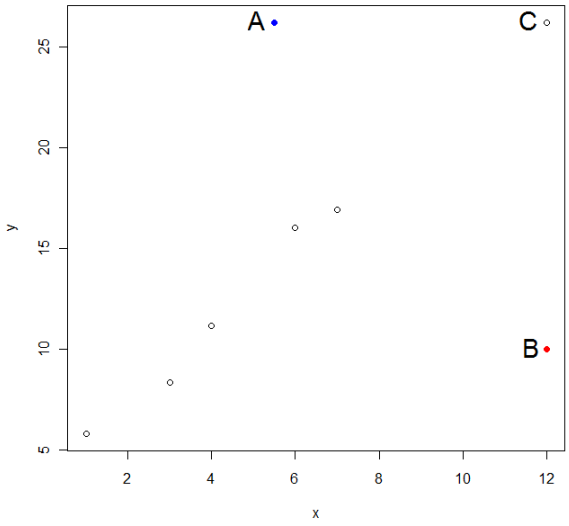
2. **Odlehlá pozorovnání ve směru osy x**

   Jinak řečeno extrémní odlehlá pozorování, která se vykazují značně odlišnými hodnotami vysvětlujících proměnných $x_1, x_2, \ldots, x_k$.

3. **Odlehlá pozorovnání ve směru osy y a osy x**

   Toto pozorování vybočuje ve směru obou os a ne vždy musí regresní model znehodnocovat či nějak ovlivňovat.

Nejlepší vysvětlení výše uvedených pojmů poskytuje obrázek 6, kde bod B je odlehlé pozorování ve směru osy x, bod A představuje odlehlé pozorování ve směru osy y a bod C je odlehlý ve směru obou os. [5, 15]



Obr. 6: Odlehlá pozorování

Problém s odlehlými pozorováními v regresi je znám již od konce 19. století. Na základě teorie robustních odhadů navržené Huberem a Hamplem byla vyvinuta celá řada dalších metod, které modifikují klasickou metodu nejmenších čtverců tak, aby byla zredukována její citlivost na odlehlá pozorování a současně, aby takové metody poskytovaly dobré odhady regresních parametrů v případech, kdy jsou splněny teoretické předpoklady použití klasických metod. [5]

4.2. Eficience

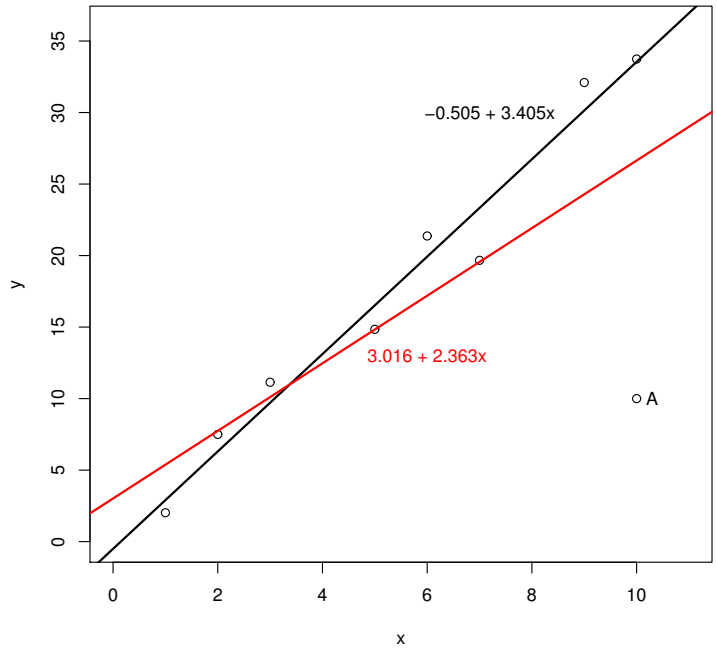
Již v úvodu této kapitoly bylo zmíněno, že od robustních odhadů se mimo jiné vyžaduje, aby poskytovaly dobré výsledky v případech, kdy data neobsahují odlehlá pozorování a rozdělení chyb je normální, tedy pokud by mohla být použita metoda nejmenších čtverců. Jinak řečeno požadujeme, aby robustní odhad byl podobný odhadu metodou nejmenších čtverců. Eficience charakterizuje míru této podobnosti a je definována jako poměr průměru reziduálních čtverců získaných metodou nejmenších čtverců a průměru reziduálních čtverců získaných použitím

jedné z robustních procedur. Čím blíže je tento podíl jedné, tím je robustní odhad eficientnější, občas se též píše vydatnější.

Odborné texty se většinou odkazují na asymptotickou eficienci, která je rovna eficienci odhadu za předpokladu, že rozsah souboru dat n jde do nekonečna. Toto je vhodná statistika pro porovnávání robustních odhadů mezi sebou, z praktického hlediska je ovšem nevhodná, jelikož soubory dat mají často malý rozsah, což způsobuje značnou odchylku pravé eficience od eficience asymptotické. Proto je v praxi oblíbenější tzv. eficience pro konečný výběr.[15]

4.3. Bod selhání

U souborů dat s konečným počtem pozorování můžeme určit takzvaný bod selhání, což je nejmenší podíl anomálních dat, který již může způsobit znehodnocení odhadů regresních koeficientů. Z dosavadních zkušeností vyplývá. že bod selhání se obvykle pohybuje mezi 1 až 10%. Nejmenší hodnota, které bod selhání může nabývat, je $1/n$, což můžeme vysvětlit tak, že jen jediné nevhodné pozorování může ovlivnit odhady koeficientů natolik, že je regresní model nepoužitelný. Právě tuto hodnotu bodu selhání vykazuje metoda nejmenších čtverců. Je to zřetelné z obrázku 7, kde již jediné špatné odlehlé pozorování v bodě A způsobilo vlivnou změnu regresní funkce. Přímka $\widehat{y} = -0.505 + 3.405x$ je výslednou regresní funkcí získanou metodou nejmenších čtverců při vynechání bodu A, který by jinak výslednou přímku vychýlil způsobem, který prezentuje přímka $\widehat{y} = 3.016 + 2.363x$. V praxi je samozřejmě výhodné používat takové metody odhadu parametrů, které mají co nejvyšší bod selhání a nejsou tedy tak náchylné na menší počet odlehlých pozorování. Tato potřeba vedla k vyvinutí odhadů s vysokým bodem selhání. [5, 15]



Obr. 7: Vliv odlehlého pozorování na směr závislosti

Bod selhání byl poprvé formulován Hampelem v roce 1971, který tento pojem definoval čistě z asymptotického hlediska. Až Donoho a Huber (1983) představili zjednodušenou verzi, která pracuje pro konečný výběr. Tuto verzi si stručně představíme. [20]

Mějme data daná vektorem $\mathbf{y}$, kterému odpovídá matice vysvětlujících proměnných $\mathbf{X}$:

$$\mathbf{X}=\begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \qquad \mathbf{y}=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad ,$$

což můžeme též zapsat jako výběr Z, který vyjádříme způsobem

$$\mathbf{Z} = \{(x_{11}, \dots, x_{1p}, y_1), (x_{21}, \dots, x_{2p}, y_2), \dots, (x_{n1}, \dots, x_{np}, y_n)\}.$$

Dále uvažujme regresní odhad T, jehož aplikací na data obdržíme odhad $\widehat{\boldsymbol{\beta}}$ vektoru regresních koeficientů $\boldsymbol{\beta}$, tedy

$$\mathrm{T}(\mathbf{Z}) = \widehat{\boldsymbol{\beta}}.$$

Nyní vezměme v úvahu všechny možné kontaminované výběry Z' úrčené m pozorováními, které v původním výběru Z nahradíme odlehlými hodnotami. Následně můžeme definovat $\mathrm{bias}(m; T, \mathbf{Z})$ jako maximální vychýlení, které může být danou kontaminací způsobeno:

$$\mathrm{bias}(m; \mathrm{T}, \mathbf{Z}) = \sup_{\mathbf{Z}'} ||\mathrm{T}(\mathbf{Z}') - \mathrm{T}(\mathbf{Z})||.$$

Pokud je toto vychýlení rovno nekonečnu, znamená to, že m odlehlých pozorování může mít značně nežádoucí účinek na odhad T, který můžeme vyjádřit jako „selhání“ odhadu. Nyní lze bod selhání odhadu T pro konečný výběr $\mathbf{Z}$ definovat vztahem

$$\epsilon_n^*(\mathrm{T}, \mathbf{Z}) = \min_{\{m|\mathrm{bias}(m;T,Z)=\infty\}} \left\{ \frac{m}{n} \right\}.$$

Jinými slovy se jedná o nejmenší podíl kontaminace, která již dokáže způsobit, že odhad T nabývá hodnot výrazně vzdálených od $\mathrm{T}(\mathbf{Z})$. [18]

4.4. Hlavní metody robustní regrese

Bylo by žádoucí, aby jedna metoda měla všechny výhodné vlastnosti a pracovala uspokojivě se všemi druhy dat, ať už s odlehlými hodnotami či bez. Taková metoda však nebyla zatím představena, protože každá situace vyžaduje metodu jinou. Metoda nejmenších čtverců pracuje výborně, pokud je použita na datech pocházejících z normálního rozdělení, ovšem poskytuje zkreslené výsledky, kdykoli je tento předpoklad porušen. Stejně tak každá robustní metoda má jisté výhody a nevýhody, a proto si zde stručně představíme základní metody robustní regrese.

L - odhady

L-odhady regresních koeficientů navrhli v roce 1978 Koenker a Basset, kteří využili regresních kvantilů. Pokud $0 < \alpha < 1$ je konstanta, regresní αkvantil je řešení minimalizační úlohy

$$\widehat{\boldsymbol{\beta}}_\alpha = \min_\beta \left[\sum_{\{i|\varepsilon_i>0\}} \alpha\varepsilon_i + \sum_{\{i|\varepsilon_i<0\}} (1-\alpha)\varepsilon_i \right]. \tag{4}$$

Tato úloha je ekvivalentní úloze

$$\widehat{\boldsymbol{\beta}}_\alpha = \min_\beta \sum_{i=1}^{n} \rho_\alpha(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta}), \tag{5}$$

kde funkce $\rho_\alpha(x)$ je daná předpisem

$$\rho_\alpha(x) = \begin{cases} x\alpha & x \geq 0 \\ x(\alpha-1) & x < 0 \end{cases}$$

Tím se vlastně dostáváme k M-odhadu definovanému funkcí $\rho_\alpha(x)$, který si uvedeme v následující podkapitole. Vyřešením minimalizační úlohy získáme nadrovinu procházející p (počet regresních parametrů) body, která dělí výběrový prostor tak, že v jedné části leží $n\alpha$ pozorování a v části druhé leží pozorování zbývající.

L-odhady založené na regresních kvantilech jsou analogické lineárním kombinacím výběrových kvantilů v odhadech polohy. Ve výsledku se nejedná o nic jiného, než o odhad metodou nejmenších čtverců z těch pozorování, jež leží mezi nadrovinami odpovídajícími α a $(1-\alpha)$-regresnímu kvantilu, ostatní pozorování jsou z výpočtu vyloučena. Velikost useknutí dat závisí na kontaminaci dat, bere se nejčastěji od 5 do 10 %. L-odhady je vhodné použít v případe odlehlých hodnot ve vysvětlované proměnné. [3, 5]

Do této třídy odhadů řadíme též L_1-odhad, též označovaný jako LAD (Least absolute deviations, v překladu nejmenší absolutní odchylka) odhad či odhad v L_1-normě. Tento odhad byl prvním krokem v používání robustních odhadů a s

myšlenkou na jeho použití přišel Francis Edgeworth roku 1887. Odhad je používán jako alternativa k metodě nejměnších čtverců, pokud se obáváme výskytu odlehlých pozorování. Podstatou odhadu v L_1-normě je minimalizovat součet absolutních hodnot reziduí

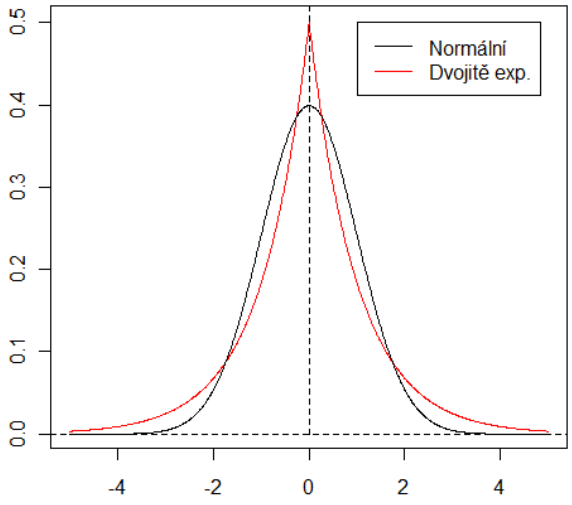
$$|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}| = \sum_{i=1}^{n} |(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\,\boldsymbol{\beta})|. \tag{6}$$

K řešení minimalizace absolutních reziduí se dostáváme většinou v případě, kdy chyby měření podléhají rozdělení s těžkými chvosty. Zaměřme se konkrétně na dvojitě exponencionální rozdělení. Uvažujme jednoduchý lineární model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde ε_i jsou náhodné chyby s dvojitě exponencionálním rozdělením

$$f(\varepsilon_i) = \frac{1}{2\sigma} exp\Big(\frac{-|\varepsilon_i|}{\sigma}\Big).$$



Obr. 8: Dvojitě exponencionální rozdělení

Z obrázku 8 tohoto rozdělení je zřejmé, že chvosty hustoty dvojitě exponencionálního rozdělení se blíží k nule pomaleji, než je tomu například u normálního

rozdělení. Od normálního rozdělení se také liší větší špičatostí, ovšem tato dvě rozdělení mají i společné vlastnosti - jako symetrii.[5, 20]

Pro odhad parametrů β_0 a β_1 použijeme metodu maximální věrohodnosti. Věrohodnostní funkce je tvaru

$$L(\boldsymbol{\varepsilon}, \sigma) = \prod_{i=1}^{n} \frac{1}{2\sigma} \exp\Big(-\frac{|\varepsilon_i|}{\sigma}\Big) = \frac{1}{(2\sigma)^n} \exp\left(-\frac{\sum_{i=1}^{n} |\varepsilon_i|}{\sigma}\right).$$

Analogicky při znalosti modelu

$$L(\beta_0, \beta_1, \sigma) = \frac{1}{(2\sigma)^n} \exp\left(-\frac{\sum_{i=1}^{n} |y_i - \beta_0 + \beta_1 x_i|}{\sigma}\right).$$

Při hledání maxima této funkce stačí minimalizovat $\sum_{i=1}^{n} |\varepsilon_i|$, tedy $\sum_{i=1}^{n} |y_i - \beta_0 + \beta_1 x_i|$.

Je zřejmé, že tato metoda odhadu regresních koeficientů je pro dané rozdělení náhodných chyb vhodnější, než by byla metoda nejmenších čtverců. U dvojitě exponencionálního rozdělení je větší pravděpodobnost výskytu odlehlých pozorování, která by odhady pořízené metodou nejmenších čtverců silně znehodnotila. Stačí si představit, jak velké by byly odchylky těchto pozorování od předpokládané optimální regresní přímky, pokud by byly při použití metody nejmenších čtverců umocněny na druhou. Pokud pracujeme s absolutní hodnotou odchylek, je logické, že takto vzniklé hodnoty nebudou mít na závěrečnou sumarizaci tak drtící vliv, jako tomu je u „sčítání čtverců“. L_1-odhad tedy nebude kvůli odlehlým hodnotám natolik zkreslený. [15]

Úloha minimalizace výrazu (6) se dá přepsat za použití regresních kvantilů. V tomto speciálním případě využijeme hodnotu $\alpha = 1/2$ a získame tzv. regresní medián $\widehat{\boldsymbol{\beta}}_{\frac{1}{2}}$. Při zápisu regresního mediánu pomocí funkce ρ z rovnice (5) dojdeme ke vztahu

$$\widehat{\boldsymbol{\beta}}_{\frac{1}{2}} = \min_{\beta} \sum_{i=1}^{n} \rho_{\frac{1}{2}}(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta}).$$

To lze také vyjádřit pomocí vztahu (4), kde pro jednodušší zápis minimalizovanou funkci označíme $\tau(1/2)$ a tím se dostáváme k regresnímu mediánu

$$\widehat{\boldsymbol{\beta}}_{\frac{1}{2}} = \min_{\beta} \tau(1/2).$$

Pomocí $\widehat{\boldsymbol{\beta}}_{\frac{1}{2}}$ lze L_1-odhad vypočítat takto:

$$\widehat{\boldsymbol{\beta}}_{L_1} = \min_{\beta}[2\ \tau(1/2)].$$

Tuto rovnost dokážeme jednoduchou úpravou funkce $\tau(\alpha)$:

$$\tau(1/2) = \left[\sum_{\{i|\varepsilon_i>0\}} \frac{1}{2}\,\varepsilon_i + \sum_{\{i|\varepsilon_i<0\}} \left(1-\frac{1}{2}\right)\varepsilon_i \right] =$$

$$= \left[\frac{1}{2} \sum_{\{i|\varepsilon_i>0\}} |\varepsilon_i| + \frac{1}{2} \sum_{\{i|\varepsilon_i<0\}} |\varepsilon_i| \right] = \frac{1}{2} \sum_{i=1}^{n} |\varepsilon_i|.$$

Tím se dostáváme k závěrečnému L_1-odhadu ve tvaru

$$\widehat{\boldsymbol{\beta}}_{L_1} = \min_{\beta} \sum_{i=1}^{n} |y_i - \beta_0 + \beta_1 x_i|.$$

Bod selhání L_1-odhadu pro konečné výběry je $1/n$. L_1 regrese je robustní vůči odlehlým pozorováním ve směru osy y, ale neposkytuje ochranu proti odlehlým pozorováním ve směru osy x. [5, 14, 25]

M - odhady

Nejpoužívanější metodou robustní regrese jsou M-odhady, představené Peterem Huberem v roce 1964. Tato třída odhadů může být chápána jako zobecnění metody maximální věrohodnosti, odtud písmeno M v názvu M-odhadů. Stejně jako L-odhady je i tato třída odhadů robustní vůči odlehlým pozorováním ve směru osy y, nikoli ve směru osy x.

Jako i v předchozím textu uvažujme klasický lineární regresní model. M-odhad regresních parametrů je definován jako rešení minimalizující součet funkcí reziduí,

$$\min_{\beta} \sum_{i=1}^{n} \rho(\varepsilon_i) = \min_{\beta} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta}), \tag{7}$$

kde ρ je rozumná funkce, která by měla mít několik základních vlastností:

1. nezáporná funkce, $\rho(x) \geq 0$,
2. funkce procházející počátkem, $\rho(0) = 0$,
3. symetrická funkce, $\rho(x) \geq \rho(-x)$,
4. monotónní funkce, $\rho(x_1) \geq \rho(x_2)$ pro $|x_1| \geq |x_2|$.

[6, 25]

M-odhady regresních koeficientů nejsou obecně invariantní vůči změně měřítka, tj. pokud jsou chyby měření násobeny nějakou konstantou, řešení úlohy (7) se tím změní. Abychom tento problém obešli, vytvoříme novou verzi odhadu, která již bude invariantní:

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right) = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta}}{s}\right), \tag{8}$$

kde s je odhad měřítka. Ten může být vypočten například oblíbeným způsobem za použití MAD (median absolute deviation, v překladu medián absolutních odchylek)

$$s = \underset{i}{\mathrm{med}}\, |\varepsilon_i - \underset{j}{\mathrm{med}}(\varepsilon_j)|/0.6745.$$

Jmenovatel je konstanta 0.6745, která z s dělá nestranný odhad parametru σ, pokud je n dostatečně velké a chyby jsou normálně rozděleny. Získáme ji jako 0.75-kvantil normovaného normálního rozdělení. Pro odhad σ je možné použít i jiných funkcí, které budou uvedeny v příkladové části u výčtu možností procedur programu SAS. [5, 15, 18]

Úlohu minimalizace (8) převedeme na řešení soustavy p rovnic tím, že součet funkcí reziduí zderivujeme podle $\beta_j, j = 0, 1, \ldots, k$, a vzniklé výrazy položíme rovny nule. To zapíšeme jako

$$\sum_{i=1}^{n} x_{ij}\, \psi\left(\frac{\varepsilon_i}{s}\right) = \sum_{i=1}^{n} x_{ij}\, \psi\left(\frac{(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\, \boldsymbol{\beta})}{s}\right) = 0, \quad j = 1, 2, \ldots, k, \tag{9}$$

kde $\psi = \rho'$ a x_{ij} je i-tá hodnota u j-tého regresoru, $x_{i0} = 1$.

