UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Metody robustní vůči prvkovým odlehlým hodnotám

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Bc. Viktorie Nesrstová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Viktorie Nesrstová

Název práce: Metody robustní vůči prvkovým odlehlým hodnotám

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Detekce odlehlých hodnot je velmi důležitá oblast datové analýzy. Tyto hodnoty mohou být chyby v pozorování, ale také mnohdy nesou důležitou informaci. Často se navíc stává, že v celém pozorování jsou odlehlé jen některé jeho složky. V této práci jsou představeny tři robustní statistické metody vytvořené pro práci s datovými soubory, které odlehlé hodnoty obsahují. Nejprve je čtenář seznámen se samotnou teorií robustní statistiky a kontaminačních modelů. Dále jsou popsány již výše zmíněné robustní statistické metody, které jsou dále aplikovány na simulované datové soubory. Závěr práce je věnován aplikaci těchto metod na reálný datový soubor.

Klíčová slova: Robustní statistika, prvkové odlehlé hodnoty, řádkové odlehlé hodnoty, metoda Detect Deviating Cells, střílející S-odhad, tříkroková regrese

Počet stran: 75

Počet příloh: 3

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Viktorie Nesrstová

Title: Methods robust against cellwise outliers

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2019

Abstract: Outlier detection is a very important part of a data analysis. Outliers can be either measurement errors or a piece of a valuable information. Moreover, it is very common that just a few components of an observation are outlying. The thesis describes three robust statistical methods designed for datasets containing outliers. Firstly, a theory of robust statistics and contamination models is explained. Secondly, the three robust methods are described, followed by applications on the simulated datasets. Finally, these three methods are applied on a real dataset as well.

Key words: Robust statistics, cellwise outliers, rowwise outliers, method Detect Deviating Cells, shooting S-estimator, three-step regression

Number of pages: 75

Number of appendices: 3

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych na tomto místě poděkovala panu doc. RNDr. Karlu Hronovi, Ph.D. za spolupráci, ochotu a cenné rady, díky kterým jsem tuto práci zdárně dokončila. Také bych chtěla poděkovat své rodině a přátelům, kteří mě po celou dobu studia podporovali.

Úvod

Cílem této diplomové práce je nastudovat, popsat a prakticky otestovat robustní statistické metody vhodné pro aplikaci na datové soubory, které obsahují odlehlá pozorování. V současné době se setkáváme s datovými soubory opravdu velkých rozměrů. Je důležité umět s těmito soubory efektivně pracovat, protože jen vzácně neobsahují odlehlé hodnoty, které je potřeba správně identifikovat.

Práci lze rozdělit do dvou hlavních částí. V první části je v rámci dvou kapitol představena robustní statistika a robustní metody vhodné pro práci s datovými soubory obsahujícími odlehlé hodnoty. Ve druhé části jsou tyto metody aplikovány na několik simulovaných a jeden reálný datový soubor.

První kapitola se věnuje robustní statistice a odlehlým pozorováním. Jejím cílem je poskytnout čtenáři základní poznatky o robustní statistice, proč se k ní přistupuje a jaké základní nástroje používá. Dále jsou popsány dva modely kontaminace datových souborů.

V další kapitole jsou uvedeny tři robustní statistické metody, které byly v nedávné době vytvořeny. Čtenáři je představena metoda Detect Deviating Cells, střílející S-odhad a tříkroková regrese. Pro všechny metody je popsán i jejich výpočetní algoritmus.

Třetí kapitola se zabývá simulační studií, ve které jsou aplikovány všechny tři metody popsané v kapitole druhé. Kapitola je doplněna o kódy ze softwaru R.

Poslední kapitola je věnovaná aplikaci zmíněných metod na reálný datový soubor. Analýza datového souboru je provedena hned dvakrát, neboť při druhé analýze byla na proměnné použita log–transformace, která reflektuje jejich měřítko.

Kapitola 1

Úvod do robustní statistiky a problematiky odlehlých pozorování

Cílem této kapitoly je seznámit čtenáře s robustní statistikou a s problematikou odlehlých pozorování. Dále jsou v této kapitole představeny dva modely kontaminace dat. V následujícím textu je čerpáno převážně z [1], [11] a [15].

1.1. Robustní statistika

Obecně můžeme přístupy ke statistice rozdělit na *klasický* a *robustní*. Problémem klasického přístupu je jeho citlivost na odlehlá pozorování. Klasické odhady jako například výběrový průměr, výběrová kovariance a korelace a dále také odhad metodou nejmenších čtverců pro lineární regresi jsou snadno ovlivnitelné, byť jen jednou ohlehlou hodnotou. Z tohoto důvodu se přistupuje právě k robustní statistice. Robustní statistické odhady jsou odolné vůči odlehlým hodnotám a poskytují tak relevantnější výsledky.

Nyní si uvedeme několik pojmů, se kterými se v této práci dále setkáme.

Pravidlo tří sigma:
Nechť x_i, $i = 1, \ldots, n$, je dané pozorování z náhodného výběru, $\overline{x}$ značí výběrový průměr a s je výběrová směrodatná odchylka. Podíl

$$t_i = \frac{x_i - \overline{x}}{s}$$

se tradičně používá pro měření „vychýlenosti“ daného pozorování. Pozorování, pro které platí $|t_i| > 3$, je označeno jako podezřelé (pravidlo tří sigma), a může jít tudíž o odlehlou hodnotu. Toto pravidlo se ale potýká s řadou nevýhod:

1. Ve velkém výběru „dobrých“ dat může dojít k nesprávnému označení dobré hodnoty jako odlehlé.

2. Pro příliš malé výběry není pravidlo tří sigma efektivní.

3. Je-li ve výběru více odlehlých hodnot, jejich efekty se mohou navzájem ovlivňovat, potažmo ovlivňovat použité statistiky, a tudíž některé (nebo všechny) odlehlé hodnoty mohou zůstat nerozpoznané.

Funkce ρ:

Definice 1.1.1. *ρ–funkcí se nazývá taková funkce ρ, pro kterou platí:*

1. *$\rho(x)$ je neklesající funkcí $|x|$*

2. *$\rho(0) = 0$*

3. *$\rho(x)$ je rostoucí pro $x > 0$ tak, že $\rho(x) < \rho(\infty)$*

4. *Je-li ρ omezená, tak se předpokládá, že $\rho(\infty) = 1$.*

Bod selhání:
Bod selhání odhadu $\hat{\theta}$ parametru θ je nejmenší kontaminace výběru taková, která způsobí, že odhad $\hat{\theta}$ začne nabývat libovolných hodnot [5].

Definice 1.1.2. *Mějme dán výběr $\{x_1, \ldots, x_n\}$, ve kterém je m hodnot $\{x_{i_1}, \ldots, x_{i_m}\}$ nahrazeno libovolnými hodnotami $\{y_1, \ldots, y_m\}$. Nový výběr označme jako $\{z_1, \ldots, z_n\}$. Bod selhání odhadu $\hat{\theta}$ je dán jako*

$$\varepsilon_n^*(\hat{\theta}; x_1, \ldots, x_n) = \min\left\{\frac{m}{n}; \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |\hat{\theta}(z_1, \ldots, z_n)| = \infty\right\}.$$

Odlehlá pozorování mohou výrazně ovlivnit klasické číselné charakteristiky či metody. Podívejme se nyní, jaký vliv mají na průměr, směrodatnou odchylku a lineární regresi.

1.1.1. Průměr a směrodatná odchylka

Uvažujme transponované vektory $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. Poznamenejme, že v tomto textu nerozlišujeme náhodné veličiny a jejich realizace. Výběrový průměr a výběrovou směrodatnou odchylku definujeme postupně jako

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Na následujícím jednoduchém příkladu si ukážeme, jak dokáže jediná odlehlá hodnota v souboru ovlivnit výběrový průměr a výběrovou směrodatnou odchylku.

Příklad 1.1.1. Mějme náhodný výběr 18 hodnot

5,60 4,90 4,75 5,20 6,10 3,90 5,45 4,60 35,30

3,55 7,20 4,30 5,15 6,25 5,10 4,35 3,75 7,45

kde 35,30 je očividně odlehlá hodnota. Vypočítáme výběrový průměr a výběrovou směrodatnou odchylku ($\overline{x} = 6{,}827778,\ s = 7{,}187514$). Můžeme si povšimnout, že průměrná hodnota je vyšší než většina hodnot ve výběru. Směrodatná odchylka navíc vyšla příliš vysoká vzhledem k daným hodnotám. Odlehlá hodnota 35,30 tedy výběrový průměr a výběrovou směrodatnou odchylku výrazně ovlivnila. Budeme-li nyní počítat bez této odlehlé hodnoty, získáme $\overline{x} = 5{,}152941,\ s = 1{,}11446$, což jsou hodnoty daleko více odpovídající zbylým hodnotám výběru.

V praxi ovšem nemůžeme odlehlé hodnoty jen tak vyřadit z datového souboru. Je třeba zvážit, jestli daná odlehlá hodnota nenese cennou informaci nebo jestli se ve skutečnosti nejedná o „dobrou“ hodnotu.

Jako určitou robustní alternativu výběrového průměru lze brát *výběrový medián*. Seřadíme-li si pozorování v $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ vzestupně podle velikosti

$$x_{(1)} \leq \cdots \leq x_{(n)}$$

tak pro nějaké celé číslo m platí

$$Med(\mathbf{x}) = \begin{cases} x_{(m)} & \text{pro } n \text{ liché, } n = 2m-1, \\ \frac{x_{(m)}+x_{(m+1)}}{2} & \text{pro } n \text{ sudé, } n = 2m. \end{cases}$$

Hodnoty průměru a mediánu jsou si přibližně rovny, je-li výběr symetricky rozdělen okolo středu, v ostatních případech to platit nemusí. Vrátíme-li se k našemu příkladu, tak hodnota mediánu pro data s odlehlou hodnotou je $Med(\mathbf{x}) = 5{,}125$ a pro data bez odlehlé hodnoty $Med(\mathbf{x}) = 5{,}1$. Hodnota mediánu se tedy téměř nezměnila.

Pro směrodatnou odchylku lze jako robustní alternativu uvažovat *mediánovou absolutní odchylku (MAD)*

$$MAD(\mathbf{x}) = Med\{|\mathbf{x} - Med(\mathbf{x})|\}.$$

Abychom mohli MAD srovnávat se směrodatnou odchylkou, je třeba nadefinovat takzvanou *normovanou MAD* jako

$$MADN(\mathbf{x}) = \frac{MAD(\mathbf{x})}{\Phi^{-1}(0{,}75)}$$

kde Φ^{-1} je inverze distribuční funkce normovaného normálního rozdělení (kvantilová funkce). Pro náš ukázkový příklad je $MAD = 1{,}18608$ pro výběr s odlehlou hodnotou
a $MAD = 1{,}11195$ pro výběr bez odlehlé hodnoty. Je zřejmé, že přítomnost odlehlé hodnoty ve výběru hodnotu MAD téměř neovlivnila. Srovnáme-li ještě obě hodnoty s příslušnými hodnotami výběrových směrodatných odchylek, tak zjistíme, že pro výběr s odlehlou hodnotou se MAD a s výrazně liší, kdežto po odstranění odlehlé hodnoty se MAD a s téměř rovnají.

Může se zdát výhodné preferovat medián a MAD, abychom si „pojistili“, že výsledné hodnoty nebudou ovlivněny hodnotami odlehlými. Nevýhodou ovšem je, že medián a MAD mají menší vypovídající hodnotu v případě, že datový soubor žádné odlehlé hodnoty neobsahuje.

Podívejme se nyní, jak přítomnost odlehlých hodnot ovlivňuje lineární regresi.