Pro výpočet M-odhadů byla navržena řada algoritmů, mimo jiné i relativně jednoduchá metoda iteračně vážených nejmenších čtverců, která je součástí výpočtu v počítačových softwarech. Odhad pořízený metodou vážených nejmenších čtverců se v mnohém podobá odhadu pořízenému obyčejnou metodou nejmenších čtverců a má tvar

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y},$$

kde $\mathbf{W}$ je diagonální matice vah o rozměrech $n \times n$.[15]

Pokud tedy uvažujeme iterační postup, první takto pořízený odhad je

$$\widehat{\boldsymbol{\beta}}^{(1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_0\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_0\mathbf{Y}.$$

Diagonální prvky matice $\mathbf{W}_0$, $w_{10}, w_{20}, \ldots, w_{n0}$, jsou váhy dané vztahem

$$w_{i0} = \begin{cases} \dfrac{\psi[(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\, \widehat{\boldsymbol{\beta}}^{(0)})/s]}{(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\, \widehat{\boldsymbol{\beta}}^{(0)})/s} & \text{pro } y_i \neq \mathbf{x}_{i\cdot}^{\mathrm{T}}\, \widehat{\boldsymbol{\beta}}^{(0)} \\ 1 & \text{pro } y_i = \mathbf{x}_{i\cdot}^{\mathrm{T}}\, \widehat{\boldsymbol{\beta}}^{(0)} \end{cases}, \tag{10}$$

kde $\widehat{\boldsymbol{\beta}}^{(0)}$ je nějaký počáteční odhad regresních koeficientů a s je odhad měřítka. V dalším kroku iteračního procesu budeme postupovat analogicky, pouze místo $\widehat{\boldsymbol{\beta}}^{(0)}$ použijeme $\widehat{\boldsymbol{\beta}}^{(1)}$. Většinou stačí pouze několik iterací k dosažení konvergence.

Výše uvedené váhy (10) jsou odvozeny z rovnice (9), kterou lze rozepsat jako

$$\sum_{i=1}^{n} x_{ij}\, \psi\left(\frac{(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\, \boldsymbol{\beta})}{s}\right) =$$

$$= \sum_{i=1}^{n} x_{ij} \psi\left(\frac{(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta})}{s} \right) \frac{(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta})}{s} \frac{s}{(y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta})} = 0 \tag{11}$$

pro $j = 0, 1, \ldots, k$. A rovnici (11) dále přenásobíme konstantou s, získáme soustavu k+1 normálních rovnic ve tvaru [15]

$$\frac{1}{s} \sum_{i=1}^{n} x_{ij} w_{i0} (Y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta}) = 0, \quad j = 1, 2, \ldots, k.$$

K posouzení kvality odhadů či pro sestavení konfidenčních intervalů je nezbytné znát rozptyl odhadů koeficientů a tedy kovarianční matici vektoru $\widehat{\boldsymbol{\beta}}$. Huber a Ronchetti za tímto účelem nabízí tři nestranné odhady kovariančních matic [11]:

$$\mathrm{K}^2 \frac{[1/(n-p)] \sum \psi(\varepsilon_i)^2}{[(1/n) \sum \psi'(\varepsilon_i)]^2} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1}, \tag{12}$$

$$\mathrm{K} \frac{[1/(n-p)] \sum \psi(\varepsilon_i)^2}{(1/n) \sum \psi'(\varepsilon_i)} \mathbf{W}^{-1},$$

$$\mathrm{K}^{-1} \frac{1}{n-p} \sum \psi(\varepsilon_i)^2 \, \mathbf{W}^{-1} (\mathbf{X}^{\mathrm{T}} \mathbf{X}) \mathbf{W}^{-1},$$

kde pro K a $\mathbf{W}$ platí

$$\mathrm{K} = 1 + \frac{p \; var(\psi')}{n \; (E(\psi'))^2},$$

$$W_{jk} = \sum \psi'(\varepsilon_i) x_{ij} x_{ik}.$$

V praxi jsou $E(\psi')$ a $var(\psi')$ neznámé, ale je možné je odhadnout pomocí výběrového průměru a rozptylu

$$E(\psi') \cong m = \frac{1}{n} \sum \psi'(\varepsilon_i),$$

$$var(\psi') \cong \frac{1}{n} \sum [\psi'(\varepsilon_i) - m]^2.$$

Funkce	$\rho(z)$	$\psi(z)$	$w(z)$	rozsah
Metoda nejmenších čtverců	$\frac{1}{2}z^2$	z	1.0	$\lvert z\rvert < \infty$
Huberova funkce	$\frac{1}{2}z^2$	z	1.0	$\lvert z\rvert \leq t$
Hampelova funkce	$\frac{1}{2}z^2$	z	1.0	$\lvert z\rvert \leq a$
	$a\lvert z\rvert - \frac{1}{2}a^2$	$a\text{sign}(z)$	$a/\lvert z\rvert$	$a < \lvert z\rvert \leq b$
	$\frac{a(c\lvert z\rvert - \frac{1}{2}z^2)}{c-b} - \frac{7}{6}a^2$	$\frac{a\text{sign}(z)(c-\lvert z\rvert)}{c-b}$	$\frac{a(c-\lvert z\rvert)}{\lvert z\rvert(c-b)}$	$b < \lvert z\rvert \leq c$
	$a(b+c-a)$	0	0	$\lvert z\rvert > c$
Anrewsova funkce	$a[1-cos(z/a)]$	$\sin(z/a)$	$\frac{\sin(z/a)}{z/a}$	$\lvert z\rvert \leq a\pi$
	$2a$	0	0	$\lvert z\rvert > a\pi$

Tabulka 1: Funkce M-odhadů

Existuje řada oblíbených funkcí ρ, které bývají při používání metody M-odhadů v robustní regresi používány. Několik z nich je uvedeno v tabulce 1. Odvození jednotlivých charakteristik funkce si můžeme ukázat například na metodě nejmenších čtverců, kde uvažujeme funkci reziduí $\rho(z) = \frac{1}{2}z^2$, jež po zderivování bude odpovídat $\psi(z) = z$. S využitím (10) již není problémem dopočítat,

že $w(z) = \frac{z/s}{z/s} = 1$. Analogicky bychom mohli odvodit charakteristiky Andrewsovy funkce. Pokud uvažujeme $\rho(z) = a[1 - cos(z/a)]$, derivací této funkce se dostáváme k $\psi(z) = (-a)(-\sin(z/a))\frac{1}{a} = \sin(z/a)$ a $w(z) = \frac{\sin(z/a)}{z/a}$.

ODHADY S VYSOKÝM BODEM SELHÁNÍ

Odhady s vysokým bodem selhání jsou potřeba, protože v řadě situací je kontaminována značná část souboru, kterou doposud uvedené metody nemusí zvládnout, aniž by odhad neselhal. Odhady s vysokým bodem selhání jsou konstruovány tak, aby byly současně vydatné v případě, kdy rozdělení chyb je normální a soubor dat neobsahuje odlehlá pozorování, a aby dosahovaly bodu selhání až 50 %. V následujícím textu se seznámíme s několika nejznámějšími odhady splňujícími tyto požadavky. [5, 15]

LMS-odhady (Least median of squares)

Jedna z prvních metod robustní regrese by se v českém překladu dala nazvat metodou nejmenšího mediánu čtverců, v angličtině Least median of squares, což dále zkrátíme na LMS. Tato metoda snižuje vliv odchylek dat od regresní funkce, čímž je vhodnou robustní metodou odhadu regresních koeficientů v případě, že soubor obsahuje odlehlá pozorování. LMS odhad koeficientů byl navrhnut Peterem J. Rousseeuwem roku 1984 a ho dostaneme jako řešení minimalizační úlohy

$$\min_{\boldsymbol{\beta}} \operatorname*{med}_{i} (y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta})^2.$$

Robustnost této metody je zřejmá z informace, že metoda nejmenších čtverců minimalizuje sumu čtverců reziduí, tudíž minimalizuje průměr čtverců reziduí. Ovšem průměr je znám svými špatnými vlastnostmi, pokud slouží jako odhad polohy za přítomnosti odlehlých pozorování. Medián je v těchto případech mnohem oblíbenější statistikou. Z geometrického pohledu lze tuto metodu interpretovat jako nalezení nejtěsnějšího pásu, jež pokrývá polovinu pozorování, přičemž polovinou se rozumí $[n/2]+1$. Regresní přímka vypočtená metodou LMS pak leží přesně uprostřed tohoto pásu.

Při počtu vysvětlujících proměnných p je bod selhání roven $([n/2]-p+2)/n$. LMS odhady jsou možná robustní, mají však špatné asymptotické vlastnosti, pomalu konvergují k normalitě a nejsou příliš vhodné v případě, kdy data splňují předpoklady pro použití klasické regrese. [5, 15]

LTS-odhady (Least trimmed squares estimator (LTSE))

Odhad metodou nejmenších useknutých čtverců, který navrhl (stejně jako LMS-odhady) Rousseeuw roku 1984, získáme jako řešení

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{h} (y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{h} \varepsilon_i^2,$$

kde $\varepsilon_1^2 < \varepsilon_2^2 < \ldots < \varepsilon_n^2$ jsou čtvercová rezidua uspořádaná dle velikosti, h je tzv. „usekávací“ konstanta, která musí splňovat $\frac{n}{2} < h \leq n$. Tato hodnota ovlivňuje bod selhání odhadu metodou LTS, což je zřejmé z minimalizační úlohy, jež je v podstatě úlohou z metody nejmenších čtverců, vyjma toho, že minimalizuje součet pouze prvních h čtverců náhodných chyb. Z toho vyplývá, že konstanta h mimo jiné udává, že $n-h$ největších reziduí nebude mít vliv na výsledný odhad.

Hodnota h může nabývat různých hodnot, pro maximální bod selhání volíme $h = [n/2] + [p/2]$, volbou $h = n$ se zase dostáváme ke klasické metodě nejmenších čtverců, bod selhání se tím ovšem sníží na $1/n$. Hodnotu h lze též volit v závislosti na tom, za jakým účelem metodu LTS provádíme. Pokud chceme zachovat vysoký bod selhání, pracujeme s malými hodnotami h, na druhou stranu stanovíme vysokou hodnotu h v případě, že chceme zachovat vydatnost odhadu, kdy je zachováno více informací při využití více dat. Vysoké h je vhodné použít při menší kontaminaci dat a nízké naopak při větší kontaminaci.[5, 15, 30]

S-odhady (S-Estimators (SE))

V roce 1984 Rousseeuw a Yohai zobecnili odhady LMS a LTS a navrhli S-odhad, který na rozdíl od předchozích metod minimalizuje robustní M-odhad

směrodatné odchylky reziduí. Ukazuje se, že tento odhad má stejné asymptotické vlastnosti jako M-odhady, navíc má výhodné robustní vlastnosti, jelikož jeho bod selhání může dostáhnout 50%. Pro rezidua s normálním rozdělením mají S-odhady nižší vydatnost než M-odhady. Pokud ovšem S-odhady porovnáváme s LMS-odhady a LTS-odhady, jejich asymptotická eficience je vyšší. [5, 15, 20]

S-odhad získáme jako minimalizaci směrodatné odchylky reziduí

$$\min_{\boldsymbol{\beta}} s[\varepsilon_1(\boldsymbol{\beta}), \varepsilon_2(\boldsymbol{\beta}), \ldots, \varepsilon_n(\boldsymbol{\beta})],$$

přičemž směrodatná odchylka $s[\varepsilon_1(\boldsymbol{\beta}), \varepsilon_2(\boldsymbol{\beta}), \ldots, \varepsilon_n(\boldsymbol{\beta})]$ je dána jako řešení rovnice

$$\frac{1}{n}\sum_{i=1}^{n} \rho(\varepsilon_i/s) = k, \tag{13}$$

kde k je konstanta závisející na zvolené funkci ρ, obvykle je volena jako $k = \mathrm{E}_\Phi(\rho)$. Tato funkce by stejně jako funkce M-odhadů měla být symetrická, neklesající a měla by procházet počátkem. Výsledný odhad měřítka bude následující

$$\hat{\sigma}^2 = s[\varepsilon_1(\widehat{\boldsymbol{\beta}}), \varepsilon_2(\widehat{\boldsymbol{\beta}}), \ldots, \varepsilon_n(\widehat{\boldsymbol{\beta}})].$$

Název S-odhadů je odůvodněn jejich odvozením od statistiky měřítka (směrodatné odchylky), v angličtině scale statistic. Dle dosavadních zkušeností nejsou S-odhady užívány v praxi příliš hojně, což je pravděpodobně způsobeno nedostatečným počtem programů, jež by uměly tento odhad spočítat. [11, 15, 18, 20]

MM-odhady (MM-Estimators (MME))

MM-odhady poprvé navrhl Yohai roku 1987 jako odhady s vysokým bodem selhání, které zároveň vykazují dobrou eficienci. MM-odhady jsou definovány pomocí třífázového postupu. V první fázi vypočteme počáteční regresní odhad, od kterého požadujeme konzistenci a vysoký bod selhání, nikoli vydatnost. Většinou volíme S-odhad. Ve druhém kroku je vypočten M-odhad směrodatné odchylky

chyb na základě reziduí vypočtených z modelu s počátečním odhadem, jedná se vlastně o směrodatnou odchylku $\hat{\sigma}^2$ definovanou vztahem (13). V poslední třetí fázi jsou iterativní procedurou vypočteny M-odhady regresních parametrů s vhodnou neklesající funkcí ψ. [5, 15]

5. Hodnocení kvality regresního modelu

Ještě před demonstrací popsaných metod na příkladech bude vhodné úvést, co rozumíme pod pojmem kvalita regresní funkce a jakými statistikami můžeme tuto kvalitu popsat.

Obecně platí, že regresní funkce je tím lepší, čím je závislost proměnných silnější a čím více jsou empirické hodnoty vysvětlované proměnné soustředěné kolem odhadu regresní křivky. Naopak, vztah je tím slabší, čím více jsou empirické hodnoty vzdáleny hodnotám vyrovnaným. Míra intenzity závislosti úzce souvisí s hodnocením účinnosti odhadnuté regresní funkce a tedy s kvalitou regresního odhadu. Pro kvalitu regresní funkce používáme zejména charakteristiky jako index determinace, rozptyl odhadnutých koeficientů a analýzu reziduí.

5.1. Reziduální součet čtverců

Rezidua jsou základní diagnostickým nástrojem při hodnocení kvality regresní funkce. Často používanou charakteristikou reziduí je reziduální součet čtverců daný vztahem

$$\mathrm{S}_R = \sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2.$$

Symbolem Y_i označujeme i-tou napozorovanou (empirickou) hodnotu a $\widehat{Y}_i$ nechť je i-tou vyrovnanou hodnotou. Pro reziduální součet čtverců intuitivně platí, že čím je menší, tím je regresní funkce kvalitnější. Ovšem tento předpoklad platí pouze pro nerobustní metody a modely, kde chyby měření mají normální rozdělení, tj. pokud soubor dat neobsahuje odlehlá pozorování.

Tato charakteristika je tedy vhodná především v případě, kde mezi sebou porovnáváme několik regresních funkcí, které se liší svým tvarem, nikoli však metodou výpočtu, kterou by měla být metoda nejmenších čtverců. Pokud ovšem počítáme s konstantním tvarem regresní funkce, v našem případě půjde o obecnou přímku, a měníme pouze metody výpočtu, je zřejmé, že nejlépe ze vzájemného porovnávání výjde metoda nejmenších čtverců, která má minimalizaci čtvercových

reziduí přímo za úkol. A to i v případě, že odhady metodou nejmenších čtverců budou naprosto nedůvěryhodné a zkreslené. Pro robustní metody by bylo třeba provést modifikaci. [4]

5.2. Index determinace

Další charakteristikou kvality regresní funkce, která trpí podobným nedostatkem jako reziduální součet čtverců, je index determinace (v angličtině R-square). Na úvod zmiňme, že v regresním modelu lze konstruovat tři základní typy součtu čtvercových odchylek:

1. celkový součet čtverců - charakterizuje celkovou variabilitu dat

$$\mathrm{S}_C = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

   kde $\overline{Y} = \frac{1}{n} \sum Y_i$,

2. teoretický součet čtverců - charakterizuje část variability závisle proměnné Y, která je zachycena regresní funkcí

$$\mathrm{S}_T = \sum_{i=1}^{n} (\widehat{Y}_i - \widetilde{Y})^2,$$

   kde $\widetilde{Y} = \frac{1}{n} \sum \widehat{Y}_i$,

3. reziduální součet čtverců - charakterizuje část variability závisle proměnné Y, kterou nelze vysvětlit regresní funkcí

$$\mathrm{S}_R = \sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2.$$

Index determinace, značíme R^2, je konstruovaný jako poměr teoretického součtu čtverců a celkového součtu čtverců.

$$R^2 = \frac{\mathrm{S}_T}{\mathrm{S}_C}.$$

Parametr může nabývat hodnot v intervalu $< 0,1 >$, navíc ho lze vyjádřit v procentech, pokud hodnotu vynásobíme stem. Index determinace vyjádřený v procentech udává část rozptylu závisle proměnné Y, kterou je možné vyjádřit zvolenou regresní funkcí. Je žádoucí, aby index nabýval hodnot blízkých k jedné, čímž je zaručena silná závislost proměnných a také vhodnost zvolené regresní funkce.

V případě použití metody nejmenších čtverců k výpočtu odhadů regresních koeficientů platí $\overline{Y} = \widetilde{Y}$ a také je možné definovat vztah mezi jednotlivými rozptyly $\mathrm{S}_C = \mathrm{S}_T + \mathrm{S}_R$, který nám dovoluje index determinace přepsat ve tvaru

$$\mathrm{R}^2 = 1 - \frac{\mathrm{S}_R}{\mathrm{S}_C}.$$

Dále je možné R^2 upravit tak, že získáme adjungovaný index determinace (označovaný v anglické literatuře adjusted R-square). Ten se hojně používá díky své schopnosti potlačit vliv rostoucího počtu koeficientů regresní funkce, což má u obyčejného indexu determinace za následek růst hodnoty R^2, ačkoliv funkce nemusí být nejvýstižnější. Definujeme ho vztahem

$$\mathrm{R}^2_{adj} = 1 - (1 - \mathrm{R}^2)\,\frac{n-1}{n-p}.$$

Index determinace je ovšem vhodnou charakteristikou pouze v případě použití metody nejmenších čtverců, tedy za přepokladu, že náhodné chyby podléhají normálnímu rozdělení. Již jediné odlehlé pozorování způsobí, že index determinace má nulovou výpovědní hodnotu o kvalitě regresní funkce. [4, 12]

Pro tyto situace byly stejně jako pro případ regrese navrženy robustní indexy determinace, které pracují dobře i za přítomnosti odlehlých pozorování v datech a zároveň vykazují dobrou eficienci v případě splnění předpokladů pro použití obyčejného R^2. Nevýhodou robustních indexů determinace je dozajista jejich vychýlenost ve většině případů.

Například robustní index determinace pro situace, kdy je k odhadu regresních parametrů použit jeden z M-odhadů, navrhl R.A. Maronna v roce 2006. Jedná

se o robustní ekvivalent R^2 daný rovnicí

$$\mathrm{R}^2_\rho = 1 - \frac{\sum_{i=1}^{n} \rho\left(\frac{Y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{s}\right)}{\sum_{i=1}^{n} \rho\left(\frac{Y_i - \widehat{\mu}}{s}\right)}.$$

Symbolem $\widehat{\mu}$ se rozumí M-odhad střední hodnoty proměnné Y, který vznikne řešením úlohy

$$\min_{\mu} = \sum_{i=1}^{n} \rho\left(\frac{Y_i - \mu}{s}\right).$$

Dále uvažujme s jako odhad směrodatné odchylky σ. Stejným způsobem bude vypočítán robustní index determinace pro MM-odhady.