1.1.2. Lineární regrese

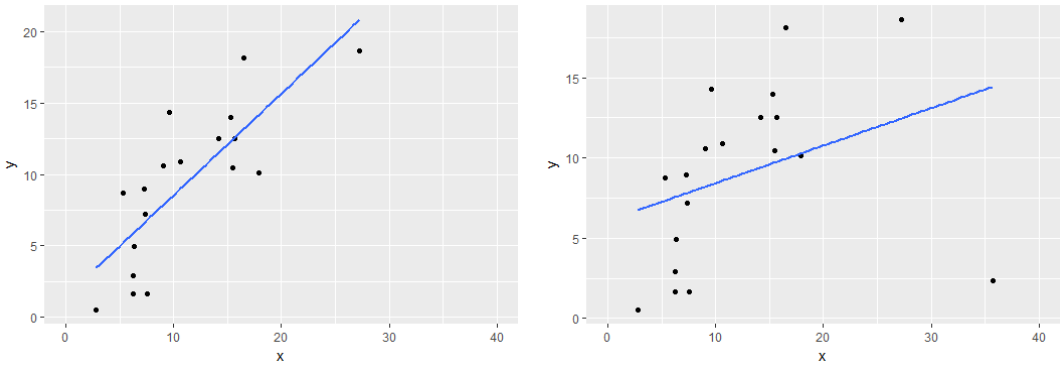
V této podkapitole si ukážeme, jak mohou odlehlé hodnoty ovlivnit odhady parametrů regresní přímky a načrtneme možné robustní alternativy. Uvažujme nyní klasický regresní model [6]

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, n \geq 2,$$

kde β_0 a β_1 jsou neznámé parametry, $x_1, \ldots, x_n$ jsou známá reálná čísla a $\varepsilon_1 \ldots, \varepsilon_n$ jsou chyby měření. Klasická metoda nejmenších čtverců minimalizuje součet čtverců odchylek pozorovaných hodnot a odhadnutých hodnot. Matematicky vyjádřeno [6], náhodné veličiny $\hat{\beta}_0, \hat{\beta}_1$ se nazývají odhady parametrů β_0 a β_1 metodou nejmenších čtverců, jestliže minimalizují výraz

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Odhady parametrů regresní přímky získané klasickou metodou nejmenších čtverců jsou citlivé na přítomnost odlehlých hodnot v datovém souboru. Následující dva grafy v obrázku 1.1 tuto citlivost na odlehlé hodnoty ilustrují.



Obrázek 1.1: Znázornění vlivu odlehlého pozorování na regresní přímku.

Na levém grafu můžeme vidět regresní přímku pro náhodně vygenerovaná data z normálního rozdělení bez odlehlé hodnoty. Regresní přímka poměrně dobře

odpovídá povaze datového souboru. Na pravém grafu je vykreslen graf pro tatáž data, ovšem s přidaným odlehlým pozorováním. Můžeme si povšimnout výrazné změny sklonu regresní přímky. Odlehlá hodnota k sobě regresní přímku „přitáhne“, což způsobí, že regresní přímka již neodpovídá povaze zbývajících dat z datového souboru. Z grafů je tedy patrné, že pouhé jedno odlehlé pozorování v datovém souboru může nepříznivým způsobem ovlivnit sklon regresní přímky, jejíž parametry jsou odhadnuty klasickou metodou nejmenších čtverců. Pokud jsou v datovém souboru odlehlá pozorování výrazně ovlivňující regresní přímku, je na místě počítat robustní odhady regresní přímky. Podívejme se nyní na výpočet regresních parametrů. Jak již bylo zmíněno výše, odhady parametrů klasickou metodou nejmenších čtverců[1], tedy

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (r_i(\boldsymbol{\beta}))^2,$$

kde $r_i(\boldsymbol{\beta})$ značí i-té reziduum, nejsou robustní a tudíž na ně mají odlehlé hodnoty velký vliv. Další možností je takzvaný *L1 odhad* ve tvaru

$$\hat{\boldsymbol{\beta}}_{L1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})| .$$

Tento odhad ale stále není příliš vhodný. Ačkoliv je odolný vůči vertikálním odlehlým hodnotám (ve směru osy y), stále na něj mají vliv odlehlé hodnoty ve směru osy x. Přistoupíme tedy k takzvaným *M-odhadům*. Vyjdeme z následujícího obecného vztahu pro regresní koeficienty

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})),$$

kde pro příslušné volby *ρ–funkce* dostaneme odhady nejmenších čtverců ($\rho(x) = x^2$) nebo L1 odhady ($\rho(x) = |x|$). Vzhledem k tomu, že chceme redukovat vliv velkých reziduí, provedeme následující úpravu výrazu na

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right),$$

[1] Zkratka *LS* z anglického *Least squares*.

kde $\hat{\sigma}$ značí robustní odhad směrodatné odchylky reziduí.

Bod selhání robustních M-odhadů je ale 0% [5], jelikož tyto odhady nejsou obecně robustní vůči vlivným hodnotám (ve směru osy x). Z tohoto důvodu se přistupuje k takzvaným *MM-odhadům*. Výpočet MM-odhadů probíhá následovně:

1. Výpočet počátečního odhadu $\hat{\boldsymbol{\beta}}_0$, který má vysokou hodnotu bodu selhání, ale není příliš přesný.

2. Výpočet robustní směrodatné odchylky $\hat{\sigma}$ pro rezidua $r_i(\hat{\boldsymbol{\beta}}_0)$. Odhad robustní směrodatné odchylky $\hat{\sigma}$ musí být M-odhad, získaný řešením rovnice

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\hat{\sigma}}\right) = \delta$$

   s $\delta > 0$.

3. Výpočet odhadu $\hat{\boldsymbol{\beta}}$ pomocí algoritmu *iterativně převážených nejmenších čtverců* [5].

Výsledný odhad má stejnou hodnotu bodu selhání jako $\hat{\boldsymbol{\beta}}_\mathbf{0}$ a je poměrně eficientní.

Problematika robustní statistiky je ovšem daleko rozsáhlejší než jak je nyní načrtnuto v této práci. Podrobnější se o ní lze dočíst například v [7] nebo [11]. Nyní se podíváme na samotná odlehlá pozorování a představíme si dva modely kontaminace datového souboru.

1.2. Odlehlá pozorování

Datové soubory tvoří obdélníkovou matici o rozměrech $n \times p$, kde n udává velikost datového souboru a p jeho dimenzi. Jinými slovy, n je počet pozorování a p je počet proměnných. V současné době se můžeme často setkat s datovými soubory s velkým počtem pozorování i proměnných a často se stává, že je počet proměnných vyšší než počet pozorování.

Mnoho datových souborů obsahuje ve větší či menší míře odlehlá pozorování neboli outliery. Jedná se o ta pozorování, která jsou nějakým způsobem

vychýlená od ostatních v daném datovém souboru. Dle [15] mohou být odlehlá pozorování v závislosti na dané situaci buď chyby v pozorování, které by mohly nepříznivě ovlivnit analýzu dat, nebo nové a nečekané informace o struktuře daného datového souboru. Může jít o extrémně vysoké nebo naopak nízké hodnoty proměnných, případně o hodnoty, které spolu s ostatními v daném pozorování představují nečekanou kombinaci. Abychom potlačili nepříznivý vliv odlehlých hodnot na datovou analýzu, je důležité tyto hodnoty správně identifikovat. S rostoucím počtem proměnných však zároveň roste obtížnost identifikace odlehlých pozorování.

Podle [15] se většinou ve statistice a datové analýze odlehlým pozorováním rozumí celý řádek v matici dat. Pro opravdu velké datové soubory jsou ale metody, které zkoumají daný soubor takto po řádcích, nevhodné. Často se totiž stává, že většina prvků v daném řádku je v pořádku a pouze několik málo z nich jsou odlehlé hodnoty. Říkáme, že tyto prvky jsou takzvaně *kontaminované*.

Podívejme se nyní na dva modely, *klasický kontaminační model* a *nezávislý kontaminační model*.

1.2.1. Klasický kontaminační model

Předpokládejme, že mnohorozměrný datový soubor je uspořádán do matice, jejíž řádky tvoří pozorování a ve sloupcích jsou proměnné, tj. $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, kde $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$. Většina postupů pro robustní analýzu mnohorozměrných dat je založena na klasickém Tukeyho-Huberově kontaminačním modelu (THCM)[2], v němž může být kontaminovaný pouze malý podíl řádků datového souboru. V tomto modelu je kontaminační mechanismus modelován jako směs dvou rozdělení: jedno odpovídá nominálnímu modelu a druhé odlehlým hodnotám. Přesněji řečeno, Tukeyho-Huberův kontaminační model uvažuje následující třídu rozdělení:

$$\mathscr{H}_\varepsilon = \{H = (1-\varepsilon)H_0 + \varepsilon\tilde{H} : \tilde{H} \text{ je jakékoliv rozdělení na } \mathbb{R}^p\}$$

[2] Z anglického *Tukey-Huber contamination model*.

kde ε je pravděpodobnost kontaminace prvku, H_0 je centrální parametrické rozdělení, například mnohorozměrné normální rozdělení $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$, a $\tilde{H}$ je nespecifikované rozdělení generující odlehlé hodnoty. Dále předpokládejme, že pro pozorování platí $\mathbf{x}_i \sim H$, kde $H \in \mathscr{H}_\varepsilon$. Klíčovou vlastností tohoto modelu je, že pro malé ε většinou dostáváme $\mathbf{x}_i \sim H_0$, tudíž detekování odlehlého pozorování a snížení jeho váhy je smysluplné a v praxi funguje velmi dobře.

V následující podkapitole si představíme nezávislý kontaminační model (ICM)[3], který je vhodný pro situace, kdy jsou jednotlivé prvky matice kontaminovány nezávisle.

1.2.2. Nezávislý kontaminační model

V praxi se můžeme setkat s takovým modelem, ve kterém budou jednotlivé prvky v $\mathbf{X}$ nezávisle kontaminované. Tento případ je typický pro vysokodimenzionální data, jejichž proměnné jsou často měřeny odděleně, případně získány z různých zdrojů. Model pro nezávisle kontaminované prvky byl poprvé popsán v [2].

Uvažujme nyní následující třídu rozdělení:

$$\mathscr{I}_\varepsilon = \{H : H \text{ je rozdělení pro } \mathbf{X} = (\mathbf{I} - \mathbf{X}_0)\mathbf{B}_\varepsilon + \tilde{\mathbf{X}}\mathbf{B}_\varepsilon\}$$

kde řádky $\mathbf{X}_0 \sim H_0$, $\tilde{\mathbf{X}} \sim \tilde{H}$ a $\mathbf{B}_\varepsilon = diag(B_1, \ldots, B_p)$, kde B_j jsou nezávislé a pocházejí z rozdělení $Bi(1, \varepsilon)$. Každý prvek z $\mathbf{X}$ je tedy nezávisle kontaminován s pravděpodobností ε. Pravděpodobnost $\bar{\varepsilon}$, že alespoň jeden prvek z $\mathbf{X}$ bude kontaminovaný je vyjádřena jako

$$\bar{\varepsilon} = 1 - (1 - \varepsilon)^p,$$

kde ε je pravděpodobnost kontaminace prvku z datové matice a p značí její dimenzi. S rostoucí hodnotou ε nebo p, případně pokud obě tyto hodnoty rostou, bude velmi rychle překročena hodnota 0,5. Na následujícím jednouchém příkladu si danou situaci ilustrujeme.

[3] Z anglického *Independent contamination model*.