V příkladové části bude vyčíslen také robustní index determinace pro LMS-odhad podle vzorce vynalezeného vědci Rousseeuwem a Leroyem

$$\mathrm{R}^2_{rob} = 1 - \left(\frac{\mathrm{med}_i \, |Y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{LMS}|}{\mathrm{mad}(Y_i)}\right)^2.$$

Výrazem v čitateli $\mathrm{mad}(y_i)$, v angličtině nazývaným median absolute deviation, zde myslíme

$$\underset{i}{\mathrm{med}}\{|Y_i - \underset{j}{\mathrm{med}}\, Y_j|\}.$$

Představme si také robustní indexy determinace pro S-odhady či metodu LTS.

$$\mathrm{R}^2_{rob} = 1 - \left(\frac{(n-p)\mathrm{S}_p^2}{(n-1)\mathrm{S}_\mu^2}\right)^2$$

pro metodu S-odhadu, kde S_p představuje S-odhad směrodatné odchylky v celém modelu, S_μ je S-odhad v regresním modelu obsahujícím pouze absolutní člen. Pro metodu LTS bude index determinace vypočítán dle následujícího vzorce

$$\mathrm{R}^2_{rob} = 1 - \frac{\mathrm{s}^2_{LTS}(X,Y)}{\mathrm{s_{LTS}}^2(1,Y)},$$

kde $\mathrm{s}_{\mathrm{LTS}}(1, Y)$ je odhad směrodatné odchylky v modelu obsahujícím pouze absolutní člen a $\mathrm{s}_{LTS}(X, Y)$ je odhad daný vztahem

$$\mathrm{s}_{LTS}(X, Y) = d_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{n}\varepsilon_i^2},$$

kde

$$d_{h,n} = 1/\sqrt{1 - \frac{2n}{h\, c_{h,n}}\,\phi\left(\frac{1}{c_{h,n}}\right)},$$

za předpokladu, že

$$c_{h,n} = 1/\Phi^{-1}\left(\frac{h+n}{2n}\right),$$

přičemž ϕ značí funkci hustoty a Φ distribuční funkci normovaného normálního rozdělení.[17, 18, 21]

5.3. Akaikeho informační kritérium, Bayesovské informační kritérium

Dalšími statistikami vyvinutými pro posouzení kvality modelu jsou informační kritéria, která bývají mnohdy využívána pro odhadnutí počtu parametrů. Nejprve si přiblime informační kritérium představené japonským statistikem Hirotsugem Akaikem v roce 1971. Kritérium je vyjádřeno rovnicí

$$\mathrm{AIC} = 2k - 2\ln L,$$

kde k je počet parametrů a L je maximální hodnota věrohodnostní funkce. Narozdíl od indexu determinace je zde preferován model, pro který toto kritérium nabývá minima.

Toto kritérium bude v následujících příkladech vypočítáno pro M a MM-odhady pro porovnání jejich přesnosti. Samozřejmě bude uvažován robustní ekvivalent, jelikož metody jsou robustní a není v nich uvažováno specifické rozdělení

náhodných chyb, není tedy možné maximalizovat věrohodnostní funkci. Robustní AIC budeme značit AICR a vyjadříme ho jako

$$\mathrm{AICR} = 2\sum_{i=1}^{n} \rho\left(\frac{Y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}}{s}\right) + \alpha p,$$

kde s jako vždy značí odhad směrodatné odchylky, $\widehat{\boldsymbol{\beta}}$ je M či MM-odhad regresních koeficientů a p je dimenze matice $\mathbf{X}$. Penalizaci pro počet koeficientů představuje parametr α vypočtený jako $\alpha = 2\,\dfrac{E\psi^2}{E\psi'}$.

Příbuzným kritériem AIC je Bayesovské informační kritérium, navržené Gideon E. Schwarzem v roce 1978, které opět může sloužit k odhadu počtu parametrů v modelu. Míra penalizace vzniklá přidaným parametrem je zde však ještě vyšší, než tomu bylo u AIC. Kritérium značíme BIC a je dáno vzorcem

$$\mathrm{BIC} = k\ln(n) - 2\ln L.$$

Pro BIC platí to samé, jako pro Akaikeho kriterium, tedy - budeme ho používat pro porovnání M a MM-odhadů v jeho upravené robustní verzi, která vypadá následujícím způsobem

$$\mathrm{BICR} = 2\sum_{i=1}^{n} \rho\left(\frac{Y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}}{s}\right) + p\,\log(n).$$

[21, 23]

6. Aplikace na datech

Poslední kapitola této práce bude věnována výpočtu regresních odhadů pomocí všech robustních i nerobustních metod popsaných v předchozích kapitolách. Výpočty budou prováděny na nasimulovaných datech, která budou kopírovat čtyři možné situce z hlediska obsahu odlehlých pozorování. Tyto situace budou také souviset s vlastnostmi jednotlivých metod odhadů, kteréžto byly diskutovány dříve v této práci - tím myslíme robustnost jako takovou a následně též vydatnost či bod selhání.

Budeme uvažovat 10 dat z klasického lineárního modelu

$$Y = 10 + 3x + \varepsilon, \quad \varepsilon \sim N(0, 2).$$

Jedná se o jednoduchou (přímkovou) regresi, která nás provázela celým předešlým textem. Malý vzorek dat je volen záměrně, aby odhalil chování jednotlivých metod, pokud jsou použity na soubor s malým rozsahem, kde každé chybné pozorování má zásadní vliv na celkovou kontaminaci souboru, a tak i na zkreslení odhadů regresních parametrů.

Data byla vygenerována za použití kódu

```
data SASUSER.DATA1;
  do x=1 to 10;
     y=10 + 3*x + rand('normal',0,2);
  output;
end;
```

Následné odhady již budou zpracovávány v programu SAS, který na rozdíl od programu R nabízí všechny námi zmíněné metody výpočtu odhadů regresních koeficientů. Před samotným provedením výpočtů si představíme, za pomoci kterých procedur budou odhady pořizovány, a uvedeme několik základních parametrů používaných pro specifikaci našich požadavků na proces odhadu koeficientů. Omezíme se však pouze na ty parametry, které budou v našich výpočtech přítomny, jelikož při obsáhlých možnostech programu SAS by popis všech možností

nastavení byl unavující a zbytečný, protože všechny informace jsou dostupné v nápovědě programu.

Pro klasickou regresi pomocí metody nejmenších čtverců bude použita procedura REG, u níž využijeme implicitního nastavení, které je pro obyčejnou MNČ dostačující. Jedinými volitelnými parametry v této proceduře budou:

DATA
nastavení datového souboru, ze kterého pochází hodnoty.

MODEL
požaduje specifikaci proměnných typu <závislá> = <nezávislé> < / volby>.

Procedura **ROBUSTREG** bude použita pro M-odhady, MM-odhady, S-odhady a LTS odhady. Její stěžejní parametry jsou následující:

DATA

PLOTS
slouží k vykreslení požadovaných grafů. Před procedurou obsahující tento příkaz musí být spuštěno prostředí ODS Graphics, po proceduře opět vypnuto(ods graphics on; ods graphics off;). Volbu grafů zapisujeme následovně plots <(požadované-grafy)>. K dispozici je celá řada grafických charakteristik, např. ddplot, fitplot, histogram, qqplot,...

METHOD - specifikuje metodu odhadu regresních koeficientů, která bude pro výpočet použita. Možnosti jsou takovéto:

METHOD=M

SCALE
nastavuje směrodatnou odchylku přímo či udává způsob jejího odhadu.

Možnosti tohoto parametru jsou následující.

Měřítko	Zápis	implicitní nastavení
konstanta	SCALE=hodnota	
Huberův odhad	SCALE=HUBER<(D=d)>	2.5
mediánový odhad	SCALE=MED	
Tukeyho odhad	SCALE=TUKEY<(D=d)>	2.5

Implicitní metoda výpočtu je MED. Jedná se o iterační algoritmus, kdy směrodatnou odchylku v $m+1$ kroku vypočteme pomocí vztahu

$$\hat{\sigma}^{(m+1)} = \underset{i}{\text{med}} \left\{ \frac{|y_i - \mathbf{x}_{i\cdot}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{(m)}|}{0.6745} \right\},$$

přičemž 0.6745 je 0.75-kvantil normovaného normálního rozdělení. U Huberova a Tukeyho odhadu jsou použity příslušné funkce, kde lze navíc stanovit parametr D.

WF
umožňuje volbu váhové funkce. SAS nabízí 10 voleb znázorněných níže, které lze dále ovlivňovat pomocí konstant A, B či C. Implicitní metodou M-odhadu je Andrewsova funkce.

Váhová funkce	Zápis	implicitní nastavení
Andrewsova	WF=ANDREWS<(C=c)>	1.339
Bisquare	WF=BISQUARE<(C=c)>	4.685
Cauchyho	WF=CAUCHY<(C=c)>	2.385
Fair	WF=FAIR<(C=c)>	1.4
Hampelova	WF=HAMPEL<(A=a B=b C=c)>	2,4,8
Huberova	WF=HUBER<(C=c)>	1.345
logistická	WF=LOGISTIC<(C=c)>	1.205
medián	WF=MEDIAN<(C=c)>	0.01
Talworthova	WF=TALWORTH<(C=c)>	2.795
Welschova	WF=WELSCH<(C=c)>	2.985

ASYMPCOV
volba způsobu pro výpočet kovarianční matice, zadáváme ASYMPCOV=H1|H2|H3. Dostupné vzorce se shodují se vztahy uvedenými u M-odhadů. Implicitně je nastaven první způsob výpočtu (12).

METHOD=LTS

H
parametr pro nastavení „usekávací“ konstanty h. Jako výchozí je nastavena hodnota $[\frac{3n+p+1}{4}]$, v důsledku čehož se v našich výpočtech dostáváme ke konstantě $h = 8$.

METHOD=LTS

ASYMPCOV
určuje možnosti pro výpočet kovarianční matice, zadáváme ASYMPCOV=H1|H2|H3|H4. První tři způsoby se shodují s volbami pro M-odhad, čtvrtá možnost je odhad kovarianční matice ve tvaru

$$\mathrm{K}^2 \frac{[1/(n-p)] \sum \psi(\varepsilon_i)^2}{[(1/n) \sum \psi'(\varepsilon_i)]^2} \mathbf{W}^{-1}. \tag{14}$$

Tento odhad je implicitní.

CHIF
specifikace funkce ρ, která bude použita při výpočtu S-odhadu. SAS nabízí dvě možnosti volby, a to Tukeyho 'bisquare' funkci a Yohaiho funkci. My budeme využívat funkci první, která je ve tvaru:

$$\rho(z) = \begin{cases} 3\left(\frac{z}{c}\right)^2 - 3\left(\frac{z}{c}\right)^4 + \left(\frac{z}{c}\right)^6 & \text{pro } |z| \leq c \\ 1 & \text{jinak} \end{cases}$$

Konstanta c kontroluje bod selhání a eficienci S-odhadu. Standardně je nastavena na hodnotu 2.9366, pro kterou je bod selhání 0.25 s odpovídající eficiencí 75,9%.

EFF
pomáhá kontrolovat eficienci pro S-odhad. Ovlivňuje konstantu c ve funkci ρ.

k0
slouží k určení konstanty c ve funkci ρ, a tím i k určení eficience a bodu selhání. Tento parametr má přednost před nastavenou hodnotou v parametru EFF.

METHOD=MM

ASYMPCOV
specifikuje způsob výpočtu kovarianční matice. Možnosti i výchozí nastavení jsou stejná jako u S-odhadu.

INITEST
umožňuje nastavit metodu odhadu regresních koeficientů v prvním kroku procesu. Možnosti jsou INITEST=LTS|S, kde LTS-odhad je výchozí volbou.

CHIF
parametr určující funkci, která bude použita v druhé fázi výpočtu MM-odhadu. V případě, že INITEST=S, nastavuje též funkci použitou pro S-odhad v první fázi. K dispozici je opět Tukeyho 'bisquare' funkce (výchozí volba) a Yohaiho funkce.

k0
slouží k určení konstanty c ve funkci ρ a tím i k určení eficience a bodu selhání. Její nastavení platí pro první i druhou fázi, pokud je INITEST=S.

EFF
specifikuje eficienci pro MM-odhad. Implicitní nastavení eficience je 0.85, což odpovídá nastavení k1=3.44 pro CHIF=TUKEY a k1=0.868 pro CHIF=YOHAI. Konstanta k1 vystupuje ve funkcí druhé fáze výpočtu, pokud je INITEST=LTS.

INITH
určuje konstantu h v případě, že INITEST=LTS.

Další procedurou použitou v příkladové části je IML procedura, v níž se budeme obracet na funkci **LAV Call** a **LMS Call**. Funkce LAV Call slouží k odhadu regresních koeficientů za použití metody L_1-odhadů, zde se ovšem setkáváme se zkratkou LAV (Least absolute value, v překladu nejmenší absolutní hodnota). Tato funkce je volána příkazem ve tvaru call lav(rc, xr, a, b, x0, opt) a vyžaduje specifikaci následujících parametrů:

rc, xr

zajišťují detailní specifikaci optimalizačního procesu a výstupů s ním souvisejících.

a
je matice plánu $\mathbf{X}_{n\times p}$, pro kterou platí, že $h(\mathbf{X}) = p$ a $n \geq p$.

b
zde zastupuje vektor $\mathbf{y}_{n\times 1}$.

x0
je volitelný vektor rozměrů $n \times 1$, který udává počáteční bod v optimalizační úloze.

opt
je volitelný vektor, kde zadáváme, jaké výstupy chceme ve výsledku zobrazit. V našich příkladech využijeme volbu opt={ . 3 0 1 }, čímž se myslí vektor opt=(0,3,0,1). Tím jsme stanovili, že ve výstupu chceme mít odhady metodou nejmenších čtverců, L_1-odhady, odhad kovarianční matice (za pomoci McKean-Schraderova odhadu) a test konvergence.

LMS Call funguje podobně jako LAV Call, jen má mnohem více možností nastavení vektoru opt, tudíž opět uvedeme pouze naše nastavení tohoto vektoru a přiblížíme, co jednotlivé volby znamenají. Funkci zavoláme příkazem call lms(sc, coef, wgt, opt, b, a) a předem nadefinujeme parametry:

sc, coef, wgt
zajišťují detailní specifikaci optimalizačního procesu a výstupů s ním souvisejících.

a
je matice plánu $\mathbf{X}_{n\times p}$, pro kterou platí, že $h(\mathbf{X}) = p$ a $n \geq p$. Pokud je ovšem opt[1]=0, předpokládá se v regresní funkci přítomnost absolutního členu a matice $\mathbf{X}$ se zadává již bez sloupce jedniček.

b
zde zastupuje vektor $\mathbf{y}_{n\times 1}$.

opt
je volitelný vektor, kde zadáváme, jaké výstupy chceme ve výsledku zobrazit. V našich příkladech využijeme volbu opt=(0,2,0,0,0,0,0,0). Tím jsme stanovili, že ve výstupu chceme mít LMS-odhad, výčet reziduí, historii op-

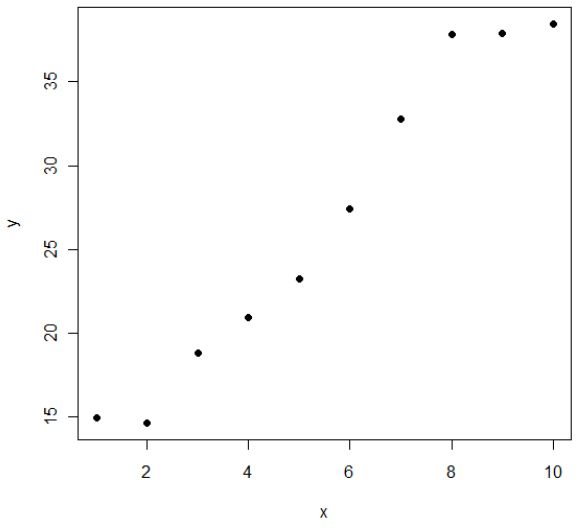
timalizačního procesu a další charakteristiky iteračního procesu, jako váhy pro RLS (vážená metoda nejmenších čtverců). [21]

6.1. Příklad 1

První soubor dat byl vygenerován za účelem, aby v případě odhadování regresních parametrů byla ideální metoda nejmenších čtverců (a metoda maximální věrohodnosti, která zde bude poskytovat stejné výsledky). Robustní metody tak budou podrobeny zkoumání jejich vydatnosti. Data jsou uvedena v tabulce 2 a graficky znázorněna na obrázku 9.

x	**y**	**x**	**y**
1	14.9320	6	27.4376
2	14.6175	7	32.7958
3	18.8207	8	37.7968
4	20.9537	9	37.9031
5	23.2528	10	38.4805

Tabulka 2: Datový soubor 1



Obr. 9: Zobrazení datového souboru 1

Výpočty budou prováděny v programu SAS 9.2, který nabízí řešení všech robustních metod, které byly přiblíženy na předchozích stránkách. Kód programu zajišťující výpočet regresních odhadů je uveden v příloze 1. Výsledky jsou zapsány v tabulce 3.

Podle výsledků se jako vhodná (ve smyslu, že odhady jsou porovnatelné s pravými parametry přímky) jeví většina metod. Nejméně se pravým hodnotám regresních koeficientů blíží LAD-odhad, LMS-odhad ši LTS-odhad, naopak nejlépe se prezentuje pravděpodobně M-odhad s použitím Huberovy funkce.

Toliko k posouzení přesnosti jednotlivých metod - toto hodnocení bylo prováděno na základě znalosti správné regresní funkce. Nyní je tedy na místě ukázat si, jak zvolit vhodnou regresní metodu v případě neznalosti správných hodnot koeficientů. K tomu nám slouží diagnostika odlehlých pozorování či charakteristiky uvedené v předchozí kapitole. Diagnostiku odlehlých pozorování provádí SAS v rámci výpočtu M-odhadu a stačí k tomu do programu vložit některé příkazy

```
proc robustreg data=SASUSER.DATA2 method=m
    plots=(ddplot);
    model Y = X / diagnostics leverage;
run;
```

Identifikace odlehlých pozorování v programu SAS je prováděna následujícím způsobem. Nejprve si přibližme, jak jsou nalezena odlehlá pozorování ve směru osy x. Ve vzorcích je takové pozorovnání označováno jako leverage a platí pro něj

$$\text{LEVERAGE} = \begin{cases} 0 & \text{pro } \mathrm{RD}(x_i) \leq \sqrt{\chi^2_{p,1-\alpha}} \\ 1 & \text{jinak} \end{cases},$$ kde $\mathrm{RD}(x_i)$ značí robustní vzdálenost. Jedná se o vzdálenost odvozenou z Mahalanobisovy vzdálenosti $\mathrm{MD}(x_i)$, která je definována jako

$$\mathrm{MD}(x_i) = [(x_i - \bar{x})^{\mathrm{T}} \, \widehat{\Sigma}^{-1} \, (x_i - \bar{x})]^{\frac{1}{2}},$$

za předpokladu, že $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ a $\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^{\mathrm{T}} (x_i - \bar{x})$ jsou empirické mnohorozměrné charakteristiky polohy (v tomto případě aritmetický průměr

a výběrová kovarianční matice). Robustní vzdálenost potom vznikne nahrazením těchto charakteristik polohy za jejich robustní ekvivalenty $\tilde{x}$ a $\widetilde{\Sigma}$, které jsou vypočteny pomocí MCD metody (minimum covariance determinant, v překladu minimální determinant kovarianční matice, viz. [19]). Robustní vzdálenost bude ve tvaru

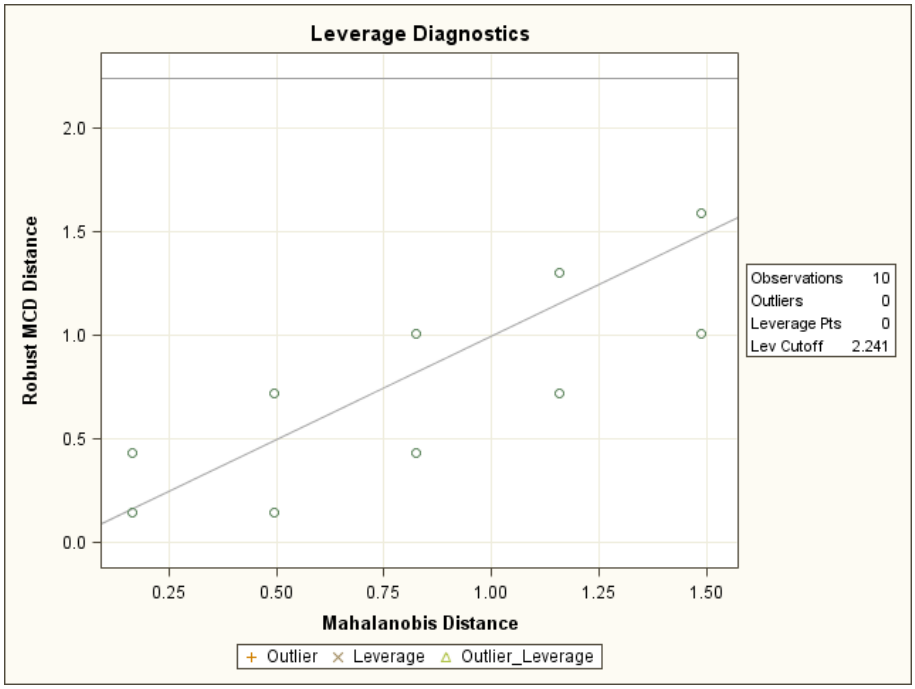
$$\mathrm{RD}(x_i) = [(x_i - \tilde{x})^{\mathrm{T}}\ \widetilde{\Sigma}^{-1}\,(x_i - \tilde{x})]^{\frac{1}{2}}.$$

Odlehlá pozorování ve směru osy y jsou detekována pomocí reziduí a ve vzorci bude takové pozorování označeno jako outlier. Odlehlé pozorování je identifikováno na základě vztahu

$$\mathrm{OUTLIER} = \begin{cases} 0 & \text{pro } |\varepsilon_i| \leq k\hat{\sigma} \\ 1 & \text{jinak} \end{cases}.$$

Zde $\hat{\sigma}$ je odhadem měřítka a implicitní nastavení hodnoty konstanty k je $k = 3$. [21]

Pro data z tabulky 2 byla tato analýza provedena a výsledkem je obrázek 10, který vypovídá, že soubor neobsahuje odlehlé hodnoty.



Obr. 10: Diagnostika odlehlých pozorování

Informace z této analýzy nás vedou k závěru, že u odhadů vycházejících z těchto dat nebudeme požadovat vysoký bod selhání, nýbrž vysokou eficienci. Tyto vlastnosti lze ovlivnit u S-odhadu, kde máme možnost nastavení buď konstanty c či přímo požadované vydatnosti odhadu. Zadáme tedy požadovanou eficienci rovnu jedné

```
proc robustreg data=SASUSER.DATA1 method=s (EFF= 1);
      model Y = X;
run;
```

Čímž získáme nové odhady parametrů a vyrovnaná přímka je nyní tvaru $\widehat{y} = \mathbf{9.6719} + \mathbf{3.0652}x$. Bod selhání se tím sníží na 0.1983, naproti standardnímu 0.25, to by však v naší situaci nemělo být problémem, jelikož žádná odlehlá pozorování nebyla identifikována.

Vypočtené hodnoty charakteristik usuzujících na kvalitu odhadu regresní přímky pro příklad 1 jsou k dispozici v tabulce 3. R^2 značí index determinace (pro metodu nejmenších čtverců bude vypočítán klasický R^2, pro ostatní metody uvedeme R^2_{rob} poskytnutý programem SAS) a označením SE(β_0) a SE(β_1) rozumíme odhad směrodatné odchylky pro odhady parametrů β_0 a β_1. SE je zkratkou z anglického názvu standard error, definujícího směrodatnou odchylku odhadů.

Některé z charakteristik program SAS ve výstupu neposkytuje, proto budou v následující tabulce vynechány. Konkrétně se jedná o odhady směrodatných odchylek pro odhady parametrů metodou LTS a LMS a o robustní index determinace pro LAD-odhad.

Metoda	$\widehat{\beta_0}$	$\widehat{\beta_1}$	SE (β_0)	SE (β_1)	R^2
MNČ	9.7148	3.0880	1.3200	0.2127	0.9634
LAD-odhad	7.9645	3.3265	2.2499	0.3626	–
M-odhad (Huber)	9.7591	3.0679	1.3920	0.2243	0.9239
M-odhad (Hampel)	9.7148	3.0880	1.3200	0.2127	0.7149
M-odhad (Andrews)	9.6884	3.0788	1.4118	0.2275	0.8845
LTS-odhad	8.7661	3.1333	–	–	0.9783
LMS-odhad	9.5230	2.8784	–	–	0.9893
S-odhad (EFF=1)	9.6719	3.0652	1.5385	0.2518	0.9650
S-odhad (EFF=0.76)	9.6668	3.0516	1.6718	0.2750	0.9657
MM-odhad	9.6753	3.0691	1.5002	0.2451	0.8028

Tabulka 3: Odhady regresní přímky pro datový soubor 1

Podle charakteristik v tabulce 3 není možné o nějakém odhadu prohlásit, že je výrazně nejvhodnější pro daná data. Index determinace je nejvyšší pro LMS-odhad či S-odhad s implicitně danou eficiencí. Musíme zde ale připomenout, že robustní indexy determinace jsou pro různé metody počítány různě, a proto se spíše hodí pro porovnávání funkcí v rámci použití jedné metody a nikoliv pro porovnávání více metod mezi sebou. S přihlédnutím k tomuto faktu a skutečnosti, že chyby jsou normálně rozděleny, můžeme prohlásit, že nejlepší volbou je metoda nejmenších čtverců.

Přímky jsou znázorněny v příloze 2. Již z prvního grafu je zřejmé, že žádná z přímek se výrazně neodchyluje od správného směru. Přímky jsou vykresleny velmi těsně u sebe a často se i překrývají. Další porovnávání umožňuje zbytek grafů v příloze, kde se jen potvrzuje, že odhadnuté přímky jsou velmi blízké správné zadané přímce.

Pro M a MM-odhady je navíc možné porovnání na základě robustního Akaikeho a Bayesovského informačního kritéria, což je znázorněno v tabulce 4.

	AICR	BICR
M-odhad (Huber)	9.2471	11.2682
M-odhad (Hampel)	50.1498	51.8550
M-odhad (Andrews)	9.0887	11.1282
MM-odhad	7.8159	10.2342

Tabulka 4: Informační kritéria pro př. 1

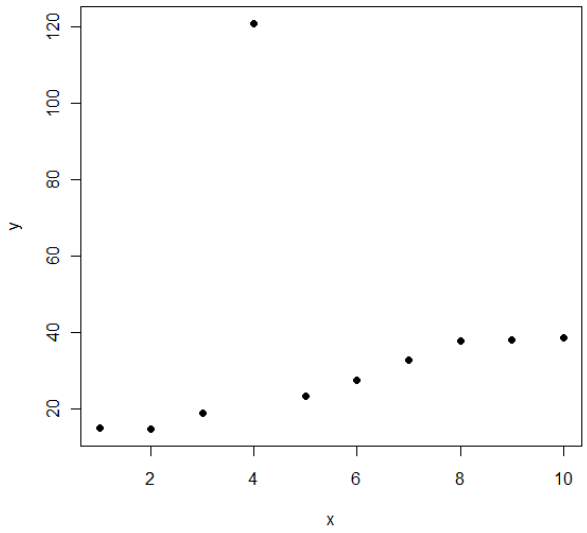
Podle těchto kritérií vyhodnotíme jako nejlepší MM-odhad, který nabývá nejmenších hodnot pro AICR a BICR charakteristiky.

6.2. Příklad 2

Druhý příklad bude obdobou příkladu prvního, pouze u jediného pozorování změníme hodnotu závisle proměnné tak, abychom vytvořili odlehlé pozorování ve směru osy y. Výsledné odhady parametrů opět vypočítáme v programu SAS 9.2 za pomoci stejného kódu, který obsahuje příloha 1. Přepsán bude pouze název datové množiny, která poskytuje potřebné vygenerované hodnoty. Důvod k zahrnutí tohoto příkladu do práce je následující - příklad by měl odhalit, jak se chovají jednotlivé metody za předpokladu malé kontaminace dat ve směru závisle proměnné. Data jsou uvedena v tabulce 5 a znázorněna na obrázku 11.

x	**y**	**x**	**y**
1	14.93197	6	27.43756
2	14.6175	7	32.79583
3	18.82073	8	37.79679
4	120.9537	9	37.90308
5	23.25283	10	38.48045

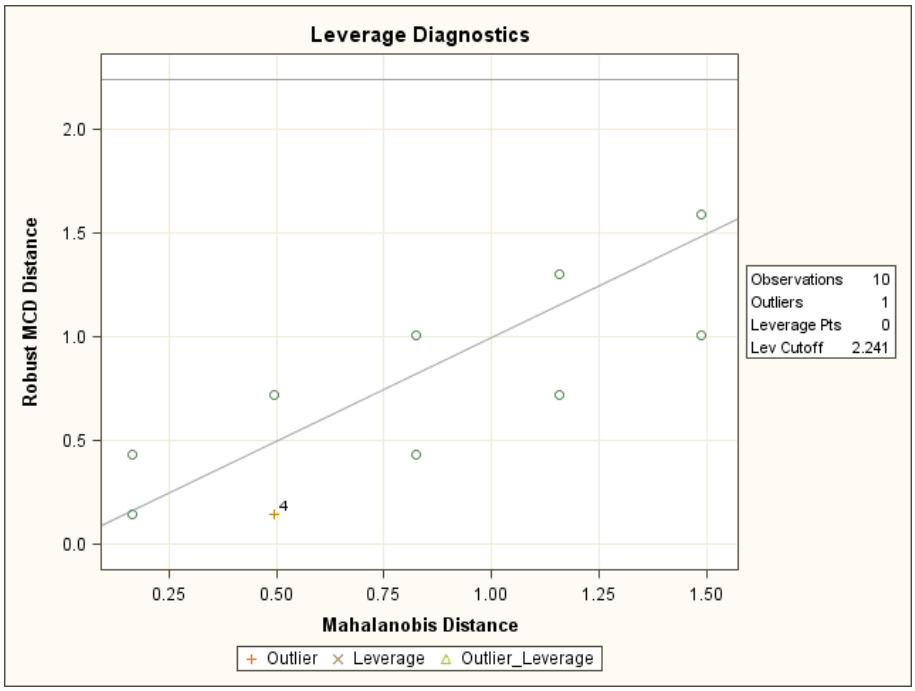
Tabulka 5: Datový soubor 2



Obr. 11: Zobrazení datového souboru 2

Zřetelné odlehlé pozorování můžeme pozorovat ve čtvrtém řádku. Výsledné odhady jsou uvedeny v tabulce 6. Dle očekávání se při kontaminaci souboru odhad metodou nejmenších čtverců stává téměř nepoužitelným díky silnému zkreslení. To se značně napraví, pokud čtvrté pozorování v souboru vynecháme - v tabulce 6 tento odhad označíme MNČ (vyn.). Ostatní odhady si již počínají mnohem lépe, všechny jsou velmi blízké správným hodnotám regresních parametrů a žádný odhad vyloženě nevyčnívá. Též je těžké určit, který z odhadů je nejlepší, protože rozdíly mezi odhadnutými parametry jsou opravdu malé.

Obrázek 12 stejně jako v předešlém příkladě znázorňuje identifikaci odlehlých pozorování provedenou programem SAS. Správně bylo za odlehlé pozorování ve směru závisle proměnné označeno pozorování čtvrté.



Obr. 12: Diagnostika odlehlých pozorování

Na základě této informace je možné změnit konstantu h pro LTS-odhad, protože máme podezření, že 9 pozorování bude v pořádku. Pokud $h = 9$, potom kód bude vypadat následujícím způsobem

```
proc robustreg data=SASUSER.DATA2 method=lts (H=9);
      model Y = X;
run;
```

Odhadnutá regresní přímka pomocí metody LTS je nyní tvaru $\widehat{y} = \mathbf{9.9699} + \mathbf{3.0648}x$, kde směrnice je odhadnuta s podobnou přesností jako u původní verze odhadu, průsečík s osou y se ovšem výrazně zlepšil. Podobně můžeme zlepšit i MNČ odhad, vynecháme-li odlehlé pozorování 4 ze souboru dat, viz tabulka 6.

Pro případ, kdy neznáme správné hodnoty parametrů, jsou v tabulce 6 uvedeny charakteristiky kvality regresní funkce, podle kterých se můžeme rozhodnout, které metodě výpočtu odhadů regresních parametrů dáme přednost.

Metoda	$\widehat{\beta_0}$	$\widehat{\beta_1}$	SE (β_0)	SE (β_1)	R^2
MNČ	29.7148	1.2699	22.312	3.5960	0.0153
MNČ(vyn.)	9.9699	3.0648	1.4432	0.2254	0.9636
LAD-odhad	9.2796	3.1804	2.4659	0.3974	–
M-odhad (Huber)	10.7443	2.9944	1.7679	0.2849	0.3369
M-odhad (Hampel)	9.9699	3.0648	1.4632	0.2358	0.5946
M-odhad (Andrews)	9.9652	3.0582	1.5446	0.2489	0.7935
LTS-odhad (h=9)	9.9699	3.0648	–	–	0.9635
LTS-odhad (h=8)	10.1882	2.9473	–	–	0.9747
LMS-odhad	9.2593	3.0455	–	–	0.9737
S-odhad	9.9636	3.0519	1.6058	0.2528	0.9599
MM-odhad	9.9648	3.0567	1.5464	0.2429	0.7372

Tabulka 6: Odhady regresní přímky pro datový soubor 2

Na základě těchto údajů by jako přijatelná metoda pro výpočet regresních koeficientů byl pravděpodobně vybrán LTS-odhad (h=9) či M-odhad (Hampel) na základě malého rozptylu koeficientů. Reziduální součet se zde stává nepodstatným ukazatelem, jelikož nepožadujeme nejlepší proložení všemi body, ale správné zachycení závislosti na základě nekontaminovaných dat. Index determinace nabývá největších hodnot pro LMS-odhad či LTS-odhad (h=8).

Odhadnuté přímky jsou znázorněny v příloze 3. Z prvního grafu lze vyčíst, že kromě metody nejmenších čtverců jsou všechny odhady velmi přesné a navzájem srovnatelné. Další grafy tuto domněnku jen potvrzují.

	AICR	BICR
M-odhad (Huber)	111.0602	112.7411
M-odhad (Hampel)	66.5482	69.1385
M-odhad (Andrews)	13.4426	16.1166
MM-odhad	7.9491	11.3205

Tabulka 7: Informační kritéria pro př. 2

Tabulka 7 by nám měla usnadnit výběr mezi M a MM-odhady. Podle indexu determinace je nejpřesnější M-odhad s použítím Andrewsovy funkce následovaný MM-odhadem, podle informačních kritérií se dostáváme k opačné situaci.

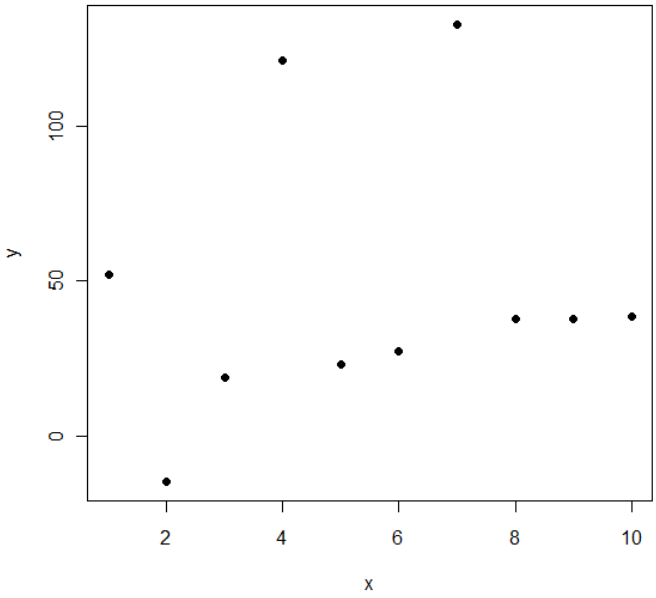
6.3. Příklad 3

Na třetím příkladě podrobíme zkoumání bod selhání námi probíraných metod. Pro tento účel použijeme nový soubor dat, který bude kontaminovaný čtyřmi odlehlými pozorováními ve směru osy y. Tyto špatné hodnoty budou tedy tvořit závažnou kontaminaci, se kterou by si měly poradit především robustní metody s vysokým bodem selhání. Data jsou uvedena v tabulce 8 a zakreslena na obrázku 13.

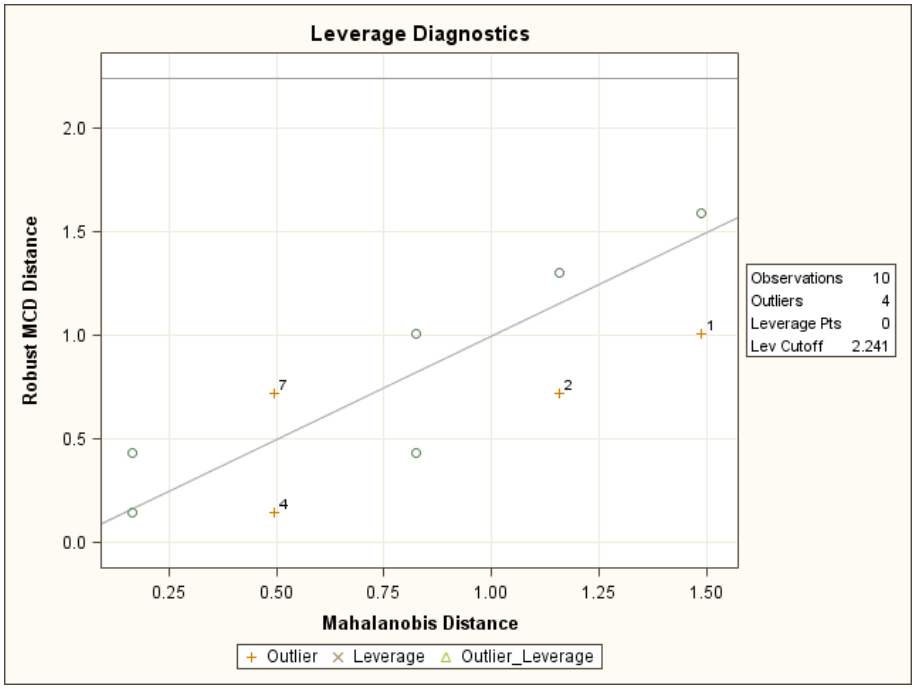
x	**y**	**x**	**y**
1	51.93197	6	27.43756
2	−14.6175	7	132.79583
3	18.82073	8	37.79679
4	120.9537	9	37.90308
5	23.25283	10	38.48045

Tabulka 8: Datový soubor 3

Nejprve se pokusíme detekovat odlehlá pozorování pomocí diagnostiky provedené programem SAS, jako tomu bylo v předchozích příkladech. Výsledkem bude obrázek 14, podle kterého je možné identifikovat všechna odlehlá pozorování.



Obr. 13: Zobrazení datového souboru 3



Obr. 14: Diagnostika odlehlých pozorování

Odhalení čtyř odlehlých pozorování nás vede k tomu, abychom u LTS-odhadu nastavili konstantu h rovnu šesti, což zapíšeme analogicky jako v předchozím příkladě. Bez této aktualizace by konstanta h byla rovna osmi a vypočtené koeficienty by byly $\widehat{y} = \mathbf{15.4395} + \mathbf{2.2157}x$.

Zároveň budeme od S-odhadu požadovat, aby bod selhání přesáhl 0.4 (4 pozorování z 10 považujeme za kontaminovaná), čehož dosáhneme snížením eficeince. Na základě několika náhodně zadaných hodnot není těžké zjistit, že již pro eficienci 0.45 je bod selhání roven 0.4054. Koeficienty určené pomocí S-odhadu s těmito parametry jsou uvedeny ve výčtu výsledků všech odhadů, pro zajímavost si však prozraďme, že za použití implicitního nastavení hodnoty eficience by odhad regresní přímky vypadal následovně $\widehat{y} = \mathbf{26.5210} + \mathbf{1.8296}x$.