Příklad 1.2.1. Mějme $\varepsilon = 0{,}05$ a p postupně rovno hodnotám 20, 25 a 30. Budeme-li postupně dosazovat do vzorce pro výpočet pravděpodobnosti, že alespoň jeden prvek z $\mathbf{X}$ bude kontaminovaný, získáme následující hodnoty:

$$\bar{\varepsilon}_1 = 0{,}641 \quad \bar{\varepsilon}_2 = 0{,}722 \quad \bar{\varepsilon}_3 = 0{,}785$$

Vidíme, že pouze s lehkým nárůstem dimenze dochází k poměrně velkému nárůstu pravděpodobnosti kontaminace alespoň jednoho prvku z $\mathbf{X}$. Ještě výrazněji je tento nárůst patrný, pokud si zvolíme pevně dimenzi p a bude růst hodnota ε. Nechť $p = 30$ a ε se postupně rovná hodnotám 0,01, 0,05 a 0,1. Vypočítáné pravděpodobnosti $\bar{\varepsilon}$ jsou následující:

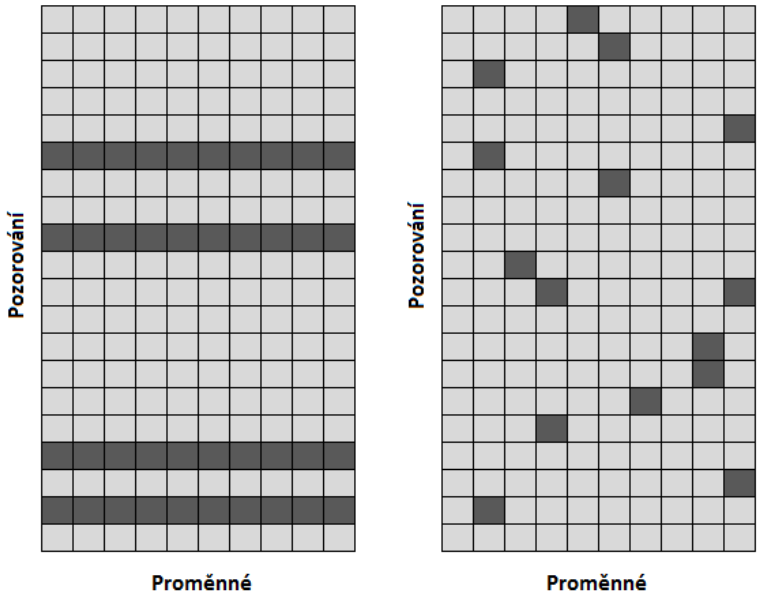
$$\bar{\varepsilon}_1 = 0{,}260 \quad \bar{\varepsilon}_2 = 0{,}785 \quad \bar{\varepsilon}_3 = 0{,}957$$

Můžeme si všimnout velmi výrazného nárůstu pravděpodobnosti kontaminace alespoň jednoho prvku, když hodnota ε vzrostla z 0,01 na 0,05.

Je tedy zřejmé, že s rostoucí pravděpodobností ε roste pravděpodobnost kontaminace celého řádku. Tento fakt je problematický v tom, že metody pro odhalování řádkových outlierů se nedokážou vypořádat s datovými soubory, které obsahující více než 50% kontaminovaných řádků.

Podívejme se nyní na jednu výraznou odlišnost výše zmíněných modelů. Bereme-li jako odlehlé pozorování celý řádek, potom toto pozorování svou strukturou vůbec nezapadá do sledovaného datového souboru [15]. V prvkovém modelu se předpokládá, že jen některé hodnoty jsou odlehlé od ostatních. Zbývající prvky v pozorování obsahují užitečné informace.

Na obrázku 1.2 je znázorněn problém, který by nastal, pokud bychom na každé pozorování obsahující alespoň jednu odlehlou hodnotu pohlíželi jako na řádkové odlehlé pozorování. V návaznosti na výše zmíněné tvrzení, že takové pozorování vůbec nezapadá do datového souboru, bychom z datového souboru v pravé části obrázku vyřadili (potlačili) většinu pozorování, čímž by byl datový soubor znehodnocen. Z daných patnácti pozorování by zbyly pouze čtyři. Takovýto datový soubor by byl prakticky nepoužitelný pro další analýzu.



Obrázek 1.2: Řádková a prvková odlehlá pozorování.

Pro řešení těchto problémů je potřeba uvést metody, které nezkoumají datové soubory pouze po řádku, ale i po jednotlivých proměnných. V následující kapitole si představíme tři různé metody, které se pro práci s odlehlými pozorováními v datovém souboru využívají.

Kapitola 2

Metody pro práci s odlehlými hodnotami

Celá následující kapitola pojednává o metodách a postupech využívaných k detekci odlehlých hodnot a potlačení jejich vlivu. Jmenovitě jsou to Detect Deviating Cells, tříkroková regrese a střílející S-odhad pro robustní regresi. V textu je čerpáno převážně z [1], [10], [13] a [15].

2.1. Detect Deviating Cells

Metoda Detect Deviating Cells (zkráceně DDC) byla představena v článku [15]. Oproti metodám, které zkoumají datové soubory po řádcích, respektive dřívějším metodám pro detekci prvkových odlehlých hodnot, disponuje několika výhodami. Tyto výhody jsou následující:

- Metoda bere v potaz korelace mezi proměnnými.
- Na metodu DDC není kladen požadavek, aby více než 50% pozorování bylo bez odlehlých hodnot. Pro tuto metodu neexistuje žádné omezení na počet kontaminovaných pozorování.
- Je vhodná i pro velké datové soubory s velkým počtem proměnných.
- Metoda rovněž počítá odhady odlehlých hodnot a zároveň imputuje chybějící hodnoty do datového souboru.

Metoda DDC využívá ve svém algoritmu robustní statistické nástroje. V následující podkapitole si tento algoritmus představíme.

2.1.1. Algoritmus metody DDC

Předpokládejme, že vstupní data tvořící matici $\mathbf{X}$ pocházejí z mnohorozměrného normálního rozdělení s vektorem středních hodnot $\boldsymbol{\mu}$ a pozitivně semidefinitní varianční maticí $\boldsymbol{\Sigma}$. Některé prvky matice $\mathbf{X}$ byly poškozeny nebo chybí. Proměnné by měly nabývat číselných hodnot a měl by jich být dostatečný počet. Je důležité, aby číselné proměnné nabývaly vícero různých hodnot (pokud by například hodnoty v některé proměnné byly pouze 0 a 1, činilo by to problém). Aby byla tato podmínka zajištěna, tak příslušný kód v softwaru nejprve identifikuje ty proměnné, které danou podmínku nesplňují, načež výpočty provede pouze pro proměnné splňující podmínku. Tyto zbývající proměnné by měly být alespoň přibližně normálně rozdělené, což lze ověřit pomocí Q-Q grafů. Je třeba provést vhodnou transformaci těch proměnných, které nejsou normálně rozdělené, na přibližně normálně rozdělené proměnné.

Nyní můžeme přistoupit k samotnému algoritmu metody DDC.

1. **Standardizace**

   Pro každý j-tý sloupec matice $\mathbf{X}$ odhadneme

   $$m_j = robLoc_i(x_{ij}) \quad \text{a} \quad s_j = robScale_i(x_{ij} - m_j), \tag{2.1}$$

   kde *robLoc* je robustní odhad polohy a *robScale* je robustní odhad směrodatné odchylky (za předpokladu, že jeho argumenty byly centrovány), viz například ty z kapitoly 1.1.1. Dále standardizujeme matici $\mathbf{X}$ na matici $\mathbf{Z}$ následujícím způsobem,

   $$z_{ij} = \frac{x_{ij} - m_j}{s_j}. \tag{2.2}$$

2. **Aplikace jednorozměrné detekce odlehlých hodnot na všechny proměnné**

   Ve druhém kroku definujeme matici **U**, jejíž prvky jsou následující

$$u_{ij} = \begin{cases} z_{ij} & \text{pro } |z_{ij}| \leq c, \\ NA & \text{pro } |z_{ij}| > c. \end{cases} \tag{2.3}$$

   Protože jsme v předchozím kroku provedli standardizaci (2.2), je vztah (2.3) takzvaným *sloupcovým detektorem* odlehlých hodnot. Prahem c rozumíme číslo

$$c = \sqrt{\chi^2_{1,p}}, \tag{2.4}$$

   kde $\sqrt{\chi^2_{1,p}}$ je p-tý kvantil chí-kvadrát rozdělení s jedním stupněm volnosti. Hodnota p je implicitně zadaná jako 0,99. V ideální situaci by tak byl vstup označen jako odlehlá hodnota s pravděpodobností 0,01.

3. **Vztahy mezi dvojicemi proměnných**

   Vypočítáme korelace pro jakékoliv dvě proměnné $h \neq j$,

$$cor_{jh} = robCorr_i(u_{ij}, u_{ih}), \tag{2.5}$$

   kde *robCorr* je robustní míra korelace. Výpočet probíhá pro všechna i pro která platí, že ani jedna z hodnot u_{ij} nebo u_{ih} není chybějící hodnota (NA). Dále budeme využívat pouze tak silný lineární vztah mezi j a h, že

$$|cor_{jh}| \geq corrlim, \tag{2.6}$$

   přičemž výchozí hodnota *corrlim* je 0,5. Proměnné j, které splňují (2.6) pro nějaká $h \neq j$, nazveme *související*. Tyto proměnné nesou důležitou informaci o sobě navzájem. Ostatní proměnné nazveme *nesouvisející proměnné*.

   Pro dvojice proměnných (j, h), které splňují (2.6), spočítáme

$$b_{jh} = robSlope_i(u_{ij} \mid u_{ih}), \tag{2.7}$$

kde *robSlope* slouží k výpočtu sklonu robustní regresní přímky bez absolutního členu (například pomocí některého z robustních odhadů z kapitoly 1.1.2). Pomocí této přímky predikujeme proměnnou j z proměnné h. Výpočet sklonu bude použit v následujícím kroku pro předpovědi souvisejících proměnných.

4. **Predikce**

   Dalším krokem je výpočet předpovědí $\hat{z}_{ij}$ pro všechny prvky. Pro každou proměnnou j předpokládáme množinu H_j, která obsahuje všechny proměnné h splňující (2.6) včetně j. Pro všechna $i = 1, \ldots, n$ položíme

$$\hat{z}_{ij} = G(\{b_{jh}u_{ih};\ h \in H_j\}), \tag{2.8}$$

   kde G je takzvané *kombinační pravidlo* [15] použité na tato čísla. Toto pravidlo vynechává chybějící hodnoty a přiřazuje nulu tam, kde již žádné hodnoty nezbývají.

   Pro G lze využít vážený průměr s vahami $w_{jh} = |cor_{jh}|$ nebo také vážený medián. Výhodou (2.8) je fakt, že podíl odlehlého prvku z_{ih} na $\hat{z}_{ij}$ je omezený, neboť $|u_{ih}| \leq c$ a má vliv pouze na jednu složku. To by ovšem neplatilo pro předpověď z mnohonásobné regrese nejmenších čtverců $z_{.j}$ na zbývajících $p-1$ proměnných $z_{.h}$ dohromady.

5. **Odstranění zmenšení škály**

   Použití předpovědí (2.8) může vést k tomu, že se škála vstupů zmenší, což je nežádoucí. Nabízí se zkusit méně výrazné zmenšení jednotlivých prvků $b_{jh}u_{ih}$. Tato možnost je ale bohužel nedostačující, neboť tyto prvky mohou mít jiná znaménka pro různá h. V [15] je tedy doporučeno řešit zmenšení až po použití kombinačního pravidla (2.8). Pro tyto účely nahradíme $\hat{z}_{ij}$

pomocí $a_j \hat{z}_{ij}$ pro všechna i, j, kde

$$a_j := robSlope_{i'}(z_{i'j} \mid \hat{z}_{i'j}) \tag{2.9}$$

je získáno z regrese pozorovaného $z_{\cdot j}$ na predikované $\hat{z}_{\cdot j}$.

6. **Vyhledávání prvkových odlehlých hodnot**

   Ve čtvrtém a pátém kroku jsme vypočítali předpovědi $\hat{z}_{ij}$ pro všechny prvky. Nyní spočítáme standardizovaná rezidua pro jednotlivé prvky pomocí vztahu

$$r_{ij} = \frac{z_{ij} - \hat{z}_{ij}}{robScale_{i'}(z_{i'j} - \hat{z}_{i'j})}. \tag{2.10}$$

   Dále potom v každém j-tém sloupci označíme všechny prvky, pro které platí $|r_{ij}| > c$, jako odlehlé. Hodnota c je určena ve druhém kroku algoritmu pomocí (2.4). Dále také sestavíme imputovanou matici $\mathbf{Z}_{imp}$. Tato matice je stejná jako matice $\mathbf{Z}$ až na to, že odlehlé hodnoty z_{ij} a chybějící hodnoty jsou nahrazeny jejich odhady $\hat{z}_{ij}$. Prvky, jež nebyly označené jako odlehlé, zůstávají zachovány beze změny. Pro zvýšení přesnosti odhadů je možné za označené prvky dosadit NA a zopakovat kroky čtyři až šest.