Podobně lze ovlivnit bod selhání též u MM-odhadu, kde jsme doposud vycházeli z implicitného nastavení, nyní ovšem vytvoříme další dva MM-odhady úpravou odhadu počátečních parametrů, což je první krok procesu při výpočtu MM-odhadu. První vylepšený odhad zapíšeme pomocí kódu

```
proc robustreg data=SASUSER.DATA3 method=mm (INITEST=S k0=1.9617
     EFF=0.45);
     model Y = X;
run;
```

Parametr INITEST rozhoduje, která metoda bude v prvním kroce použita, v tomto případě jsme stanovili S-odhad. Konstanta k0 vystupuje v druhé fázi výpočtu, kde si zvolíme příslušnou funkci, která bude vystupovat v M-odhadu směrodatné odchylky. V našem případě se jedná o Tukeyho funkci, jejíž tvar byl uveden již v příkladě 1. Konstantou je možné ovlivnit výslednou eficienci a bod selhání, který zde požadujeme vyšší než 0.4, konstantu k0 tedy zvolíme rovnu 1.9617. Třetí fáze obsahuje opět volbu z dvou funkcí a znovu volíme Tukeyho, tentokrát však nelze zadat hodnotu konstanty, musíme stanovit námi požadovanou eficienci 0.45, která vyhovuje našemu cíli dosáhnout bodu selhání přes 0.4.

Druhý upravený odhad bude vycházet z implicitně nastavené metody LTS-odhadu a do programu ho zapíšeme jako

```
proc robustreg data=SASUSER.DATA3 method=mm (INITH=6 k0=1.9617
     EFF=0.45);
     model Y = X;
run;
```

Kde hodnota pro INITH udává, s jakou konstantou h bude odhad LTS pracovat. Konstantu k0 udáváme stejnou jako v prvním upraveném MM-odhadu, stejně to platí pro eficienci v třetím kroku.

Bez těchto úprav bychom dostali MM-odhad regresní přímky ve tvaru $\widehat{y} = \mathbf{30.3077}+\mathbf{1.8818}x$, což není přesně to, co bychom od robustní metody s vysokým bodem selhání čekali, jelikož parametry nejsou ani zdaleka blízké pravé hodnotě koeficientů.

Očividně jsou všechny neupravené odhady výrazně zkreslené v porovnání se svými protějšky, které jsme byli schopni na základě informací z obrázku 14 patřičně přeformulovat. Všechny odhadnuté hodnoty regresních parametrů jsou zapsány v tabulce 9.

Po úpravě S-odhadu se tato metoda jeví v tomto příkladě nejrobustnější, velmi uspokojivý odhad poskytuje také metoda LMS, LTS, LAD či MM-odhad. Přijatelně by se také daly hodnotit odhady pořízené M-odhadem za použití Andrewsovy funkce či odhady metodou nejmenších čtverců, která předpokládá vynechání kontaminovaných dat v datovém souboru. Zbytek M-odhadů je silně zkreslený, nejvíce ovlivněné kontaminací se potom jeví parametry odhadnuté metodou nejmenších čtverců.

Nejlépe by v této sekci měly působit odhady s vysokým bodem selhání, což se potvrdilo. Nepřesnost M-odhadů a metody nejmenších čtverců byla očekávaná.

Opět je na místě jako v přechozích příkladech uvést základní charakteristiky při porovnávání jednotlivých regresních funkcí (viz tabulka 9).

Metoda	$\widehat{\beta_0}$	$\widehat{\beta_1}$	SE (β_0)	SE (β_1)	R^2
MNČ	34.7698	2.3102	32.573	5.2497	0.0236
MNČ(vyn.)	8.9513	3.1703	2.7349	0.3775	0.9463
LAD-odhad	9.2796	3.1804	50.0580	8.0676	–
M-odhad (Huber)	29.4744	1.3558	21.4137	3.4511	0.0104
M-odhad (Hampel)	0.1644	4.2386	9.6512	1.5554	0.0743
M-odhad (Andrews)	8.9521	3.1664	2.0460	0.3297	0.3431
LTS-odhad (h=6)	8.9513	3.1703	–	–	0.9463
LTS-odhad (h=8)	15.4395	2.2157	–	–	0.1368
LMS-odhad	10.7289	3.0455	–	–	0.9243
S-odhad (EFF=0.76)	26.5210	1.8296	32.0844	5.0512	0.0000
S-odhad (EFF=0.45)	9.3988	3.0431	14.7385	2.1952	0.0000
MM-odhad	30.3077	1.8818	33.0654	5.2470	0.0159
MM-odhad (S)	9.3990	3.0431	14.7386	2.1952	0.0373
MM-odhad (LTS)	9.4377	3.0381	14.7431	2.1961	0.0373

Tabulka 9: Odhady regresní přímky pro datový soubor 3

Podle těchto výsledků by se při neznalosti pravého tvaru přímky nejvěrohodněji jevil LTS odhad za použití „usekávací“ konstanty $h = 6$, jehož směrodatná odchylka odhadnutých regresních parametrů je velmi malá a zároveň má vysokou hodnotu robustního indexu determinace. Vysoký index determinace můžeme vidět též u LMS-odhadu. Velmi dobré SE pro odhady regresních koeficientů vykazuje též M-odhad za použití Andrewsovy funkce, též má nejvyšší index determinace mezi všemi M-odhady.

Grafy odhadnutých přímek jsou uvedeny v příloze 4. Již první graf ukazuje, že odhady přímek se mezi sebou liší výrazněji, než tomu bylo u předešlých příkladů. Detailně to lze zkoumat na zbývajících grafech.

	AICR	BICR
M-odhad (Huber)	30.9791	31.1627
M-odhad (Hampel)	112.9228	113.1939
M-odhad (Andrews)	30.5154	34.3974
MM-odhad	7.8214	10.6310
MM-odhad (S)	4.5789	8.7651
MM-odhad (LTS)	4.5780	8.7637

Tabulka 10: Informační kritéria pro př. 3

Podle informačních kritérií je z třídy M a MM-odhadů nejpřesnější upravený MM-odhad za použití LTS-odhadu v první fázi, následovaný druhým upraveným MM-odhadem.

Na konec této podkapitoly si uveďme, že i M-odhady, které jsou primárně odhady s nízkým bodem selhání, mohou být naprogramovány tak, že i při vysoké kontaminaci podávají relativně dobré a nezkreslené výsledky. Dosáhnout toho lze pomocí nastavení konstant, které vystupují ve funkcích ρ. Implicitní hodnoty konstant zajišťují eficienci 0,95 a jsou uvedeny na začátku kapitoly, kde byly představeny použité procedury a jejich parametry.

Program SAS ve výsledcích u M-odhadů neuvádí, jaká je eficience a bod selhání použitého odhadu, tudíž není možné konstanty nastavovat podobně, jako je to možné u S-odhadů či MM-odhadů, kde stačí upravovat hodnotu konstant, dokud bod selhání nedosáhne požadované velikosti. Proto bude proveden experiment zabývající se vlivem změny konstant na přesnost odhadu regresních koeficientů. Každá váhová funkce bude použita pro odhad koeficientů čtyřikrát, pokaždé s jiným nastavením konstant.

Zápis do programu provedeme takto:

```
proc robustreg method=m(wf=andrews) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=andrews(c=1)) data=SASUSER.DATA3;
     model Y = X;
```

```
run;
proc robustreg method=m(wf=andrews(c=0.7)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=andrews(c=0.4)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=hampel) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=hampel(a=1.5 b=3 c=6)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=hampel(a=1 b=2 c=4)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=hampel(a=0.5 b=1 c=2)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=huber) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=huber(c=1)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=huber(c=0.7)) data=SASUSER.DATA3;
     model Y = X;
run;
proc robustreg method=m(wf=huber(c=0.4)) data=SASUSER.DATA3;
     model Y = X;
run;
```

Výsledné odhady pro přehlednost zobrazíme v tabulce 11.

	$\widehat{\boldsymbol{\beta}}_0$	$\widehat{\boldsymbol{\beta}}_1$
Andrews (c=1.339)	8.9521	3.1664
Andrews (c=1)	8.9544	3.1629
Andrews (c=0.7)	8.9667	3.1525
Andrews (c=0.4)	9.1659	3.0645
Huber (c=1.345)	29.4744	1.3558
Huber (c=1)	23.5676	1.7087
Huber (c=0.7)	12.2959	2.8359
Huber (c=0.4)	11.2795	2.8929
Hampel (a=2 b=4 c=8)	0.1644	4.2386
Hampel (a=1.5 b=3 c=6)	8.9513	3.1703
Hampel (a=1 b=2 c=4)	8.9513	3.1703
Hampel (a=0.5 b=1 c=2)	9.0633	3.1100

Tabulka 11: Vliv konstant váhových funkcí na odhad regresních koeficientů

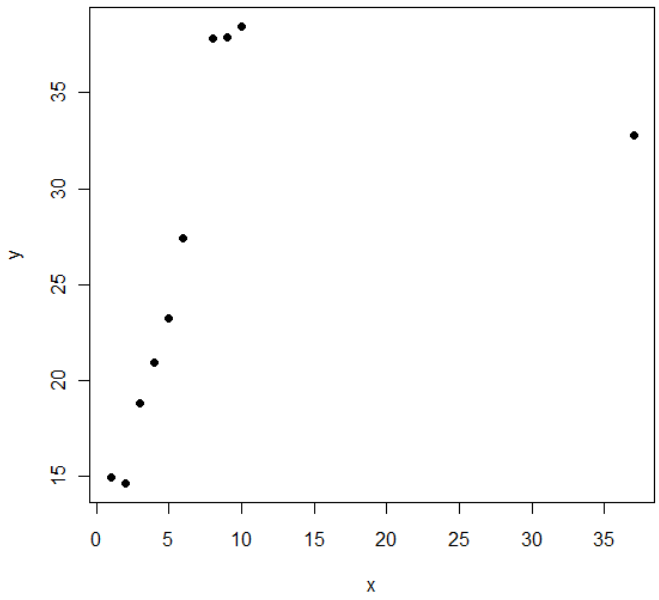
Dle těchto výsledků lze předpokládat, že se snižujícími se konstantami bod selhání roste, což usuzujeme podle odhadů koeficientů, které se postupně přibližují správným parametrům regresní přímky. Tato domněnka vychází také z tabulky 1, kde je z rozsahu váhových funkcí očividné, že jak snižujeme konstanty, odlehlá pozorování jsou penalizována a jsou jim přiděleny menší (až nulové) váhy.

6.4. Příklad 4

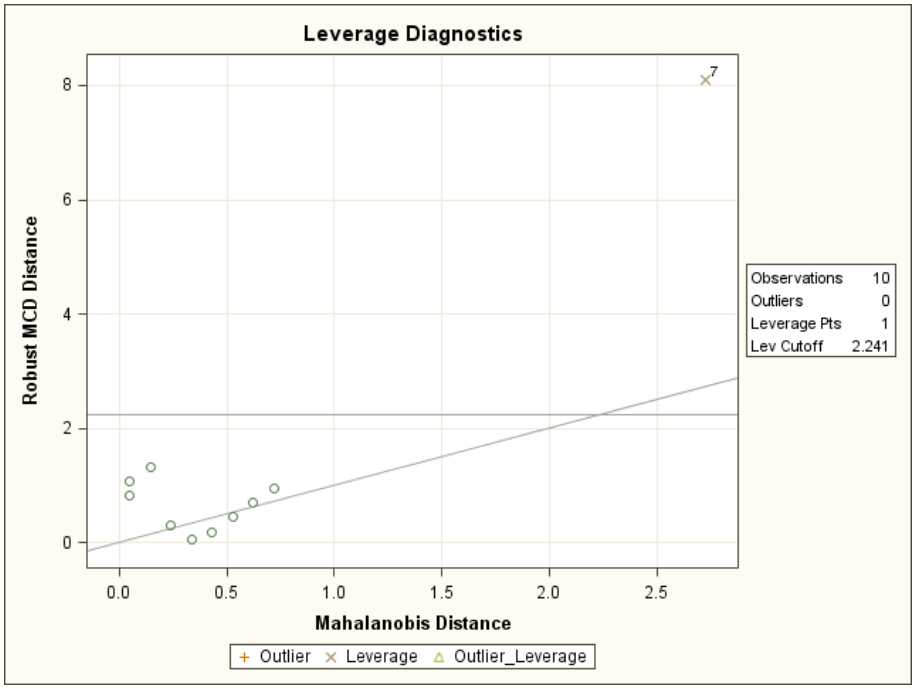
Nakonec je třeba prozkoumat chování odhadů v případě odlehlého pozorování ve směru osy x. Data si můžeme prohlédnout v tabulce 12 a na obrázku 15.

x	**y**	**x**	**y**
1	14.93197	6	27.43756
2	14.6175	37	32.79583
3	18.82073	8	37.79679
4	20.9537	9	37.90308
5	23.25283	10	38.48045

Tabulka 12: Datový soubor 4



Obr. 15: Zobrazení datového souboru 4



Obr. 16: Diagnostika odlehlých pozorování

Jako vždy se budeme chtít přesvědčit, že SAS toto odlehlé pozorování odhalí. Že tomu tak je, vidíme na obrázku 16.

Ponechme tentokrát výpočet odhadů regresních parametrů na programu SAS a jeho předem daných hodnotách některých parametrů, protože robustnost vůči odlehlému pozorování ve směru osy x nelze nijak výrazně ovlivnit např. vyšším bodem selhání, o který jsme usilovali v předchozím příkladě. Díky tomu se dostáváme k výčtu výsledných odhadů.

Naprosto dle očekávání lze z výsledných odhadů vyčíst, že LAD-odhad, M-odhady, a metoda nejmenších čtverců jsou jediným špatným pozorováním velmi zkresleny a odhady jsou tedy nepoužitelné. Naproti tomu odhady s vysokým bodem selhání se prezentují dostatečně přesnými hodnotami. Také metoda nejmenších čtverců po vynechání odlehlého pozorování poskytuje uspokojivé odhady. Teoretické předpoklady vyslovené v kapitole o robustních odhadech jsou tedy potvrzeny.

Metoda	$\widehat{\beta_0}$	$\widehat{\beta_1}$	SE (β_0)	SE (β_1)	R^2
MNČ	22.9126	0.4455	3.67210	0.28122	0.2388
MNČ(vyn.)	9.7148	3.0575	1.3518	0.2212	0.9649
LAD-odhad	21.7617	0.2982	6.2453	0.4783	–
M-odhad (Huber)	22.9126	0.4455	3.6721	0.2812	0.2388
M-odhad (Hampel)	22.9126	0.4455	3.6721	0.2812	0.0337
M-odhad (Andrews)	22.7420	0.4415	3.9636	0.3035	0.2330
LTS-odhad (h=8)	10.0166	2.9066	–	–	0.9761
LMS-odhad	9.5230	2.8784	–	–	0.9893
S-odhad	9.7150	3.0272	1.4155	0.2345	0.9323
MM-odhad	9.7113	3.0390	1.3938	0.2300	0.7382

Tabulka 13: Odhady regresní přímky pro datový soubor 4

Na základě vysokých hodnot SE bychom zajisté nevybrali jako vhodnou metodu LAD-odhad, stejně by tomu bylo i u M-odhadů a metody nejmenších čtverců, kde je tento závěr podpořen také nízkými hodnotami indexu determinace. Ostatní

odhady se jeví značně věrohodně, především potom MM-odhad, S-odhad či LTS-odhad.

Odhady přímek jsou znázorněny v příloze 5. První graf poskytuje dostatečnou představu o tom, jakou změnu závislosti odlehlé pozorování ve směru osy x způsobilo u některých odhadnutých přímek. Bohužel se některé vychýlené přímky opět překrývají, tento nedostatek však vynahradí následující grafy pro jednotlivé odhady, kde je již zřetelně vidět, která přímka vybočuje ze správného směru.

	AICR	BICR
M-odhad (Huber)	7.0152	9.6160
M-odhad (Hampel)	47.0152	49.6160
M-odhad (Andrews)	7.2668	9.7492
MM-odhad	7.8783	11.3285

Tabulka 14: Informační kritéria pro př. 4

Informační kritéria jako nejlepší volbu nabízí M-odhad za použití Huberovy funkce, ačkoliv díky naší znalosti pravých hodnot parametrů víme, že odhady touto metodou pořízené jsou prakticky nepoužitelné.

Závěr

V práci byl položen teoretický základ různých metod odhadů regresních parametrů. Zvláštní prostor byl věnován běžně nejvíce používané metodě nejmenších čtverců, která byla představena též z geometrického hlediska. Dále zde byly uvedeny základní robustní metody odhadu a jejich elementární principy.

Jednotlivé metody byly použity při hledání odhadů parametrů regresní přímky. Výsledky byly analyzovány pro různě kontaminované malé datové soubory. Konkrétně se jednalo o čtyři situace, a to datový soubor bez odlehlých pozorování, soubor dat s malou kontaminací ve směru osy y, třetí datový soubor byl již kontaminován o poznání více, a testoval tak bod selhání různých metod, a čtvrtý příklad podrobil metody zkoumání, nakolik jsou schopny pracovat s malou kontaminací ve směru osy x.

Všechny výpočty byly prováděny v programu SAS 9.2, který mimo jiné poskytuje diagnostiku odlehlých pozorování, jež by nám při neznalosti poškození datového souboru značně pomohla při nastavování parametrů jednotlivých metod ovlivňujících např. eficienci a bod selhání. Ve výpočetní části byly též uvedeny charakteristiky hodnotící kvalitu modelu, které by při neznalosti pravé závislosti měly hrát hlavní roli při výběru vhodné metody. Bohužel robustní indexy determinace fungovaly lépe v rámci jedné metody, ne však jako obecný ukazatel přesnosti odhadů, a směrodatné odchylky odhadů byly značně ovlivněny volbou parametrů, takže při zvyšování bodu selhání tyto ukazatelé občas nabývaly značně vysokých hodnot, přestože odhad koeficientů se naopak zlepšil. Ukazuje se tedy, že pro výběr vhodné metody odhadu je zapotřebí hlubší analýzy jednotlivých metod, což by ovšem výrazně překročilo rozsah bakalářské práce. Na základě našich znalostí můžeme metodu odhadu volit např. dle jejího bodu selhání či obecných vlastností.

Příloha 1

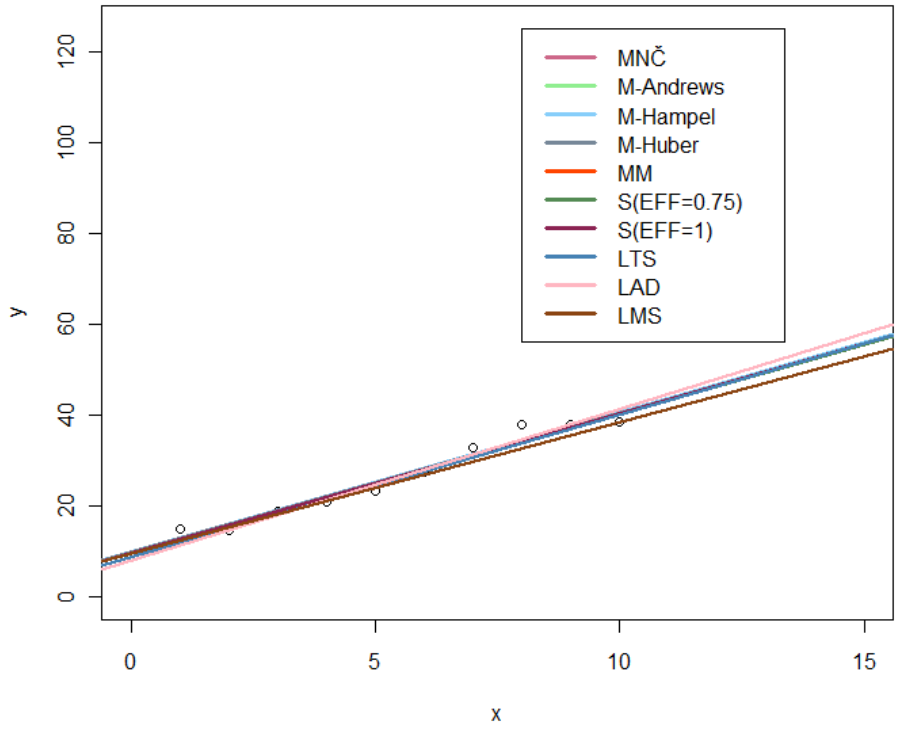
Zápis programu pro výpočet odhadů parametrů v příkladu 1 v prostředí SAS 9.2.

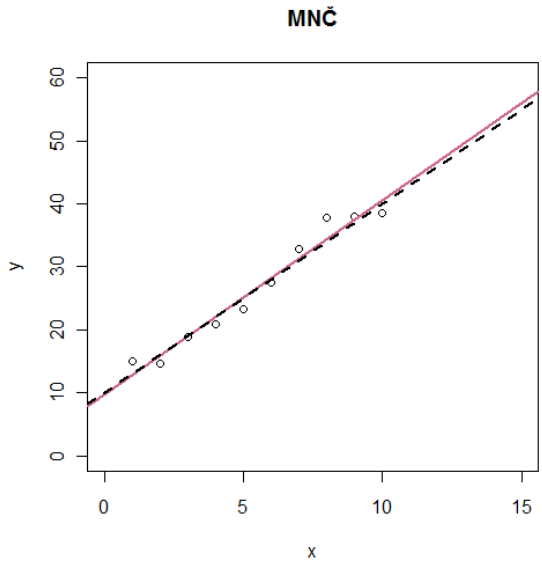
```
/*Least squares*/
proc reg data=SASUSER.DATA1;
      model Y = X ;
run;

/*M - estimator (Andrews)*/
proc robustreg method=m(wf=andrews) data=SASUSER.DATA1;
      model Y = X;
run;

/*M - estimator (Hampel)*/
proc robustreg method=m(wf=hampel) data=SASUSER.DATA1;
      model Y = X;
run;

/*M - estimator (Huber)*/
proc robustreg method=m(wf=huber) data=SASUSER.DATA1;
      model Y = X;
run;

/*MM - estimator*/
proc robustreg data=SASUSER.DATA1 method=mm;
      model Y = X;
run;

/*S - estimator*/
proc robustreg data=SASUSER.DATA1 method=s ;
      model Y = X;
run;

/*Least trimmed squares*/
proc robustreg data=SASUSER.DATA1 method=lts ;
      model Y = X;
run;

/*Least absolute values*/
proc iml ;
```

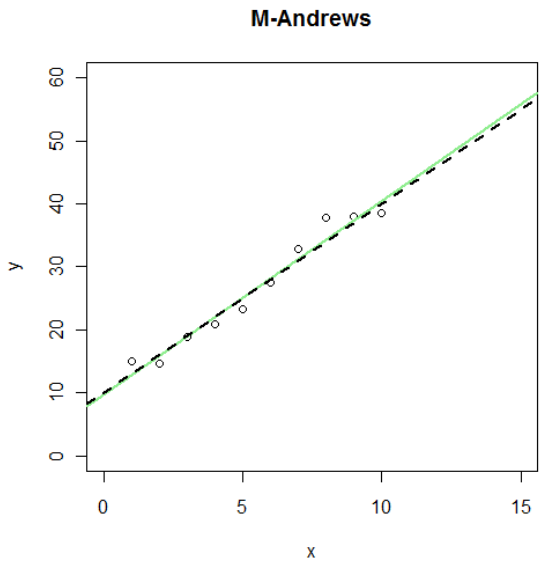
```
  use SASUSER.DATA1;
  read all;
  a = x;
  m = nrow(a);
  a = j(m,1,1.) || a;
  b = y;
  opt= { . 3  0 1 };
  call lav(rc,xr,a,b,,opt);

/*Least median of squares*/
proc iml ;
  use SASUSER.DATA1;
  read all;
  a = x;
  b = y;
  opt = j(8,1,0);
  opt[2]= 2;
  call lms(sc, coef, wgt, opt, b, a);
```

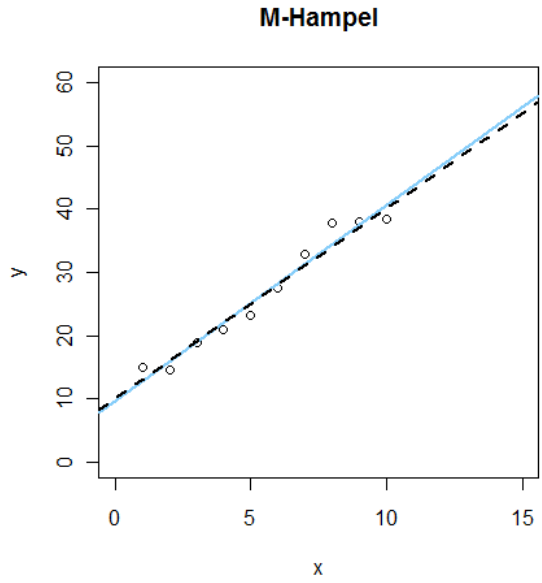
Příloha 2

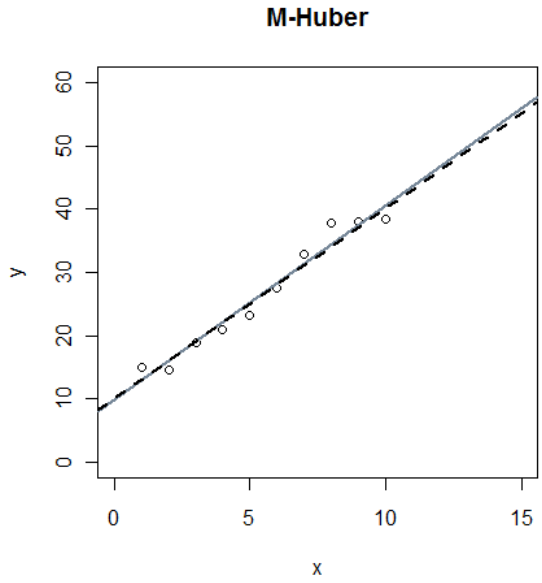
První graf obsahuje všechny odhadnuté přímky, které se ovšem díky malým ochylkám často překrývají. Lepší představu o jednotlivých odhadech regresních koeficientů získáme z následujících grafů, které porovnávají vždy jednu odhadnutou regresní přímku (vykreslenou barevně) se správnou přímkou, jejíž parametry známe a chceme se jim co nejvíce přiblížit (vykreslena čárkovaně).