7. **Vyhledávání řádkových odlehlých hodnot**

   Jedna z možností, jak označit i-tý řádek jako odlehlý, je jednoduše spočítat prvky, pro něž $|r_{ij}|$ překročí práh c. Tím by ale byly vynechány řádky s mnoha poměrně velkým $|r_{ij}| < c$. Další možností by bylo porovnat průměr přes čtvercová standardizovaná rezidua v i-tém řádku $ave_j(r_{ij}^2)$ s hodnotou c. To by ale vedlo k tomu, že by řádek s jedním výrazně vyčnívajícím prvkem byl označen jako odlehlé pozorování, což není žádoucí. Řešení tohoto problému je následující:

   Za předpokladu nulové hypotézy o mnohorozměrném normálním rozdělení bez odlehlých hodnot je rozdělení veličiny r_{ij} blízké normovanému nor-

málnímu rozdělení, tudíž distribuční funkce pro r_{ij}^2 je přibližně distribuční funkce F pro χ_1^2 rozdělení, což vede k následujícímu kritériu:

$$T_i = \underset{j=1}{\overset{p}{\text{ave}}} F\left(r_{ij}^2\right) \tag{2.11}$$

Dále provedeme standardizaci T_i tak jako v (2.2) a označíme ty řádky i, pro něž standardizovaná hodnota T_i překročí práh c z (2.4). Najdeme-li i-tý řádek, pro který je hodnota T_i neobvykle vysoká, tak to automaticky neznamená, že celé pozorování nezapadá mezi ostatní pozorování z dané skupiny. Je třeba ale takovému pozorování věnovat pozornost.

Ačkoliv můžeme pomocí hodnoty T_i identifikovat některé řádkové odlehlé hodnoty, existují i takové, které i přesto zůstanou skryty. Z tohoto důvodu je doporučeno použít v pozdější fázi analýzy dat, po vyloučení označených řádků, řádkové robustní metody.

8. **Destandardizace**

   Posledním krokem je destandardizace imputované matice $\mathbf{Z}_{imp}$ na matici $\mathbf{X}_{imp}$. Hlavním výstupem metody Detect Deviating Cells je matice $\mathbf{X}_{imp}$ spolu s indexy prvků i celých řádků, které byly označené jako odlehlé.

Metoda Detect Deviating Cells je tedy schopná odhalit jednotlivé prvkové odlehlé hodnoty v datovém souboru. Klíčovým bodem této metody je výpočet předpovědi každého prvku datového souboru. Zde je také výhodou vyšší dimenze datové matice. Máme-li více proměnných, zvýší se nám množství informace o daném datovém souboru, a tím pádem můžeme predikovat hodnoty s větší přesností. Metoda je výpočetně náročnější než v případě, kdy by se pro detekci prvkových odlehlých hodnot brala každá proměnná zvlášť. Výhodou je ovšem schopnost odhalit i ty odlehlé hodnoty, které by jinak zůstaly skryté. Grafickým výstupem metody je takzvaná *cellmap*, která přehledně označí prvkové i řádkové odlehlé hodnoty. Příklady jsou uvedeny ve čtvrté kapitole. Kód pro metodu DDC je dostupný jak v softwaru R (knihovna **cellWise**) [14], tak v Matlabu.

2.2. Střílející S-odhad pro robustní regresi

V této podkapitole si ukážeme regresní odhad nazvaný *střílející S-odhad*, který byl představen v článku [13]. Tento odhad je rovněž vhodný právě pro případy, kdy jsou v pozorování kontaminovány jen některé prvky. Spojuje v sobě takzvaný souřadnicový spádový algoritmus a jednoduchý regresní S-odhad.

2.2.1. Idea metody

Střílející S-odhad využívá souřadnicový spádový algoritmus, taktéž zvaný střílející algoritmus. Tato metoda původně používala lasso regresi. Problémem je, že lasso odhady nejsou robustní. Proto je ve střílejícím S-odhadu lasso odhad nahrazen nepenalizovaným S-odhadem. V porovnání s běžnou S-regresí umožňuje souřadnicový spádový algoritmus při výpočtu odhadů regresních parametrů zvážit jednotlivé složky pozorování různě.

Lasso odhad je v [13] definován jako

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + 2\lambda\sum_{j=1}^{p}|\beta_j|.$$

Po aktualizaci odhadu lasso koeficientu $\hat{\beta}_j$ v souřadnicovém spádovém algoritmu pro $j = 1, \ldots, p$ zůstanou všechny ostatní koeficienty stejné, a to jako $\hat{\beta}_k$, $k \neq j$

$$\begin{aligned}\hat{\beta}_{j,Lasso} &= \underset{\beta_j\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}((y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k) - x_{ij}\beta_j)^2 + 2\lambda\sum_{k\neq j}|\hat{\beta}_k| + 2\lambda|\beta_j| = \\ &= \underset{\beta_j\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}((y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k) - x_{ij}\beta_j)^2 + 2\lambda|\beta_j|.\end{aligned} \tag{2.12}$$

což můžeme vzít jako jednoduchou lasso regresi, kde nová vysvětlovaná proměnná

$$\tilde{y}_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k \quad i = 1, \ldots, n \tag{2.13}$$

je v regresi na x_{ij} pro pevnou hodnotu j.

V případě střílejícího S-odhadu chceme zajistit, aby nová vysvětlovaná proměnná $\tilde{y}_i^{(j)}$ nebyla ovlivněna odlehlými hodnotami v prvcích x_{ik}. Z tohoto důvodu nadefinujeme regresní váhy

$$w_{ik} = w\left(\frac{|\tilde{y}_i^{(k)} - x_{ik}\hat{\beta}_k|}{\hat{\sigma}_k}\right)$$

kde argumentem váhové funkce $w(\cdot)$ je reziduum z regrese $\tilde{y}_i^{(k)}$ na x_{ik} normované robustní směrodatnou odchylkou reziduí $\hat{\sigma}_k$. Tudíž w_{ik} určuje odlehlost prvku x_{ik} v regresi $\tilde{y}_i^{(k)}$ na x_{ik}. Váhová funkce by měla být nerostoucí na množině kladných čísel a nabývat hodnot z intervalu $[0, 1]$. V [13] bylo zvoleno pro zachování jednoduchosti silné (binární) zamítnutí, kde pro $r \leq c$ je $w(r) = 1$, jinak se rovná 0. Dle [13] by při volbě $c = 3$ bylo očekáváno, že méně než 0,3% čistých pozorování budou označena jakou odlehlá v regresním modelu s normálně rozdělenými chybami. Jiné volby pro funkci vah jsou samozřejmě možné.

Nová vysvětlovaná proměnná $\tilde{y}_i^{(j)}$ je tedy definovaná jako

$$\tilde{y}_i^{(j)} = y_i - \sum_{k \neq j} \tilde{x}_{ik}\hat{\beta}_k \quad \text{kde} \quad \tilde{x}_{ik} = w_{ik}x_{ik} + (1 - w_{ik})\hat{x}_{ik} \tag{2.14}$$

Narozdíl od (2.13), v (2.14) jsou ve výpočtu nové vysvětlované proměnné nahrazeny hodnoty x_{ik} konvexní kombinací $\tilde{x}_{ik}$ pozorované hodnoty x_{ik} a její „opravené“ hodnoty $\hat{x}_{ik}$. Vzhledem k tomu, že známe $\tilde{y}_i^{(k)}$ a $\hat{\beta}_k$, můžeme „opravenou“ hodnotu $\hat{x}_{ik}$ vypočítat pomocí kalibrace

$$\hat{x}_{ik} = \frac{\tilde{y}_i^{(k)}}{\hat{\beta}_k}. \tag{2.15}$$

Aby nedošlo k výpočetním problémům, je stanoveno $\hat{x}_{ik} = 0$ pro malá $|\hat{\beta}_k|$. Hodnota $\tilde{x}_{ik}$ může být interpretována jako očištěná hodnota x_{ik} v designové matici. Je-li pozorování označeno jako odlehlé a má nulovou váhu, potom se $\tilde{x}_{ik}$ rovná „opravené“ hodnotě $\hat{x}_{ik}$. Naopak pokud je pozorování čisté (tedy ve shodě

s převažujícím rozdělením datového souboru) a má váhu jedna, potom se očištěná hodnota rovná té pozorované. Je důležité poznamenat, že hodnoty $\hat{x}_{ik}$ a w_{ik} jsou závislé na $\hat{\beta}_k$ pro $k \neq j$.

K výpočtu regresního odhadu $\hat{\beta}_j$ nepoužijeme metodu lasso (2.12), ale robustní nepenalizovaný jednoduchý S-regresní odhad [13]. To přímo vede ke střílejícímu S-odhadu, který je definován po jednotlivých proměnných v závislosti na tom, že známe ostatní odhady $\hat{\beta}_k$, $k \neq j$, jako

$$\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}} \hat{\sigma}_j(\beta), \tag{2.16}$$

kde $\hat{\sigma}_j(\beta)$ je definováno jako řešení s rovnice

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{\tilde{y}_i^{(j)} - x_{ij}\beta}{s}\right) = \delta. \tag{2.17}$$

Tudíž $\hat{\sigma}_j(\hat{\beta}_j)$ je M-odhad variability vypočítaný z reziduí. Hodnota δ je rovna střední hodnotě funkce ρ při normálním rozdělení, t.j. $\delta = \mathbb{E}[\rho(Z)]$ pro $Z \sim N(0,1)$. Volba je provedena tak, aby hodnota bodu selhání odhadu nebyla příliš nízká a zároveň aby bylo docíleno dostatečně vysoké eficience. Vyšší hodnota δ znamená vyšší hodnotu bodu selhání, ale nižší eficienci.

Za funkci ρ se volí buď *Tukey's biweight*

$$\rho_{BI}(z) = \begin{cases} \frac{k_{BI}^2}{6}(1-(1-(\frac{z}{k_{BI}})^2)^3), & \text{pokud } |z| \leq k_{BI}, \\ \frac{k_{BI}^2}{6}, & \text{pokud } |z| > k_{BI}, \end{cases} \tag{2.18}$$

nebo funkce pojmenovaná jako *skipped Huber*

$$\rho_{skH}(z) = \begin{cases} \frac{1}{2}z^2, & \text{pokud } |z| \leq k_{skH}, \\ \frac{k_{skH}^2}{2}, & \text{pokud } |z| > k_{skH}. \end{cases} \tag{2.19}$$

Funkce *Tukey's biweight* a *skipped Huber* se od sebe poměrně dost liší. Huberova funkce je kvadratická okolo středu na intervalu $[-k_{skH}, k_{skH}]$ a konstantní mimo tento interval. Oproti tomu *Tukey's biweight* funkce je hladká, ale omezuje efekt

extrémních hodnot. Podle [13] se kromě těchto dvou funkcí dají zvolit i jiné funkce ρ, o kterých se blíže píše v [11]. Hodnoty k_{BI} a k_{skH} jsou dle [13] voleny tak, aby byl bod selhání 20%, a to jako $k_{BI} = 3{,}420$ a $k_{skH} = 2{,}177$. Tato volba zajišťuje dobrý poměr mezi robustností a eficiencí.

2.2.2. Algoritmus metody

K výpočtu střílejícího S-odhadu se používá iterativní postup, který je podobný souřadnicovému spádovému algoritmu. Samotný algoritmus výpočtu je podrobně rozepsán v článku [13], zde si uvedeme jen jeho obecný popis. Kód lze potom získat z webových stránek jeho autorky [12].

Vyjdeme z předpokladu, že máme model s absolutním členem α.

1. **Cyklus algoritmu**

   Za předpokladu pevného j spočítáme v každém kroku hodnotu $\tilde{y}_i^{(j)}$ pomocí (2.14) a (2.15). Dále spočítáme jednoduchý regresní S-odhad pro $\tilde{y}_i^{(j)}$ na x_{ij}. K tomuto výpočtu využijeme iterativně převážený algoritmus nejmenších čtverců uvedený v [11]. Sestává z dalšího iterativního algoritmu, v jehož každé iteraci je vypočítán vážený odhad nejmenších čtverců parametru β_j. Spolu s ním je vypočítána i nová hodnota M-odhadu směrodatné odchylky $\hat{\sigma}_j(\hat{\beta}_j)$, a to nalezením pevného bodu rekurzivní verze (2.17),
   $f(s) = \frac{1}{n\delta}\sum_{i=1}^{n}\rho(\frac{\tilde{y}_i^{(j)} - x_{ij}\hat{\beta}_j}{s})s = s$. Nevýhodou ovšem je, že není zaručena konvergence tohoto cyklu.

2. **Vstupní hodnoty**

   Nejprve použijeme Huberovu funkci na predikované hodnoty a získáme tak „přibližně čisté“ prediktory $\tilde{x}_{ij}^0$. Dále využijeme MM-odhad pro výpočet výchozích koeficientů $\hat{\beta}_j^{(0)}$ s *lineární kvadratickou kvadratickou (lqq)* funkcí ρ [8] a ladícími konstantami nastavenými pro 50% bod selhání a 95% eficienci.

Uvedená metoda je tedy vhodná pro zkoumání kontaminace jednotlivých složek pozorování. Střílející S-odhad lze také využít také jako diagnostický nástroj. Po výpočtu střílejícího S-odhadu jsou prvky w_{ij} matice vah schopné rozlišit čistá data a odlehlé hodnoty. Navíc dokážou rozpoznat prvkovou a řádkovou kontaminaci. Nízká váha je přiřazena odlehlé hodnotě, vysoká váha hodnotě čisté. Pokud mají všechmy složky pozorování nízké váhy, pak jsou kontaminované nebo se jedná o vertikální odlehlou hodnotu.

Je důležité zmínit, že střílející S-odhad má problémy s prvkovými dobrými vlivnými pozorováními, tedy extrémními pozorováními ve směru regresního trendu, jelikož má tendenci označovat kontaminované složky dobrých vlivných pozorování jako odlehlé hodnoty při výpočtu počátečních hodnot algoritmu.

Vzhledem k tomu, že střílející S-odhad přistupuje ke každé proměnné zvlášť, může být použit i na datové soubory malého rozsahu, dokonce i na takové, kde počet proměnných převyšuje počet pozorování, tedy $n < p$.

2.3. Tříkroková regrese

Poslední metoda, kterou si zde představíme, je *tříkroková regrese*, která byla navržena v článku [10]. Tato metoda se zabývá jak prvkovými odlehlými hodnotami, tak celými odlehlými řádky.

V [10] byl uveden tříkrokový regresní odhad, jenž v sobě kombinuje filtrování prvkových odlehlých hodnot a robustní regresi prostřednictvím odhadu varianční matice. Stručný algoritmus výpočtu je dán následovně:

1. Použití konzistentního jednorozměrného filtru k nalezení a odstraněné extrémních prvkových odlehlých hodnot.

2. Aplikace robustního odhadu mnohorozměrné polohy a variability na filtrovaná data za účelem snížení váhy řádkových odlehlých hodnot.

3. Výpočet robustních regresních koeficientů z odhadů získaných v kroku 2.

Filtrováním dat je myšleno detekování odlehlých hodnot a následně jejich

nahrazení chybějícími hodnotami (NA). Je-li zvolen filtr, který je schopný detekovat odlehlé hodnoty a zároveň zachovává všechna čistá data, nebo alespoň jejich většinu, výsledný odhad může být odolný vůči prvkovým i řádkovým odlehlým hodnotám zároveň. Navíc docílí konzistence a asymptotické normality pro čistá data.

2.3.1. Konzistentní filtr

Při předzpracování mnohorozměrných dat a hledání prvkových odlehlých hodnot se jako jedna z možností nabízí pravidlo tří sigma, kde by byly nalezené odlehlé hodnoty nahrazeny hodnotami NA. K nalezení odhadu mnohorozměrné polohy a variability by byl následně použit takzvaný *EM algoritmus* pro uměle vytvořená nekompletní data. Problémem tohoto přístupu je ale nekonzistence a problém se zpracováním řádkových odlehlých hodnot. Z tohoto důvodu je vhodnější použít takzvaný *konzistentní filtr*.

Filtrování je metoda pro předzpracování dat prováděná za účelem kontroly vlivu potenciálních prvkových odlehlých hodnot. V [10] je filtrování prováděno tak, že jsou odlehlé hodnoty označeny a následně nahrazeny hodnotami NA. Konzistentním filtrem se rozumí filtr, který dokáže asymptoticky zachovat všechna data, jsou-li čistá, což je žádoucí.

Nechť X je náhodná proměnná se spojitou distribuční funkcí $G(x)$. Nadefinujeme horní a dolní chvost rozdělení distribuční funkce $G(x)$ jako

$$F^u(t) = P_G\left(\frac{X-\eta^u}{med(X-\eta^u \mid X>\eta^u)} \leq t \mid X>\eta^u\right), \tag{2.20}$$

respektive

$$F^l(t) = P_G\left(\frac{\eta^l - X}{med(\eta^l - X \mid X<\eta^l)} \leq t \mid X<\eta^l\right), \tag{2.21}$$

kde med označuje medián, $\eta^u = G^{-1}(1-\alpha)$, $\eta^l = G^{-1}(\alpha)$ a $0 < \alpha < 0,5$. Pro jednoduchost zápisu budeme dále psát $s^u = med(X-\eta^u \mid X > \eta^u)$ a $s^l = med(\eta^l - X \mid X < \eta^l)$.

Nechť $\{X_1, \ldots, X_n\}$ je náhodný výběr z G a nechť $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ je odpovídající uspořádaný výběr. Konzistentní odhady pro η^u, s^u, η^l, s^l jsou dány jako

$$\hat{\eta}_n^u = \hat{G}_n^{-1}(1-\alpha), \quad \hat{s}_n^u = med(\{X_i - \hat{\eta}_n^u \mid X_i > \hat{\eta}_n^u\})$$

$$\hat{\eta}_n^l = \hat{G}_n^{-1}(\alpha), \quad \hat{s}_n^l = med(\{\hat{\eta}_n^l - X_i \mid X_i < \hat{\eta}_n^l\})$$

kde $\hat{G}_n^{-1}(a) = X_{([na])}$, $0 < a < 1$ je empirický kvantil a $med(\{Y_1, \ldots, Y_m\}) = Y_{([\frac{m}{2}])}$ je výběrový medián. Empirické distribuční funkce pro normované horní a dolní chvosty jsou dány jako

$$\hat{F}_n^u(t) = \frac{\sum_{i=1}^n I(0 < \frac{X_i - \hat{\eta}_n^u}{\hat{s}_n^u} \leq t)}{\sum_{i=1}^n I(X_i > \hat{\eta}_n^u)}, \tag{2.22}$$

$$\hat{F}_n^l(t) = \frac{\sum_{i=1}^n I(0 < \frac{\hat{\eta}_n^l - X_i}{\hat{s}_n^l} \leq t)}{\sum_{i=1}^n I(X_i < \hat{\eta}_n^l)}, \tag{2.23}$$

kde I je indikátorová funkce.

Odlehlé hodnoty z horního a dolního chvostu jsou odhaleny tak, že srovnáme empirické distribuční funkce normovaných chvostů s jejich očekávanou distribucí. Nechť obecně $\{a\}^+ = \max(0, a)$ značí kladnou část čísla a. Podíl označených odlehlých hodnot z horního a dolního chvostu definujeme jako

$$\hat{d}_n^u = \sup_{t \geq t_0} \{F_0(t) - \hat{F}_n^u(t)\}^+$$

$$\hat{d}_n^l = \sup_{t \geq t_0} \{F_0(t) - \hat{F}_n^l(t)\}^+$$

kde $F_0(t) = 1 - e^{-\log(2)t}$ a $t_0 = \frac{1}{\log(2)}$. Je-li $X - \eta^u \mid X > \eta^u$ exponenciálně rozdělená s parametrem $\lambda^u > 0$, tak standardizovaný chvost $\frac{X - \eta^u}{s^u | X > \eta^u}$ je taktéž exponenciálně rozdělen s parametrem $\log(2)$, což spěje k výše popsanému vyjádření $F_0(t)$ a t_0. Následně již můžeme filtrovat odlehlé hodnoty. Položíme-li

$$\hat{t}_n^u = \min\{t : \hat{F}_n^u(t) \geq 1 - \hat{d}_n^u\} \quad \text{a} \quad \hat{t}_n^l = \min\{t : \hat{F}_n^l(t) \geq 1 - \hat{d}_n^l\},$$

potom filtrujeme hodnoty X_i takové, pro které platí

$$X_i < \hat{\eta}_n^l - \hat{s}_n^l \hat{t}_n^l \quad \text{nebo} \quad X_i > \hat{\eta}_n^u + \hat{s}_n^u \hat{t}_n^u.$$

Zvolený model pro těžké chvosty se jeví jako nejvhodnější, neboť dosahuje dobrého poměru mezi robustností a konzistencí procesu filtrace. V [10] je taktéž dokázána konzistence filtru.

2.3.2. Robustní odhad

Uvažujme model

$$Y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{2.24}$$

pro $i = 1, \ldots, n$, kde ε_i jsou nezávislé, stejně rozdělené rozdělené chyby a $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Odhady získané metodou nejmenších čtverců[1] $(\hat{\alpha}_{LS}, \hat{\boldsymbol{\beta}}_{LS})$ jsou definovány jako minimalizace součtu čtverců reziduí:

$$(\hat{\alpha}_{LS}, \hat{\boldsymbol{\beta}}_{LS}^T) = \underset{(\alpha, \boldsymbol{\beta}^T) \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \alpha - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

Řešení je následující:

$$\hat{\boldsymbol{\beta}}_{LS} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy}, \quad \hat{\alpha}_{LS} = \hat{\mu}_y - \hat{\boldsymbol{\mu}}_x^T \hat{\boldsymbol{\beta}}_{LS}. \tag{2.25}$$

Výše zmíněné $\hat{\boldsymbol{\Sigma}}_{xx}, \hat{\boldsymbol{\Sigma}}_{xy}, \hat{\mu}_y$ a $\hat{\boldsymbol{\mu}}_x$ jsou prvky empirické varianční matice a střední hodnoty:

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{xx} & \hat{\boldsymbol{\Sigma}}_{xy} \\ \hat{\boldsymbol{\Sigma}}_{yx} & \hat{\Sigma}_{yy} \end{pmatrix} \quad \text{a} \quad \begin{pmatrix} \hat{\boldsymbol{\mu}}_x \\ \hat{\mu}_y \end{pmatrix} \tag{2.26}$$

pro sdružená data $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, kde $\mathbf{Z}_i = (\mathbf{x}_i^T, Y_i)^T$.

[1] Index *LS* je zkratkou pro *least squares*.

K docílení robustní regrese pro řádkové odlehlé hodnoty bylo navrhnuto provést robustifikaci komponent v (2.25). Empirickou varianční matici a průměr nahradil mnohorozměrný S-odhad [10]. Tento přístup se nazývá *dvoukroková regrese*, zkráceně 2S-regrese. Bylo dokázáno, že za slabých podmínek (symetrie rozdělení ε_i a nezávislost ε_i) je dvoukroková regrese Fisherovsky konzistentní a asymptoticky normální, a to i tehdy, kdy S-odhady mnohorozměrné polohy a variability samy o sobě konzistentní nejsou. Dvoukroková regrese je navíc odolná jak vůči vertikálním odlehlým hodnotám, tak i špatným vlivným pozorováním a dobrým vlivným pozorováním. Pokud je dobrým vlivným pozorováním dávána menší váha, může to vést k určité ztrátě eficience. Zároveň to ale zamezuje podhodnocení rozptylu odhadu.

V [10] doporučují zabývat se prvkovými i řádkovými odlehlými hodnotami s využitím zobecněného S-odhadu (GSE)[2], který využívá výše vysvětlený konzistentní filtr. Globálně robustní regresní odhad, nazývaný 3S-regrese, je dán jako

$$\hat{\boldsymbol{\beta}}_{3S} = \hat{\mathbf{S}}_{xx}^{-1}\hat{\mathbf{S}}_{xy}, \quad \hat{\alpha}_{3S} = \hat{m}_y - \hat{\mathbf{m}}_x^T\hat{\boldsymbol{\beta}}_{3S} \tag{2.27}$$

kde $\hat{\mathbf{S}}$ a $\hat{\mathbf{m}}$ jsou příslušné robustní odhady. Výpočet zobecněného S-odhadu $(\hat{\mathbf{m}}, \hat{\mathbf{S}})$ je následující:

1. Filtrace extrémních prvkových odlehlých hodnot tak, aby prvkově kontaminovaná pozorování neměla velké robustní Mahalanobisovy vzdálenosti ve druhém kroku.