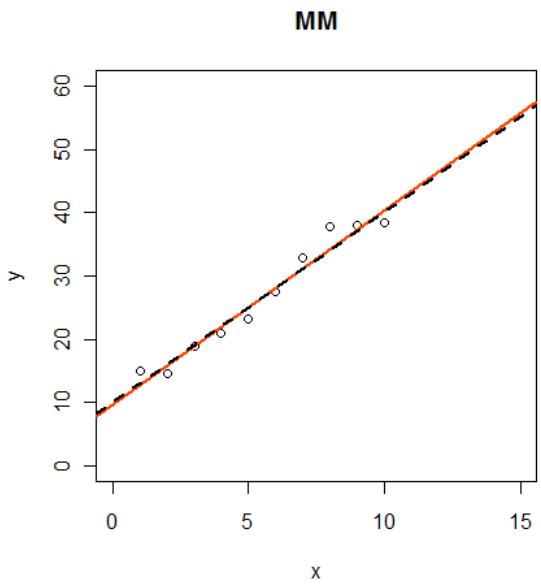


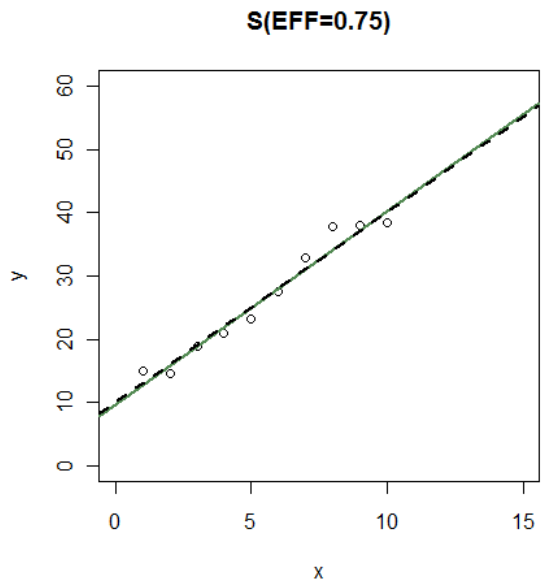

MNČ
y
x
0
10
20
30
40
50
60
0
5
10
15

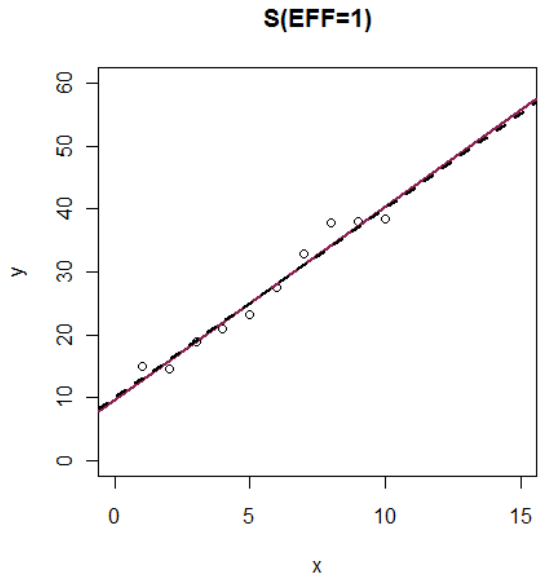

M-Andrews
y
x
0
10
20
30
40
50
60
0
5
10
15

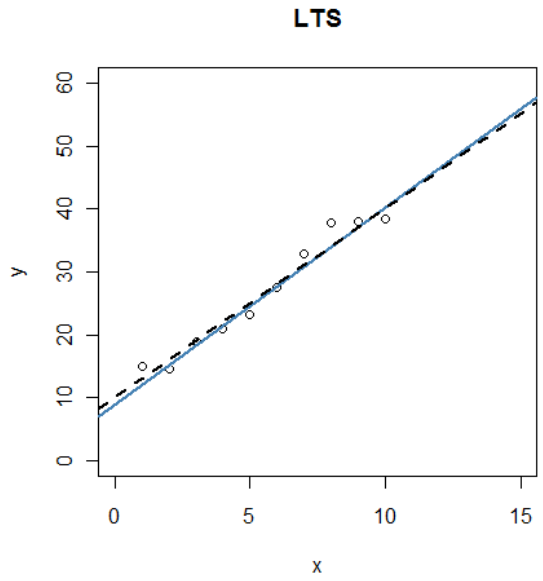

M-Hampel
y
x
0
10
20
30
40
50
60
0
5
10
15

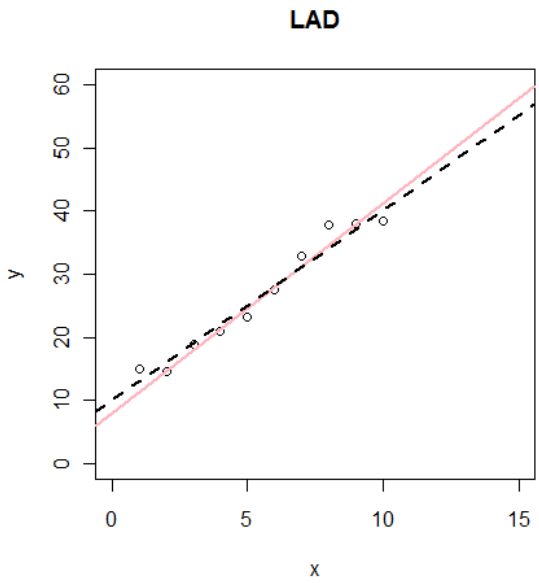

M-Huber
y
x
0
10
20
30
40
50
60
0
5
10
15

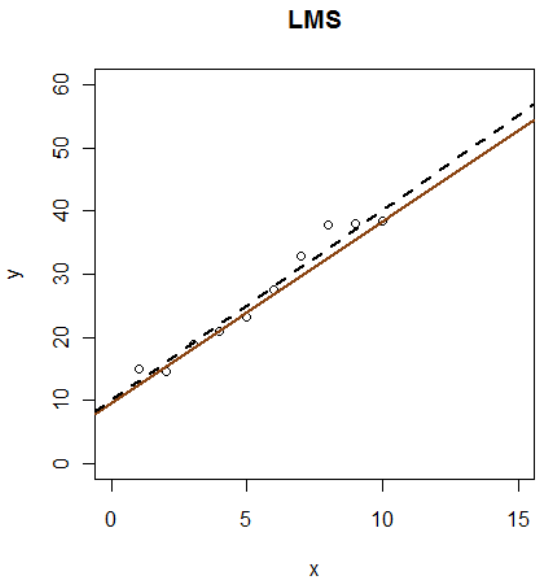

MM
y
x
0
10
20
30
40
50
60
0
5
10
15


S(EFF=0.75)
y
x
0
10
20
30
40
50
60
0
5
10
15

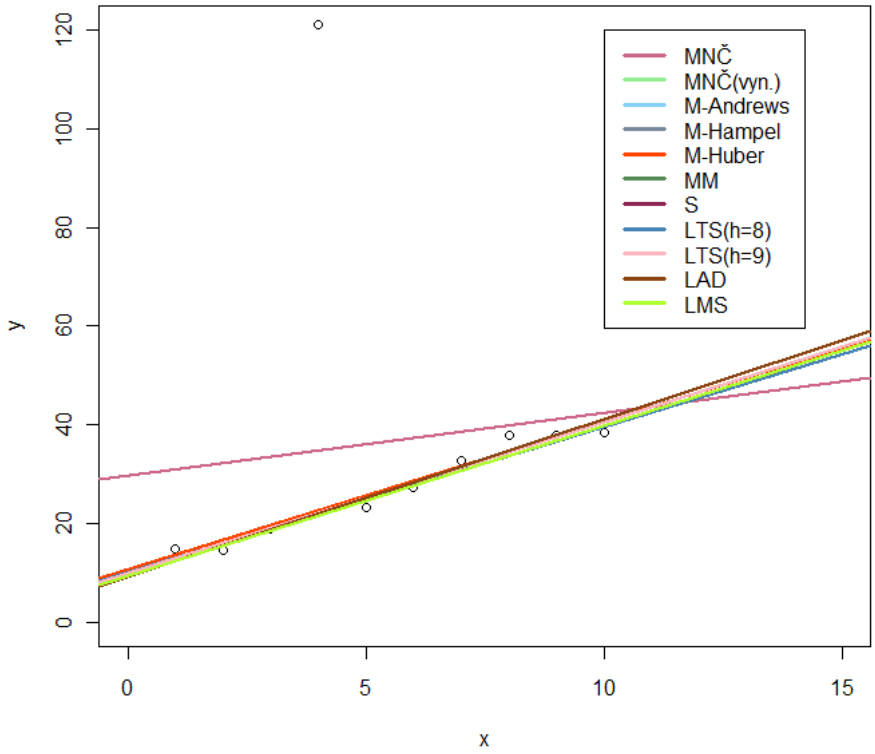

S(EFF=1)
y
x
0
10
20
30
40
50
60
0
5
10
15

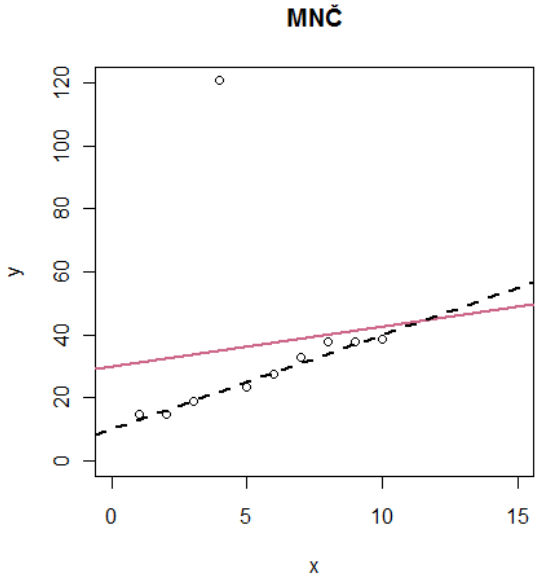

LTS
y
x
0
10
20
30
40
50
60
0
5
10
15

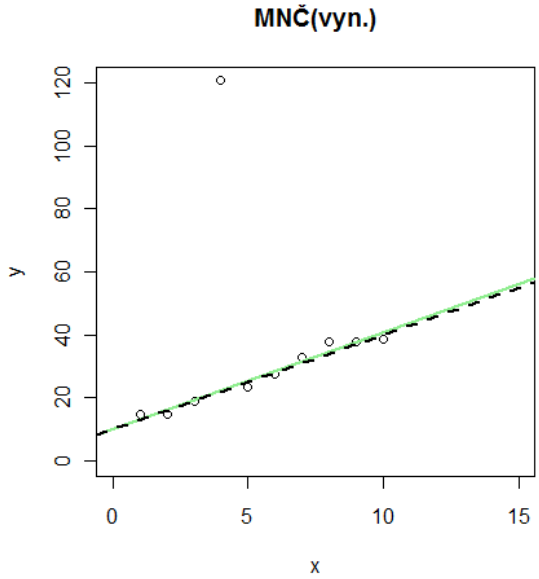

LAD
y
x
0
10
20
30
40
50
60
0
5
10
15

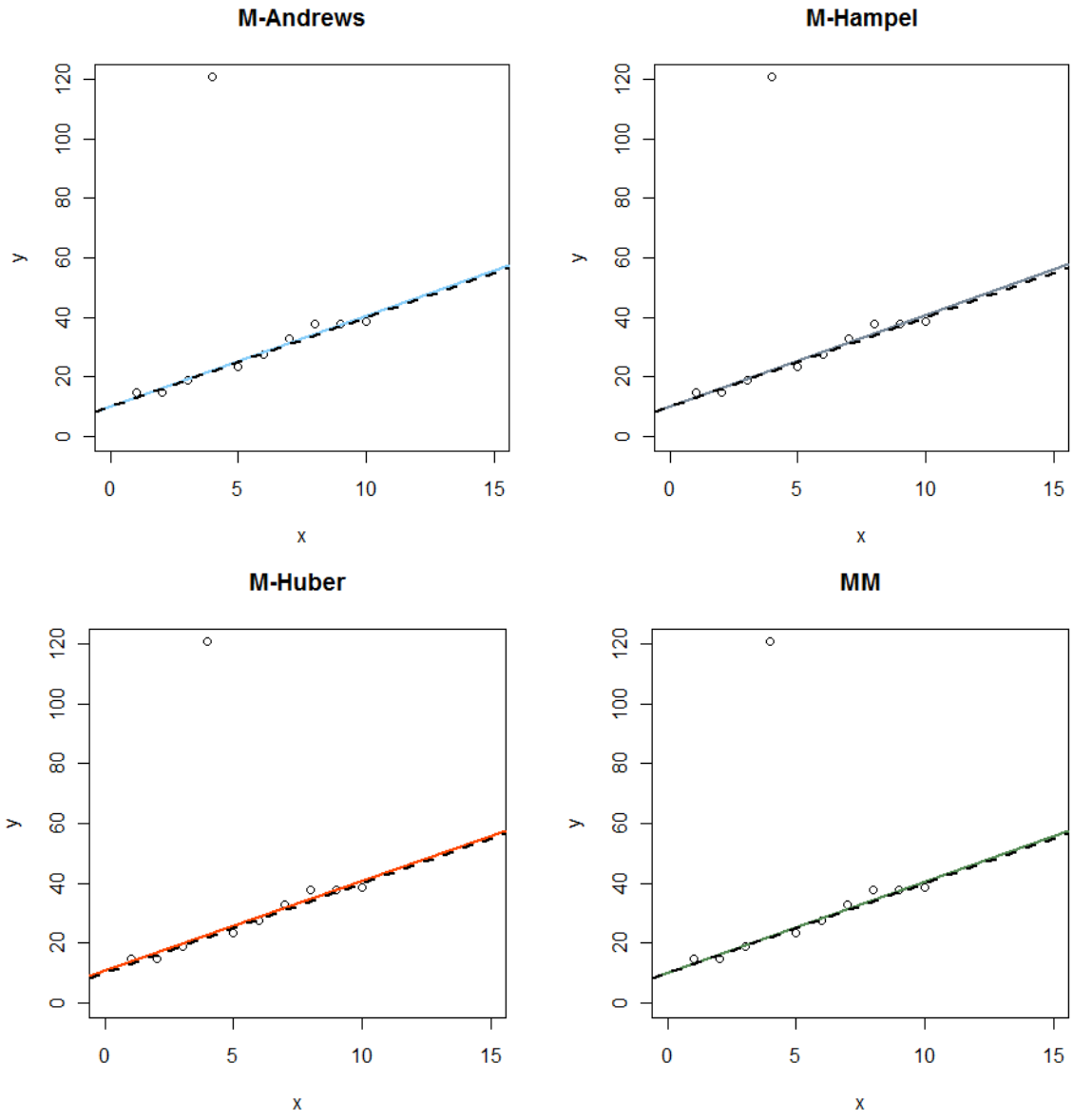

LMS
y
x
0
10
20
30
40
50
60
0
5
10
15

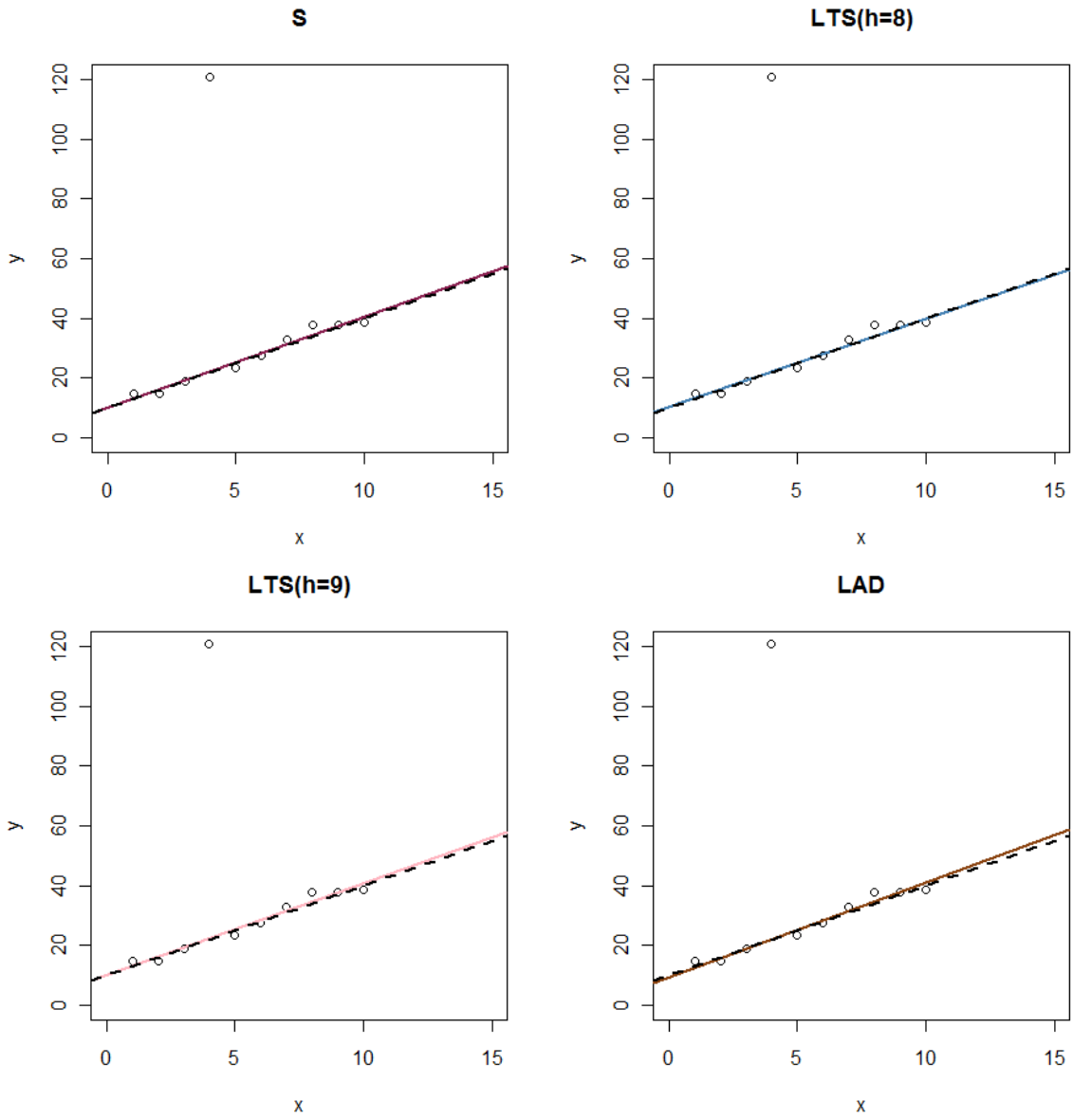
Příloha 3

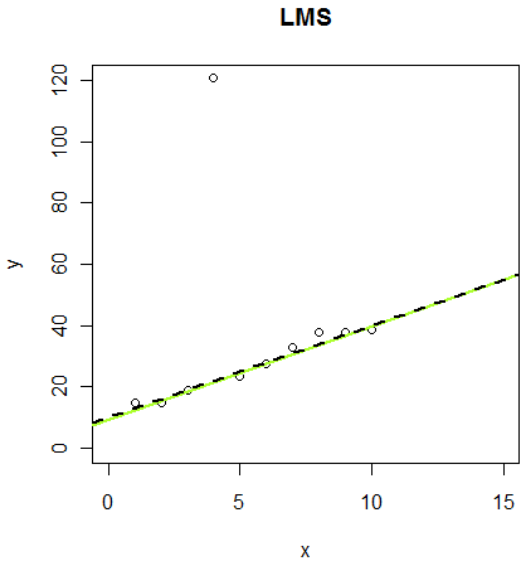

MNČ
MNČ(vyn.)
M-Andrews
M-Hampel
M-Huber
MM
S
LTS(h=8)
LTS(h=9)
LAD
LMS
y
x
0
20
40
60
80
100
120
0
5
10
15


MNČ
y
x
0
20
40
60
80
100
120
0
5
10
15

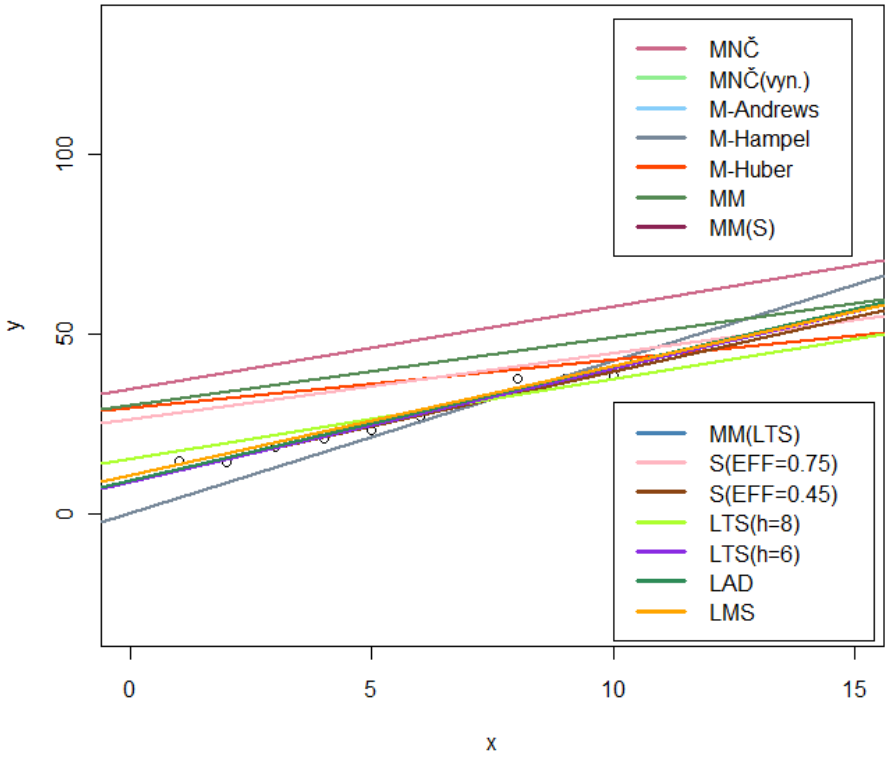

MNČ(vyn.)
y
x
0
20
40
60
80
100
120
0
5
10
15

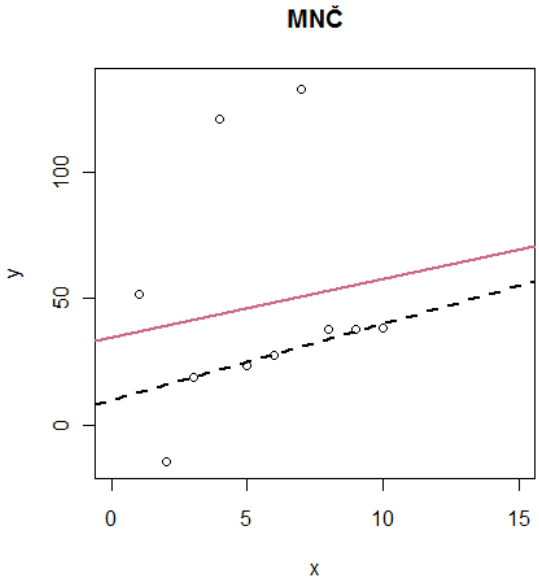

M-Andrews
M-Hampel
M-Huber
MM
y
x
0
20
40
60
80
100
120
0
5
10
15

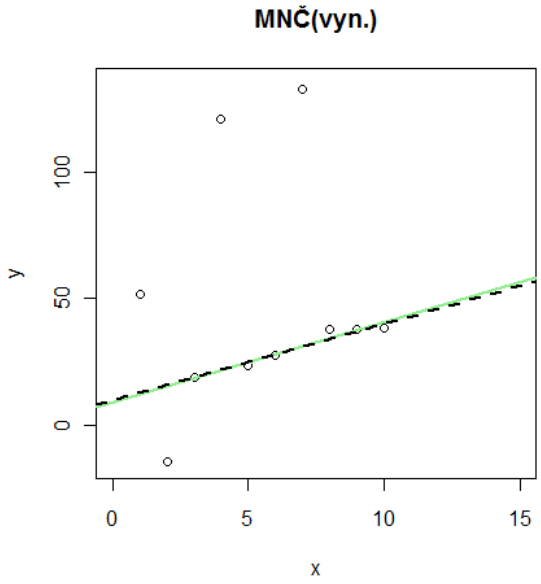

S
LTS(h=8)
LTS(h=9)
LAD
y
x
0
20
40
60
80
100
120
0
5
10
15

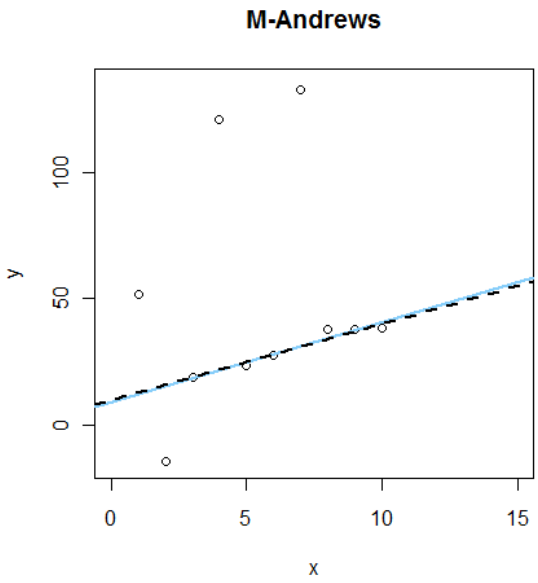

LMS
y
x
0
20
40
60
80
100
120
0
5
10
15

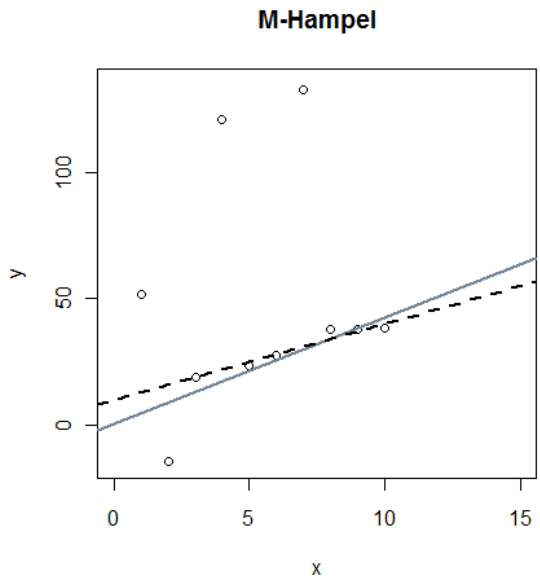
Příloha 4

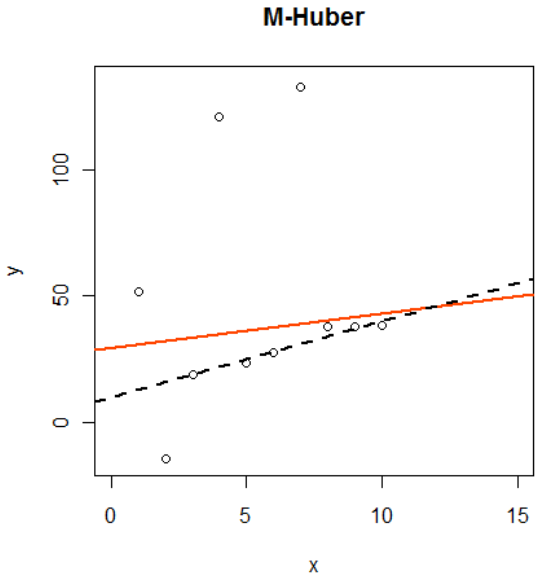

MNČ
MNČ(vyn.)
M-Andrews
M-Hampel
M-Huber
MM
MM(S)
MM(LTS)
S(EFF=0.75)
S(EFF=0.45)
LTS(h=8)
LTS(h=6)
LAD
LMS
y
x
0
50
100
0
5
10
15

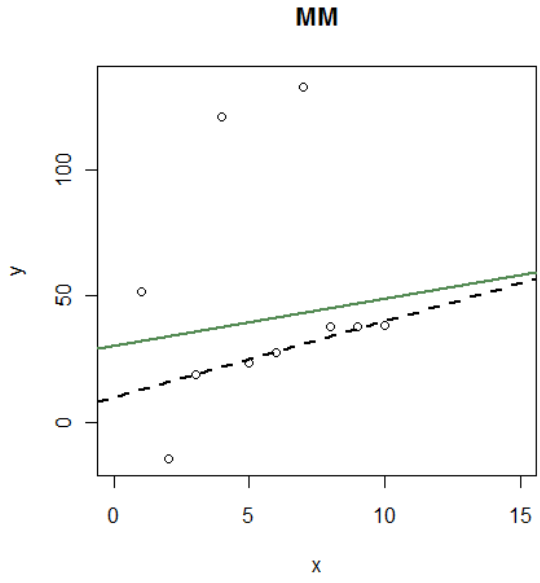

MNČ
y
x
0
50
100
0
5
10
15

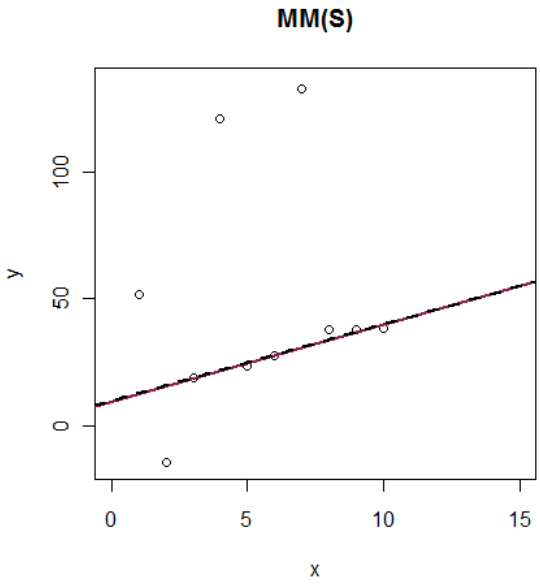

MNČ(vyn.)
y
x
0
50
100
0
5
10
15

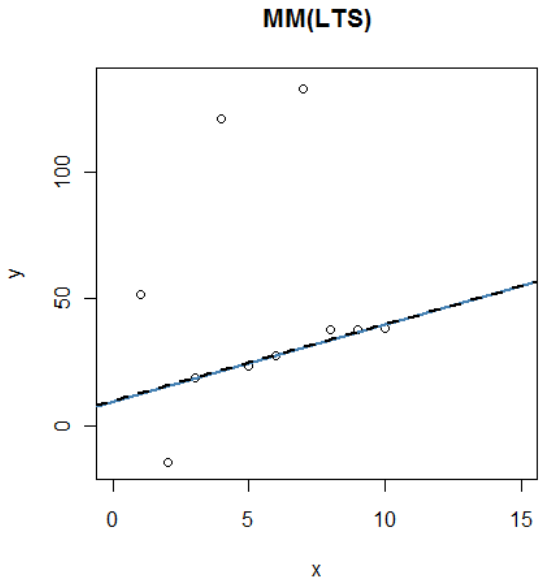

M-Andrews
y
x
0
50
100
0
5
10
15

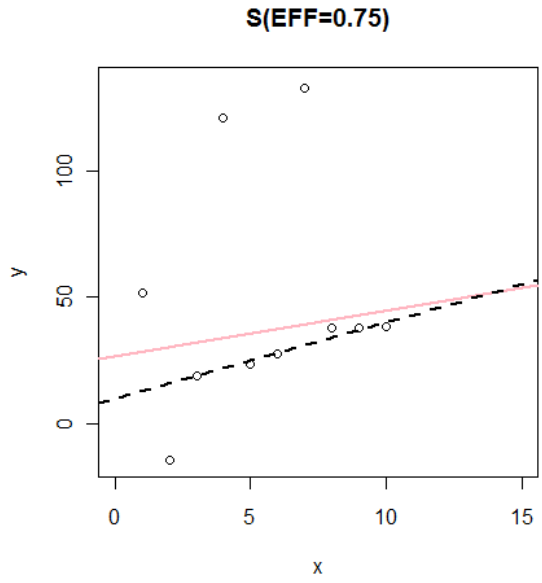

M-Hampel
y
x
0
50
100
0
5
10
15

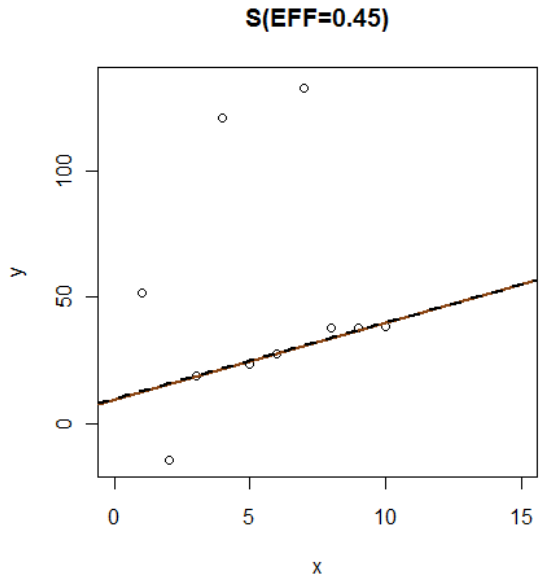

M-Huber
y
x
0
50
100
0
5
10
15

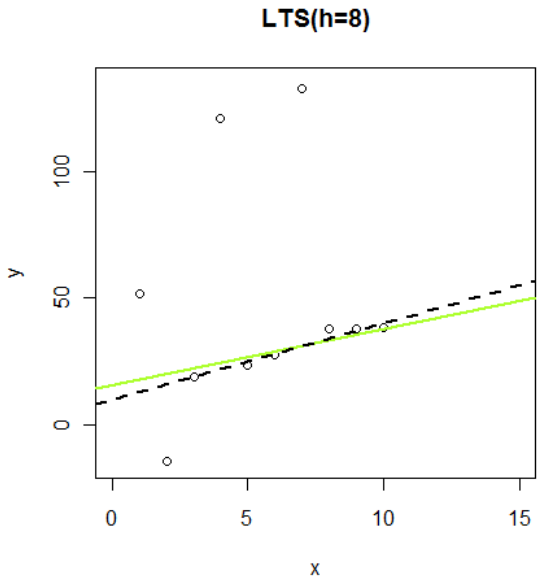

MM
y
x
0
50
100
0
5
10
15

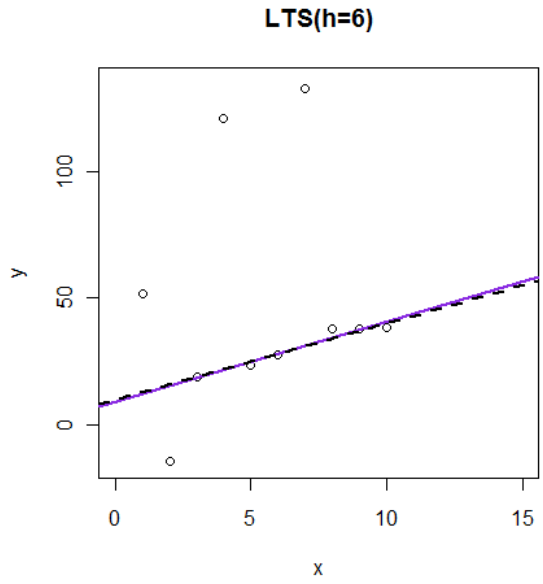

MM(S)
y
x
0
50
100
0
5
10
15

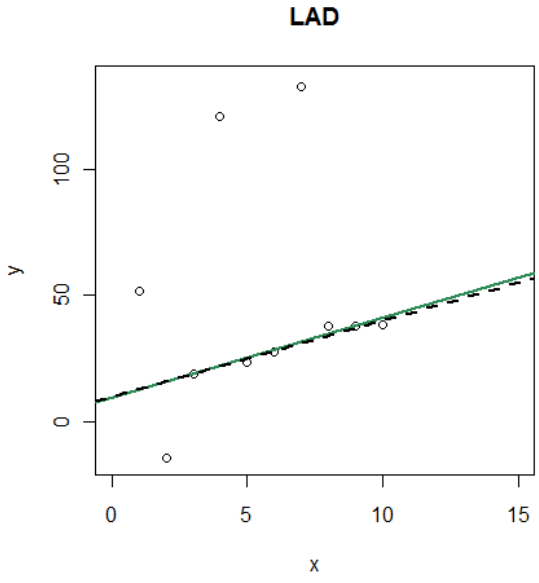

MM(LTS)
y
x
0
50
100
0
5
10
15

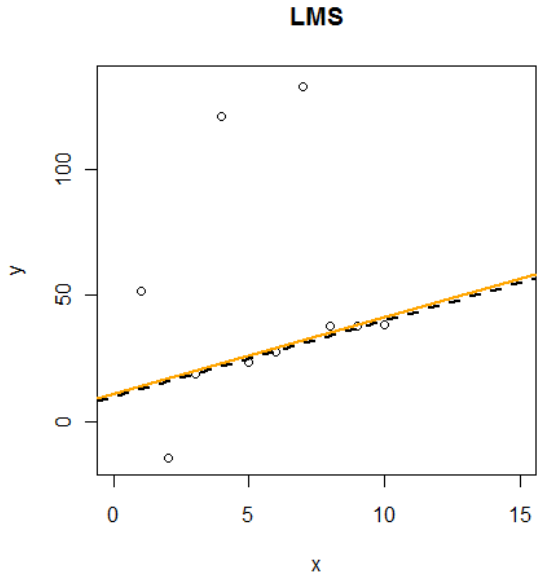

S(EFF=0.75)
y
x
0
50
100
0
5
10
15


S(EFF=0.45)
y
x
0
50
100
0
5
10
15

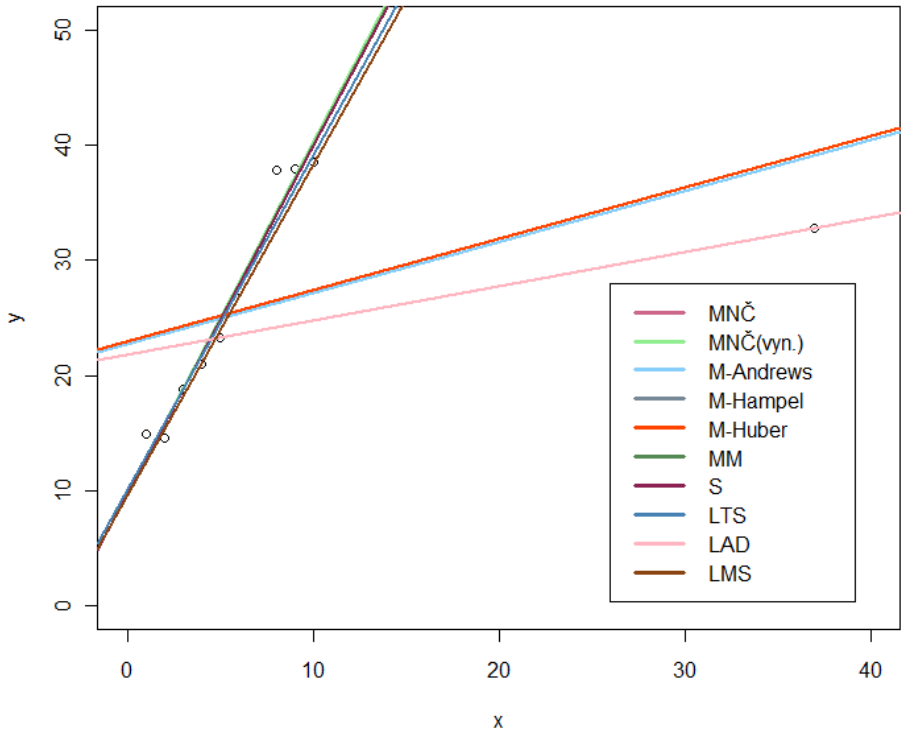

LTS(h=8)
y
x
0
50
100
0
5
10
15

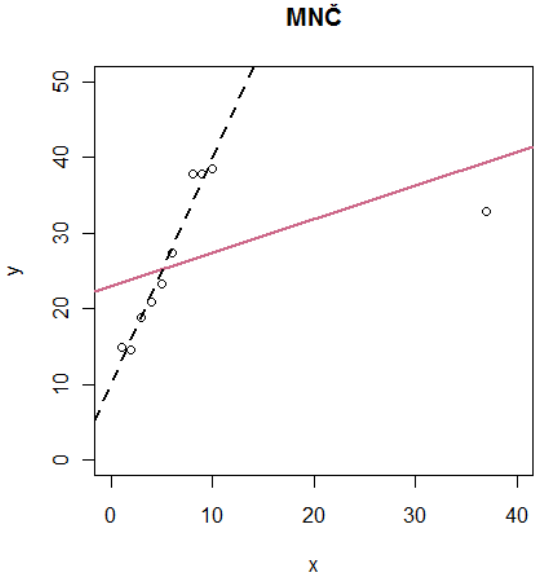

LTS(h=6)
y
x
0
50
100
0
5
10
15

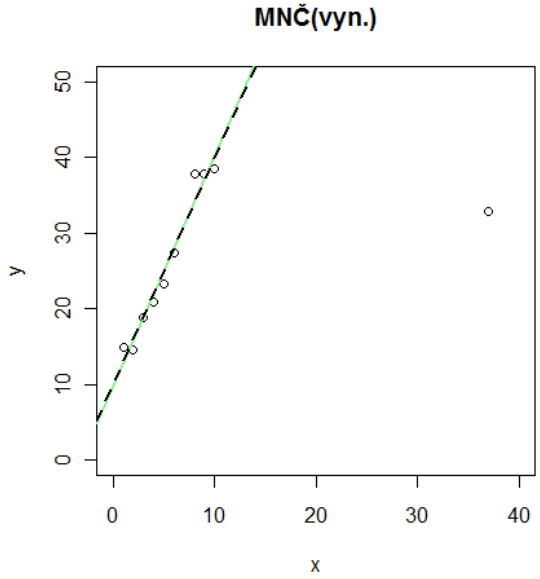

LAD
y
x
0
50
100
0
5
10
15

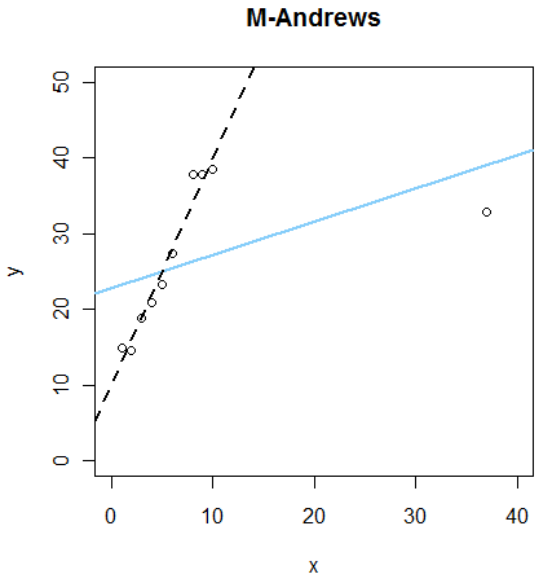

LMS
y
x
0
50
100
0
5
10
15

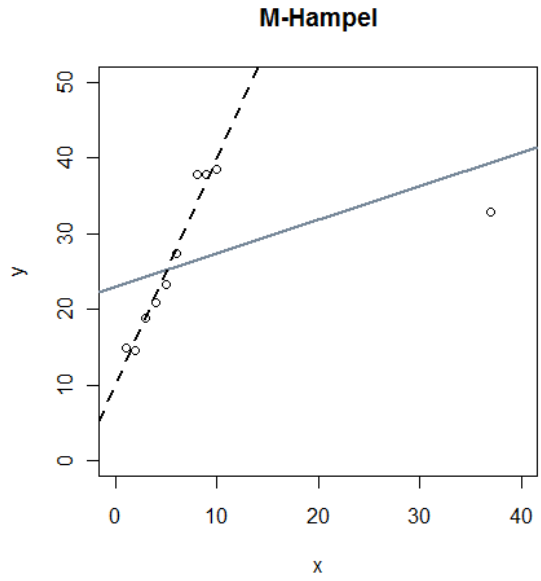
Příloha 5

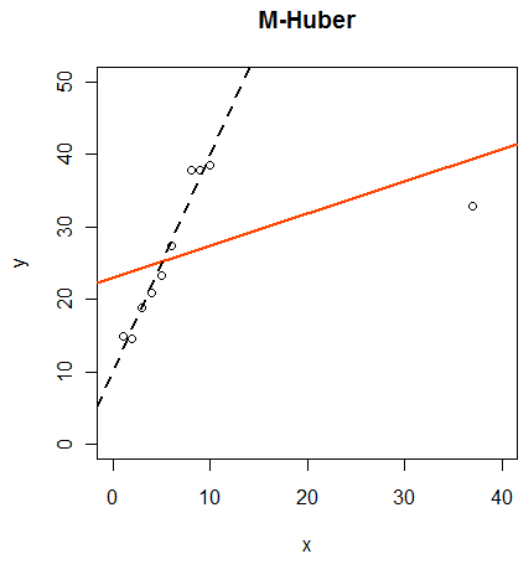

MNČ
MNČ(vyn.)
M-Andrews
M-Hampel
M-Huber
MM
S
LTS
LAD
LMS
y
x
0
10
20
30
40
50

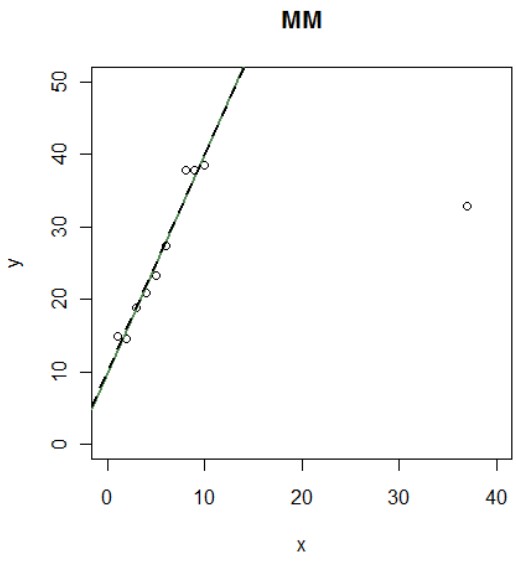

MNČ
y
x
0
10
20
30
40
50

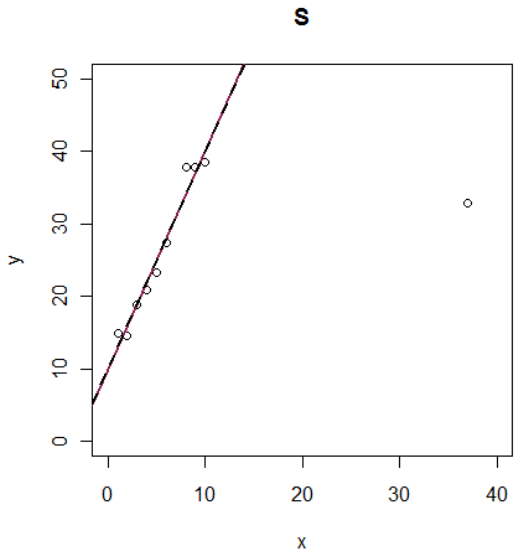

MNČ(vyn.)
y
x
0
10
20
30
40
50

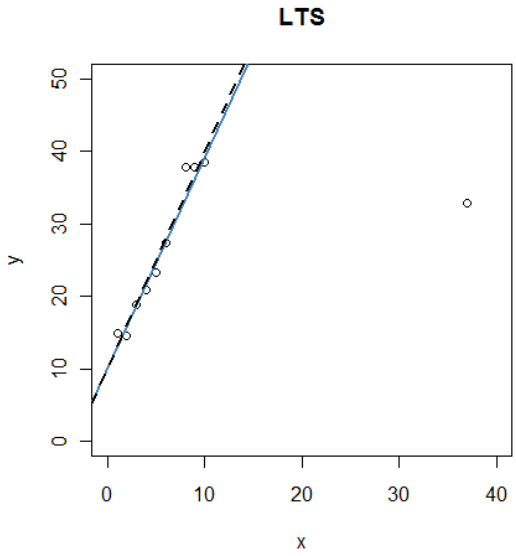

M-Andrews
y
x
0
10
20
30
40
50

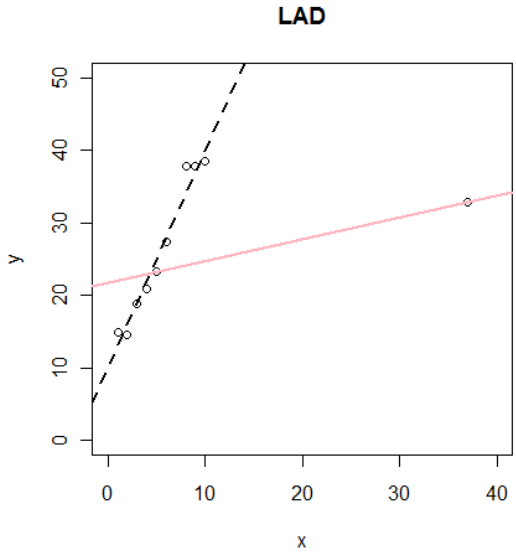

M-Hampel
y
x
0
10
20
30
40
50

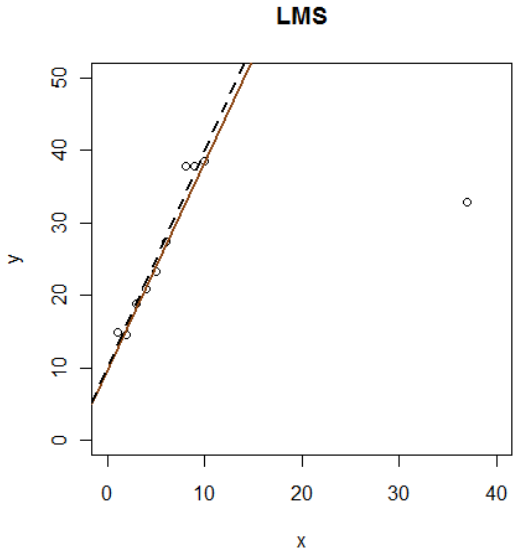

M-Huber
y
x
0
10
20
30
40
50


MM
y
x
0
10
20
30
40
50


S
y
x
0
10
20
30
40
50


LTS
y
x
0
10
20
30
40
50


LAD
y
x
0
10
20
30
40
50


LMS
y
x
0
10
20
30
40
50

Literatura

[1] Anděl, J., Statistické metody, 2. přeprac. vyd., Praha: Matfyzpress, 1998.

[2] Anděl, J., Základy matematické statistiky, 1. vyd., Praha : Matfyzpress, 2005.

[3] Antoch, J., Vorlíčková, D., Vybrané metody statistické analýzy dat, 1. vyd., Praha : Academia, 1992.

[4] Blatná, D., Metody statistické analýzy, 4. vyd., Praha : Bankovní institut vysoká škola, 2009.

[5] Blatná, D., Robustní přístup v lineární regresi (Praktické důvody pro použití robustní regrese)
http://panda.hyperlink.cz/cestapdf/pdf08c3/blatna.pdf [online 3. 3. 2012]

[6] Chen, C., Robust Regression and Outlier Detection with the ROBUSTREG Procedure, SAS Institute Inc., Cary, NC
http://www2.sas.com/proceedings/sugi27/p265-27.pdf [online 3. 3. 2012]

[7] Draper, N.R., Smith, H., Applied regression analysis, 3. vyd., New York, N.Y., Chichester, Weinheim, John Wiley & Sons, 1998.

[8] Fišerová, E., Text k přednáškám předmětu Statistické lineární modely, 2011.

[9] Hebák, P., Regrese. Část 1, 1. vyd., Praha : Vysoká škola ekonomická, 1998.

[10] Hellebrandová, I., Regrese s AR(p) chybami, bakalářská práce
http://is.muni.cz/th/211585/prif_b/Regrese_s_AR_p_chybami.pdf [online 30. 3. 2012]

[11] Huber, P.J., Ronchetti, E.M., Robust Statistics, 2.vyd., John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[12] Hron, K., Text k přednáškám předmětu Pravděpodobnost a matematická statistika 2, 2011.

[13] Jukl, M., Lineární algebra: Homomorfismy a Euklidovské vektorové prostory, VUP Olomouc, 2006.

[14] Kuan, CH.M., An Introduction to quantile regression http://jinhe.xjtu.edu.cn/upfiles/1185633662.pdf [online 30. 3. 2012]

[15] Montgomery, D.C., Peck, E.A. a Vining, G.G., Introduction to linear regression analysis, 4th ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[16] Motl, T., Regresní analýza, bakalářská práce http://is.muni.cz/th/175869/prif_b/Bakalarska_prace_Tomas_Motl_175869.pdf [online 8. 3. 2012]

[17] Renauda, O., Victoria-Feser, M.P., A robust coefficient of determination for regression http://www.hec.unige.ch/www/dms/hec_en/victoriafeser/recherche/paper-rob-R2.pdf [online 30. 3. 2012]

[18] Rousseeuw, P.J., Leroy, A.M., Robust Regression and Outlier Detection, New York, John Wiley & Sons, 1987.

[19] Rousseeuw, P.J., Driessen, K.V., A Fast Algorithm for the Minimum Covariance Determinant Estimator, 1999.

[20] Rousseeuw, P.J, Yohai, V.J., Robust regression by means of S-estimators ftp://ftp.win.ua.ac.be/pub/preprints/84/Robreg84.pdf [online 13. 3. 2012]

[21] SAS 9.2 Help and Documentation

[22] Zvára, K., Štěpán, J., Pravděpodobnost a matematická statistika, 3. vyd., Praha : Matfyzpress, 2002.

[23] Žigárdy, M., Matematické modelování závislosti mezi ekonomickými veličinami
http://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=17616
[online 30. 3. 2012]

[24] http://www.le.ac.uk/users/dsgp1/COURSES/MATHSTAT/13mlreg.pdf
[online 8. 3. 2012] (Maximum-Likelihood Estimation of the Classical Linear Model)

[25] http://socserv.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Robust-Regression.pdf [online 3. 3. 2012] (Robust Regression in R; An Appendix to An R Companion to Applied Regression, Second Edition, John Fox & Sanford Weisberg)

[26] http://www.utd.edu/ herve/Abdi-LeastSquares06-pretty.pdf [online 8. 3. 2012] (The method of least squares)

[27] http://prf.osu.cz/doktorske_studium/dokumenty/Multivariable_Data_Analysis.pdf
[online 8. 3. 2012] (Tvrdík, J., Analýza vícerozměrných dat, Učební texty Ostravské Univerzity, 2003)

[28] http://www.informs-sim.org/wsc06papers/011.pdf [online 8. 3. 2012] (White Noise Assumptions Revisited: Regression Metamodels and Experimental Designs in Practise)

[29] http://aiolos.um.savba.sk/ṽiktor/Econ/Lecture3.pdf [online 8. 3. 2012] (Witkovský, V., Prednáška 3, Klasický lineárny model)

[30] http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/xaghtmlnode12.html#eq:lts:def [online 9. 3. 2012] (XploRe Tutorial System, Least Trimmed Squares)