2. Snížení váhy prvkových odlehlých hodnot použitím GSE pro mnohorozměrnou polohu a variabilitu na data filtrovaná v prvním kroku. GSE je zobecnění S-odhadu pro neúplná data, kde jednotlivé chybějící hodnoty vznikají náhodně a nezávisle na sobě (tzv. *missing at random*). Vzhledem k tomu, že nezávislý kontaminační model předpokládá, že prvky jsou náhodně odlehlé, tak je předpoklad náhodně chybějících dat pro tento model splněn.

[2] Z anglického *generalized S-estimator*.

Podívejme se nyní na celý postup podrobněji. Uvažujme $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Pro každou proměnnou $\mathbf{x}_j = \{x_{1j}, \ldots, x_{nj}\}$, $j = 1, \ldots, p$, provedeme filtraci popsanou v podkapitole zabývající se konzistentním filtrem. Nechť $\{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ jsou pomocné výsledné vektory jedniček a nul, kde nula značí filtrovaný vstup v $\mathbf{x}_i$, tedy $\mathbf{U}_i = (U_{i1}, \ldots, U_{ip})^T$, kde

$$U_{ij} = I(\hat{\eta}^l_{j,n} - \hat{s}^l_{j,n}\hat{t}^l_{j,n} \leq X_i \leq \hat{\eta}^u_{j,n} + \hat{s}^u_{j,n}\hat{t}^u_{j,n}).$$

Filtrace je provedena proto, aby se eliminoval vliv prvkových odlehlých hodnot. V případě, že je podíl pozorování, která obsahují alespoň jeden označený prvek, velmi malý (kupříkladu méně než 1%), se odlehlé hodnoty výrazně neprojeví a můžeme přestat používat filtr. Označme si ξ jako práh, jenž určuje, kdy se má filtr přestat používat. Nyní se blíže podíváme na postup, který zastaví používání filtru v případě, že podíl pozorování obsahujících odlehlé hodnoty je menší než ξ. Tento postup si navíc zachovává všechny robustní vlastnosti filtru. Nechť $n_0 = \#\{1 \leq i \leq n : \mathbf{U}_i = \mathbf{1}\}$ je počet úplných pozorování po provedení filtrace. Určíme

$$\mathbf{U}^*_i = \mathbf{1I}\left(\frac{n - n_0}{n} \leq \xi\right) + \mathbf{U}_i\mathbf{I}\left(\frac{n - n_0}{n} > \xi\right), \quad i = 1, \ldots, n, \tag{2.28}$$

kde ξ je dostatečně malý práh (lze vzít například hodnotu $\xi = 0{,}01$). Nyní se dostáváme k samotnému zavedení zobecněného S-odhadu $(\hat{\mathbf{m}}, \hat{\mathbf{S}})$. Nechť
$\mathbb{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$ a $\mathbb{U} = ((\mathbf{U}^*_1, \ldots, \mathbf{U}^*_n)^T, \mathbf{1})$. Zobecněný S-odhad je definován jako

$$\hat{\mathbf{m}} = \hat{\mathbf{m}}_{GS}(\mathbb{Z}, \mathbb{U}), \quad \hat{\mathbf{S}} = \hat{\mathbf{S}}_{GS}(\mathbb{Z}, \mathbb{U}), \tag{2.29}$$

kde $\hat{\mathbf{m}}_{GS}$ a $\hat{\mathbf{S}}_{GS}$ jsou robustní mnohorozměrné zobecněné S-odhady polohy a variability pro nekompletní data $(\mathbb{Z}, \mathbb{U})$. Pokud by vstupní data byla kompletní, tedy $\mathbb{U} = (\mathbf{1}, \ldots, \mathbf{1})$, zobecněný S-odhad by se zredukoval na S-odhad.

Tak jako předchozí dvě metody, i pro tříkrokovou regresi je k dispozici kód v softwaru R. Existuje pro ni knihovna s názvem **robreg3S** [9].

Kapitola 3

Aplikace metod na simulovaná data

V této kapitole se budeme věnovat simulační studii, jejímž cílem je prozkoumat chování metod popsaných v předchozí kapitole. Pro účely simulací bylo nagenerováno šest datových souborů o rozměrech $n \times p$, jež všechny pocházejí z mnohorozměrného normálního rozdělení, tj. pro každý datový soubor platí $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}_p = \mathbf{1}$ a $\boldsymbol{\Sigma}$ je jednotková matice. Vytvořené datové soubory jsou následujících rozměrů:

n	50	100	100	200	200	500
p	5	10	20	10	20	10

Tabulka 3.1: Rozměry nagenerovaných datových souborů.

Poznámka 3.0.1. Je třeba volit datové soubory vhodných rozměrů. V případě, že datový soubor obsahuje příliš mnoho proměnných ku malému počtu pozorování (například $n = 50$ a $p = 10$) tříkroková regrese výpočetně selhává.

Tyto datové soubory byly následně kontaminovány, tzn. byly v nich uměle vytvořeny odlehlé hodnoty. Prvky byly kontaminovány s pravděpodobností 0,05 následujícím způsobem

$$x_{ij} = x_{ij} \cdot \alpha,$$

kde $\alpha \in \{2, 3, \ldots, 10\}$. Každý datový soubor byl postupně kontaminován všemi vybranými hodnotami α. Z každého nagenerovaného datového souboru tak vzniklo

devět kontaminovaných. Vždy byly kontaminovány stejné prvky, ale jinou hodnotou α. S pravděpodobností 0,95 zůstaly prvky takzvaně „čisté“.

Jako první si ukážeme simulační studií pro metodu DDC. V ní nás bude zajímat převážně úspěšnost metody při hledání uměle vytvořených odlehlých hodnot.

3.1. Metoda DDC - simulační studie

Pro tuto simulaci byl zvolen nagenerovaný datový soubor o rozměrech 100×10. Cílem bylo sledovat úspěšnost metody při detekci odlehlých hodnot, tj. jestli metoda správně označila jako odlehlé hodnoty ty prvky, které opravdu byly uměle kontaminovány, a zároveň čisté prvky nebyly chybně označeny jako odlehlé.

Poznámka 3.1.1. Kontaminace probíhala následovně: bylo nagenerováno 100 různých kontaminačních matic, jejichž prvky jsou rovny 1 s pravděpodobností 0,95 a dané α s pravděpodobností 0,05. Prvky datové matice na pozici kontaminovaných prvků kontaminační matice byly vynásobeny příslušným α (nejprve $\alpha = 2$). Výsledkem je tedy 100 různě kontaminovaných datových souborů, kde kontaminovaný prvek je $2 \cdot x_{ij}$. V dalším kroku byly kontaminovány ty samé prvky, ale tentokrát číslem 3. Stejně se postupovalo i pro zbývající hodnoty α.

Metoda DDC požaduje, aby datový soubor pocházel z mnohorozměrného normálního rozdělení. Tento předpoklad je splněn už z podstaty toho, jakým způsobem byla data generována, není tedy nutné jej ověřovat.

Nyní přistoupíme k samotným výpočtům. Na všech 100 kontaminovaných datových souborů aplikujeme metodu DDC. K tomu slouží funkce `DetectDeviatingCells` z knihovny **cellWise** [14]. Vzorový kód je pro simulaci, ve které prvky kontaminujeme vynásobením číslem 2. Pro ostatní hodnoty α je postup analogický.

```
> Matice<-list()
# 100 prázdných matic
> for (i in 1:100){
+  Matice[[i]]<-matrix(0,nrow=n,ncol=p)
```

```
+}
# 100 kontaminačních matic
> for(i in 1:100){
+  for (k in 1:n){
+   for (l in 1:p){
+    Kont.Matice.2[[i]][k,l]<-rbinom(1,1,prob=0.95)
+  }
+}
+  Kont.Matice.2[[i]][which(Kont.Matice.2[[i]] == 0)]<-alfa[1]
+}
# 100x různě kontaminovaná data
> Data.2<-list()
> for(i in 1:100){
+  Data.2[[i]]<-Data*Kont.Matice.2[[i]]
+}
# Výpočet DDC
> library(cellWise)
> for(i in 1:100){
+   DDC.2[[i]]<-DetectDeviatingCells(Data.2[[i]],DDCpars)
+ }
```

Poznámka 3.1.2. `DDCpars` stanovuje parametry metody. V této práci byly ve výpočtech použity výchozí parametry metody, tj.:

```
> DDCpars = list(fracNA = 0.5,numDiscrete = 3,precScale = 1e-12,
+                tolProb = 0.99, corrlim = 0.5, combinRule = "wmean",
+                includeSelf = TRUE, rowdetect = TRUE,
+                returnBigXimp = F, fastDDC = FALSE)
```

Další možnosti pro parametry metody jsou popsány v [14].

`Data` je nagenerovaný datový soubor, n značí počet pozorování, p dimenzi. Po ukončení výpočtů metoda vypíše, kolik řádků a sloupců z datového souboru vy-

užila. V tomto případě vypadá výstup následovně:
`The input data has 100 rows and 10 columns.`

Poznámka 3.1.3. Metoda DDC před výpočty provede vlastní předzpracování datového souboru. Pokud by datový soubor obsahoval i jiné než pouze číselné proměnné, tak by tyto proměnné metoda sama z analýzy vyloučila. Datový soubor využitý pro simulaci obsahuje pouze číselné proměnné, a tudíž z něj metoda žádné sloupce ani řádky neodstranila.

Poznámka 3.1.4. Funkce `DetectDeviatingCells` v sobě přímo zahrnuje funkci `checkDataSet`. Tato funkce slouží k očištění datového souboru od řádků a sloupců, které nesplňují dané podmínky. Mezi tyto podmínky patří určitý maximální podíl hodnot NA v pozorování či proměnné a minimální počet různých hodnot, kterých musí proměnné nabývat. Více se dá o funkci `checkDataSet` nalézt v [14]. Využitím funkce `DetectDeviatingCells` je tedy zároveň využita i funkce `checkDataSet`. Tuto funkci lze ale také použít zvlášť na očištění datového souboru před jeho následnou analýzou pomocí jiných metod.

Výstupem metody je seznam obsahující několik složek. Nás budou především zajímat tyto:

- `remX`: očištěná data po použití funkce `checkDataSet`
- `stdResid`: matice reziduí
- `indcells`: indexy prvků, které byly označené jako odlehlé hodnoty
- `indrows`: indexy řádků, které byly označené jako odlehlé

Grafickým výstupem metody DDC je *cellmap*. Vzhledem k velkému rozsahu provedené simulace nebyly v této fázi grafické výstupy vytvořeny, a proto si jejich zobrazení a interpretaci necháme až do následující kapitoly.

Během simulací jsme sledovali několik ukazatelů, a to hlavně senzitivitu (true positive rate), míru falešné pozitivity (false positive rate) a míru falešného objevu (false discovery rate). Tyto hodnoty se vypočítají následovně [3], [4]:

- senzitivita (True positive rate)

$$TPR = \frac{TP}{TP + FN},$$

- míra falešné pozitivity (False positive rate)

$$FPR = \frac{FP}{FP + TN},$$

- míra falešného objevu (False discovery rate)

$$FDR = 1 - \frac{TP}{TP + FP},$$

kde TP značí *true positive*, což v našem případě znamená počet odhalených odlehlých hodnot, které opravdu odlehlé jsou (uměle vytvořené při kontaminaci datového souboru). Naopak TN neboli *true negative* jsou hodnoty, které nejsou uměle vytvořené jako odlehlé a byly tak správně označeny jako hodnoty čisté. FP neboli *false positive* jsou ty hodnoty, které byly označeny jako odlehlé, ale ve skutečnosti odlehlé nejsou. Hodnota FN značí *false negative*, což jsou ty hodnoty, které byly uměle vytvořené jako odlehlé, ale byly při výpočtech označeny jako čisté.

Senzitivita (TPR) tedy vyjadřuje proporci vytvořených odlehlých hodnot, které byly správně označeny jako odlehlé. Na druhou stranu FPR vyjadřuje pravděpodobnost chyby 1. druhu. Míru falešného objevu (FDR) definují v [3] jako očekávanou proporci chybně zamítnutých hypotéz. V našem případě je to tedy očekávaná proporce chybně označených čistých hodnot jako odlehlé.

Podívejme se nyní blíž na výsledky provedené simulace. V tabulce 3.2 je uveden medián a minimální a maximální hodnoty TPR, FPR a FDR pro všechny kontaminace datového souboru.

Nejprve se podíváme na hodnotu TPR. Můžeme si povšimnout velkého rozdílu v její minimální hodnotě pro $\alpha = 2$ a $\alpha = 3$. Při kontaminaci datového souboru číslem 3 tato hodnota znatelně vzrostla. Vzrostla tedy minimální proporce odlehlých hodnot správně označených jako odlehlé, z čehož můžeme usoudit, že se

	TPR			FPR			FDR		
α	**Min**	**Med**	**Max**	**Min**	**Med**	**Max**	**Min**	**Med**	**Max**
2	0,0851	0,2324	0,4286	0,0074	0,0107	0,0137	0,3438	0,4791	0,7500
3	0,3519	0,4722	0,6667	0,0063	0,0105	0,0136	0,1818	0,3000	0,5000
4	0,4203	0,5893	0,7436	0,0053	0,0104	0,0135	0,1364	0,2472	0,4483
5	0,5000	0,6796	0,8333	0,0063	0,0096	0,0134	0,1176	0,2222	0,4062
6	0,5652	0,7287	0,8679	0,0053	0,0095	0,0134	0,1087	0,2043	0,4062
7	0,6304	0,7813	0,8776	0,0053	0,0095	0,0134	0,1020	0,1890	0,3714
8	0,6957	0,8043	0,9184	0,0053	0,0095	0,0134	0,0962	0,1875	0,3714
9	0,7255	0,8298	0,9388	0,0053	0,0095	0,0134	0,0943	0,1818	0,3714
10	0,7255	0,8509	0,9464	0,0053	0,0095	0,0134	0,0943	0,1747	0,3611

Tabulka 3.2: Srovnání hodnot TPR, FPR a FDR.

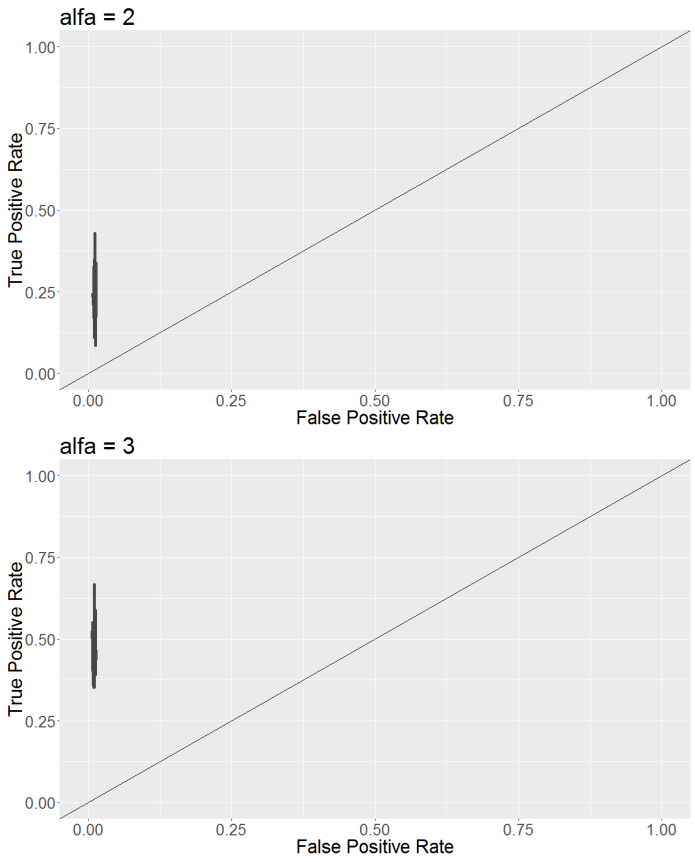
zlepšila schopnost správně identifikovat skutečné odlehlé hodnoty. Dle minimálních hodnot TPR také můžeme říci, že od kontaminace číslem 5 byl vždy správně označen nadpoloviční počet skutečných odlehlých hodnot. Výrazný skok pozorujeme i u mediánu pro první dvě hodnoty α, kde hodnota mediánu pro $\alpha = 3$ je dvonásobkem té pro $\alpha = 2$. Podíváme-li se na maximální hodnoty TPR, tak si můžeme povšimnout, že i při kontaminaci malým číslem ($\alpha = 2$) se metodě v jednom opakování simulace podařilo najít téměř polovinu skutečných odlehlých hodnot. Jedná se ale spíše o zajímavou výjimku, neboť se výrazně liší od výsledků ostatních.

Zaměřme se nyní na hodnotu FPR. Jak bylo uvedeno výše, vyjadřuje tato hodnota pravděpodobnost chyby 1. druhu. Je tedy žádoucí, aby její hodnota byla nižší než hladina významnosti 0,05. Tento požadavek je pokaždé splněn, neboť nejvyší hodnota je $FPR = 0{,}0137$. Navíc se maximální hodnota FPR od kontaminace souboru číslem 5 ustálila na hodnotě $FPR = 0{,}0134$. Obdobně to platí i pro medián, jehož hodnota se od $\alpha = 6$ ustálila na 0,0095.

Nakonec se podíváme ještě na míru falešného objevu (FDR). Její minimální hodnota s rostoucím α podle očekávání klesá, nejvýrazněji opět mezi $\alpha = 2$ a $\alpha = 3$. Pro poslední dvě hodnoty α se ustálila. Co se týče maximální hodnoty

FDR, tak při kontaminaci nejmenším α dosáhla maximální hodnoty 0,75. To znamená, že v jednom opakování simulace se metodě podařilo správně označit pouze 25% skutečných odlehlých hodnot. Jedná se ale opět o jeden výrazný extrém. Více vypovídající je hodnota mediánu. Můžeme si také všimnout, že od kontaminace číslem $\alpha = 4$ už byla míra falešného objevu vždy menší než 0,5.

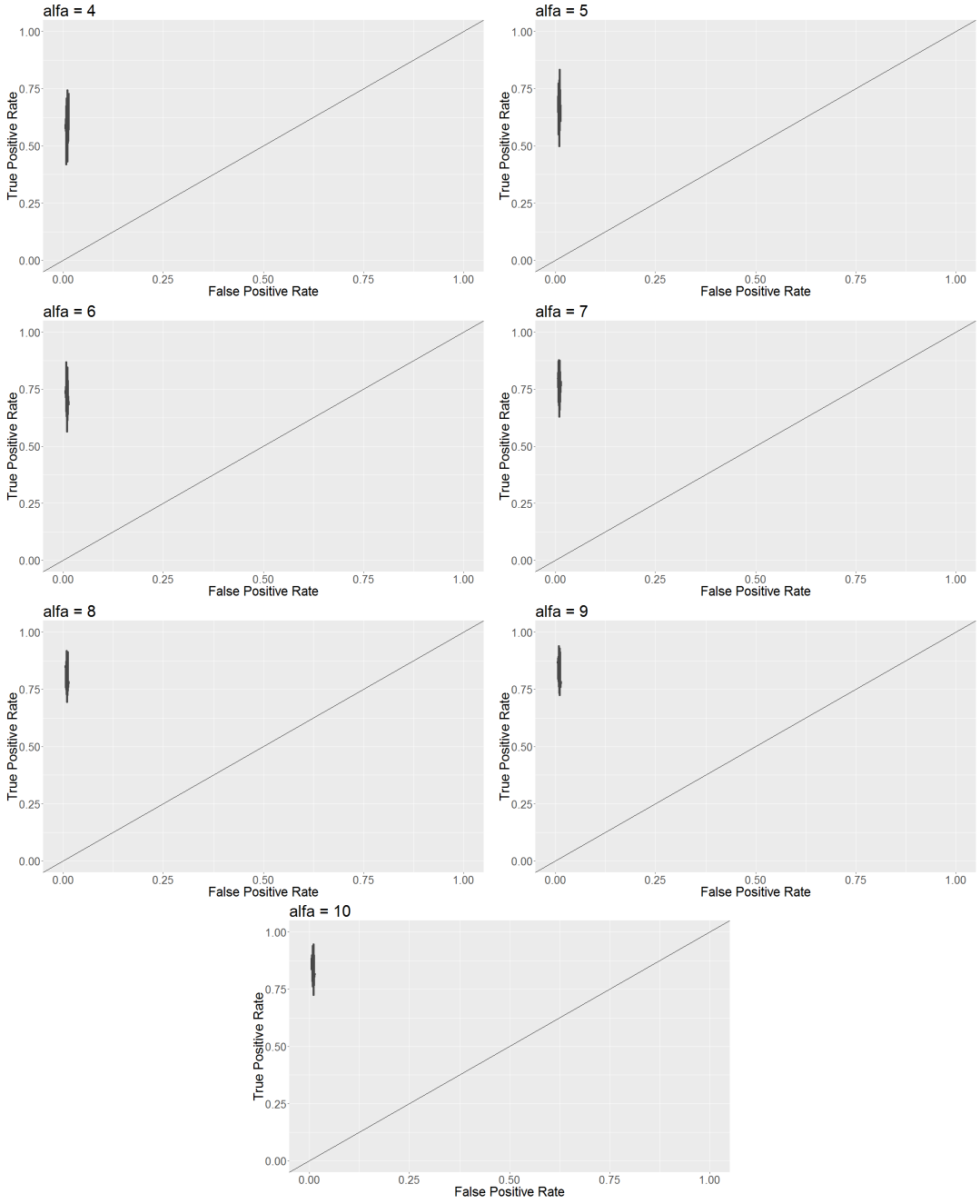
Nyní si ještě graficky znázorníme senzitivitu (TPR) ku míře falešné pozitivity (FPR). Čím více se hodnoty blíží levému hornímu rohu grafu, tím byla simulace úspěšnější v hledání odlehlých hodnot. Naopak, pokud jsou hodnoty blízko diagonále, tak simulace nebyla příliš úspěšná.



Obrázek 3.1: Srovnání TPR a FPR pro první dva kontaminované datové soubory.

V grafech 3.1 máme výsledky pro kontaminaci čísly $\alpha = 2$ a $\alpha = 3$. Mezi hodnotami zobrazenými v těchto dvou grafech došlo k výraznému skoku, při kontaminaci číslem 3 hodnoty viditelně „poskočily“ směrem k levému hornímu rohu. S rostoucí kontaminačním číslem α potom docházelo k očekávanému zlepšení a tedy i úspěšnější identifikaci odlehlých hodnot. Na obrázcích 3.2 můžeme vidět vývoj pro zbývající hodnoty α.

Poznámka 3.1.5. Dle očekávání byla metoda DDC schopná najít téměř všechny odlehlé hodnoty po kontaminaci číslem 10. Je nutné si ale uvědomit, že kontaminované prvky byly opravdu znatelně odlišné od prvků čistých, což analýzu výrazně usnadňuje.



Obrázek 3.2: Srovnání TPR a FPR pro zbývající kontaminované datové soubory.

3.2. Tříkroková regrese a střílející S-odhad–simulační studie

Nyní se blíže podíváme na další dvě metody uvedené v této práci. Budeme tentokrát pracovat se všemi šesti vytvořenými datovými soubory, ne pouze s jedním jako v simulační studii pro metodu DDC. Uvažujme regresní model

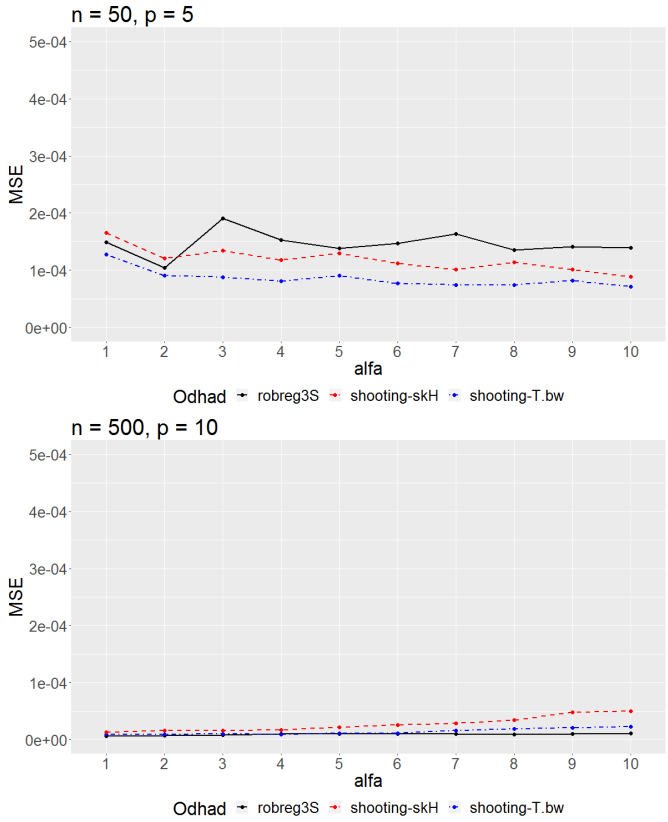
$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

pro $i = 1, \ldots, n$, kde chyby ε_i jsou nezávislé a stejně rozdělené a $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Absolutní člen uvažujeme jako nulový. Koeficienty $\boldsymbol{\beta}$ byly nagenerovány z rovnoměrného rozdělení. Postupně byly vypočítány odhady koeficientů pomocí tříkrokové regrese a střílejícího S-odhadu (s využitím obou ztátových funkcí) pro všechny nagenerované datové soubory, a to jak pro datové soubory postupně kontaminované všemi hodnotami α, tak pro ty čisté. Vliv odlehlých pozorování na vypočítané koeficienty budeme sledovat pomocí *střední čtvercové chyby* vypočítané ze vztahu

$$\overline{MSE} = \frac{1}{N} \sum_{m=1}^{N} \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j^{(m)} - \beta_j^{(m)})^2,$$

kde N značí počet simulací a $\hat{\beta}_j^{(m)}$ je odhad koeficientu $\beta_j^{(m)}$ pro m-tou simulaci, $m = 1, \ldots, N$. Provedených simulací bylo opět 100 pro každé α.

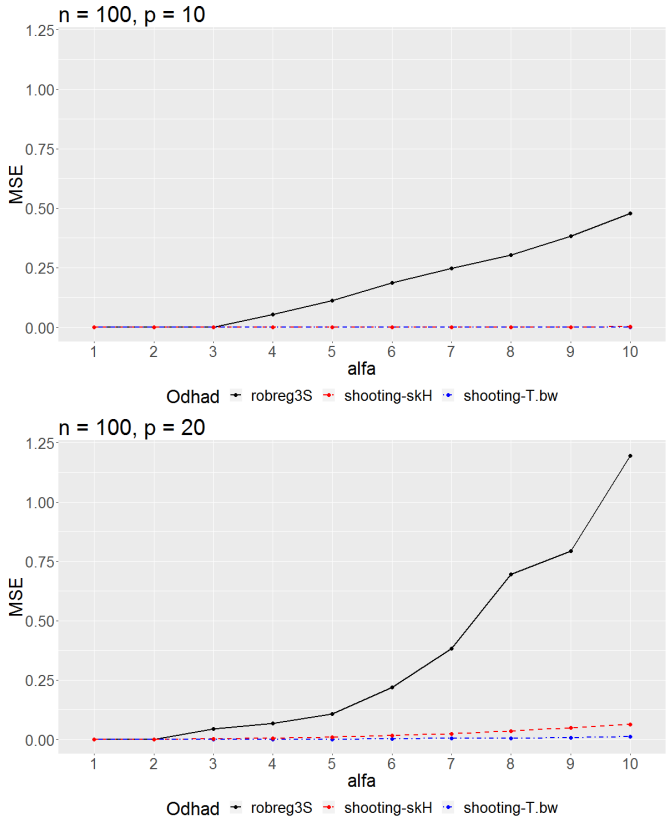
Podívejme se nejprve na dva „nejextrémnější“ datové soubory, a to nejmenší o rozměrech 50×5 a největší o rozměrech 500×10. V grafech na obrázku 3.3 je vykresleno, jak se s rostoucím číslem α měnila střední čtvercová chyba odhadů parametrů. Pro oba dva datové soubory vycházela střední čtvercová chyba velmi malá postupně pro všechna α při výpočtech všemi třemi metodami. V případě menšího datového souboru ovšem můžeme pozorovat větší výkyvy v hodnotách. Větší chyba vycházela v případě výpočtů odhadů koeficientů pomocí tříkrokové regrese. Střílející S-odhad dával nepatrně lepší výsledky při využití *Tukey's biweight* ztrátové funkce. Na druhou stranu, v případě datového souboru 500×10



Obrázek 3.3: MSE pro datové soubory o rozměrech 50×5 a 500×10.

vycházela nejmenší chyba pro odhady vypočítané pomocí tříkrokové regrese. Je nutné si ale uvědomit, že v obou případech se střední čtvercová chyba pohybovala v opravdu velmi malých hodnotách, tudíž i rozdíly mezi vypočítanými odhady jsou velmi malé.

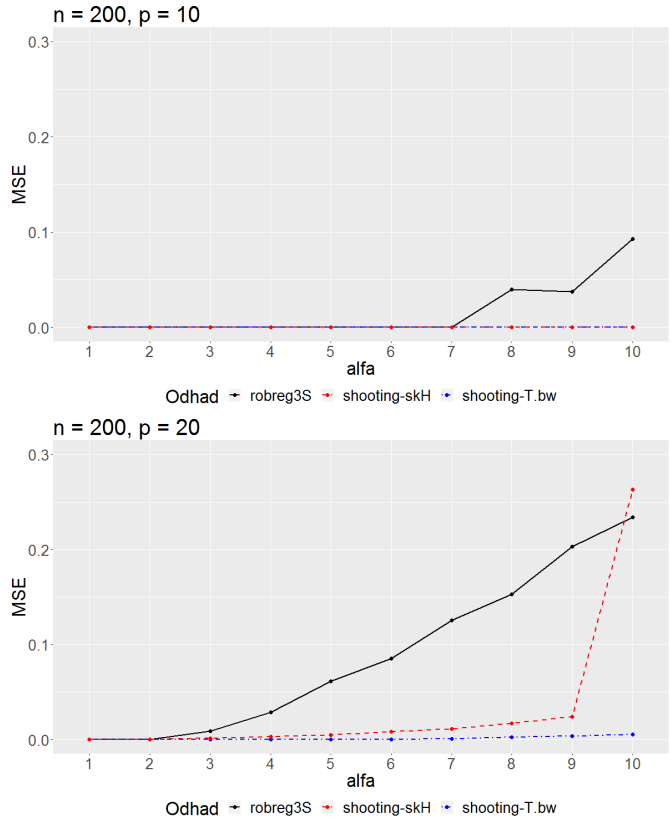
Dále se zaměříme na datové soubory se 100 pozorováními a s postupně 10 a 20 proměnnými. Střední čtvercová chyba pro tyto datové soubory je znázorněna v grafech na obrázku 3.4. Zde je situace zajímavější. V obou případech s rostoucím α výrazně rostla chyba u odhadů tříkrokové regrese. Pro datový soubor s menší dimenzí začala chyba růst od $\alpha = 3$, kdežto v případě většího datového souboru můžeme pozorovat nárůst chyby již od $\alpha = 2$. Nejvyšší hodnota střední čtvercové chyby pro menší datový soubor je 0,4790, kdežto pro datový soubor s 20 proměnnými vzrostla až na hodnotu 1,1953 (hodnoty obou chyb v případě $\alpha = 10$). Navíc



Obrázek 3.4: MSE pro datové soubory o rozměrech 100×10 a 100×20.

u datového souboru s větším počtem proměnných můžeme pozorovat lehký nárůst chyby pro střílející S-odhad se ztrátovou funkcí *skipped Huber*. V menším datovém souboru jsou chyby pro střílející S-odhad velmi malé a téměř stejné pro oba typy odhadů.

Poslední datové soubory jsou ty o rozměrech 200×10 a 200×20. Tak jako v předchozím případě, i zde jsou chyby pro střílející S-odhad v menším datovém souboru téměř shodné a velmi malé. Opět vzrostla chyba pro odhady tříkrokové regrese, ale tentokrát až od hodnoty $\alpha = 7$, pro menší hodnoty α byly chyby srovnatelné jako chyby pro střílející S-odhady. Nejvyšší hodnota chyby byla 0,0927 pro $\alpha = 10$ pro odhad tříkrokové regrese. V datovém souboru s 20 proměnnými opět pozorujeme výraznější a rychlejší nárůst chyby pro odkady tříkrokové regrese. Co se týče střílejícího S-odhadu, tak s rostoucím α chyba pro oba odhady



Obrázek 3.5: MSE pro datové soubory o rozměrech 200×10 a 200×20.

mírně narůstá (více pro odhad se *skipped Huber* ztrátovou funkcí). Velmi zajímavý je markantní nárůst chyby pro střílející S-odhad se *skipped Huber* ztrátovou funkcí pro $\alpha = 10$, kdy hodnota dosáhla 0,2633. V žádné provedené simulaci takto „skokově“ chyba nenarostla.

Závěrem tedy můžeme říct, že téměř ve všech simulacích chyba výrazněji rostla pro odhady vypočítané tříkrokovou regresí (výjimkou je datový soubor 500×10, nicméně to je datový soubor s mnoha pozorováními a málo proměnnými). Větší chybu pozorujeme také v případě, kdy vzroste počet proměnných v datovém souboru.

Kapitola 4

Aplikace metod na reálná data

V této kapitole aplikujeme popsané metody na reálný datový soubor. Tím je soubor **bodyfat** přímo implementovaný do softwaru R v knihovně **mfp**.

Datový soubor **bodyfat** obsahuje dohromady 252 pozorování (muži ve věku od 22 let do 81 let) a 16 proměnných. Zaměřuje se na procento tuku v těle jedince stanovené pomocí váhy člověka a naměřených obvodů různých částí těla. Pro analýzu byly zvoleny následující proměnné:

- brozek: hodnoty tělesného tuku v procentech
- age: věk pozorovaného jedince
- neck (N): obvod krku
- chest (CH): obvod hrudníku
- abdomen (ABD): obvod břicha
- hip (H): obvod boků
- thigh (T): obvod stehna
- knee (K): obvod kolena
- ankle (A): obvod kotníku
- biceps (B): obvod bicepsu

- forearm (F): obvod předloktí
- wrist (W): obvod zápěstí

Všechny hodnoty naměřených obvodů jednotlivých částí těla jsou uvedeny v centimetrech. Proměnná *brozek* byla před samotnou analýzou transformována pomocí *logitové transformace*, tj.:

```
> logit(bodyfat$brozek, percents=TRUE)
```

Vybraný datový soubor byl využit ke dvěma analýzám. Nejprve byl zkoumám soubor, na kterém byla provedena pouze logitová transformace proměnné *brozek*. Dále byl analyzován soubor, kde kromě transformace první proměnné byly ostatní proměnné transformovány pomocí *log-transformace*, což se jeví jako vhodnější vzhledem k původnímu (podílovému) měřítku proměnných. Datový soubor bez použití log-transformace je k nahlédnutí v příloze A.

4.1. Analýza datového souboru bez log-transformace proměnných

Nejprve tedy budeme zkoumat vybraný datový soubor, aniž bychom jeho proměnné transformovali (výjimku tvoří proměnná *brozek*). Abychom mohli použít metodu DDC, je třeba nejprve ověřit, zda datový soubor pochází z mnohorozměrného normálního rozdělení. Abychom si o tomto udělali alespoň hrubou představu, vykreslíme si Q-Q grafy pro jednotlivé proměnné. Tyto grafy jsou k nahlédnutí v příloze B.

Na základě vykreslených Q-Q grafů normalitu jednotlivých proměnných i přes některé nedostatky nezamítáme, a tudíž můžeme na datový soubor aplikovat metodu DDC. Pomocí jednoduchého příkazu

```
> DDC1<-DetectDeviatingCells(bodyfat,DDCpars)
```

provedeme příslušný výpočet. Podívejme se nyní, kolik prvků metoda označila jako odlehlé hodnoty a jestli byly jako odlehlé označeny i některé řádky.

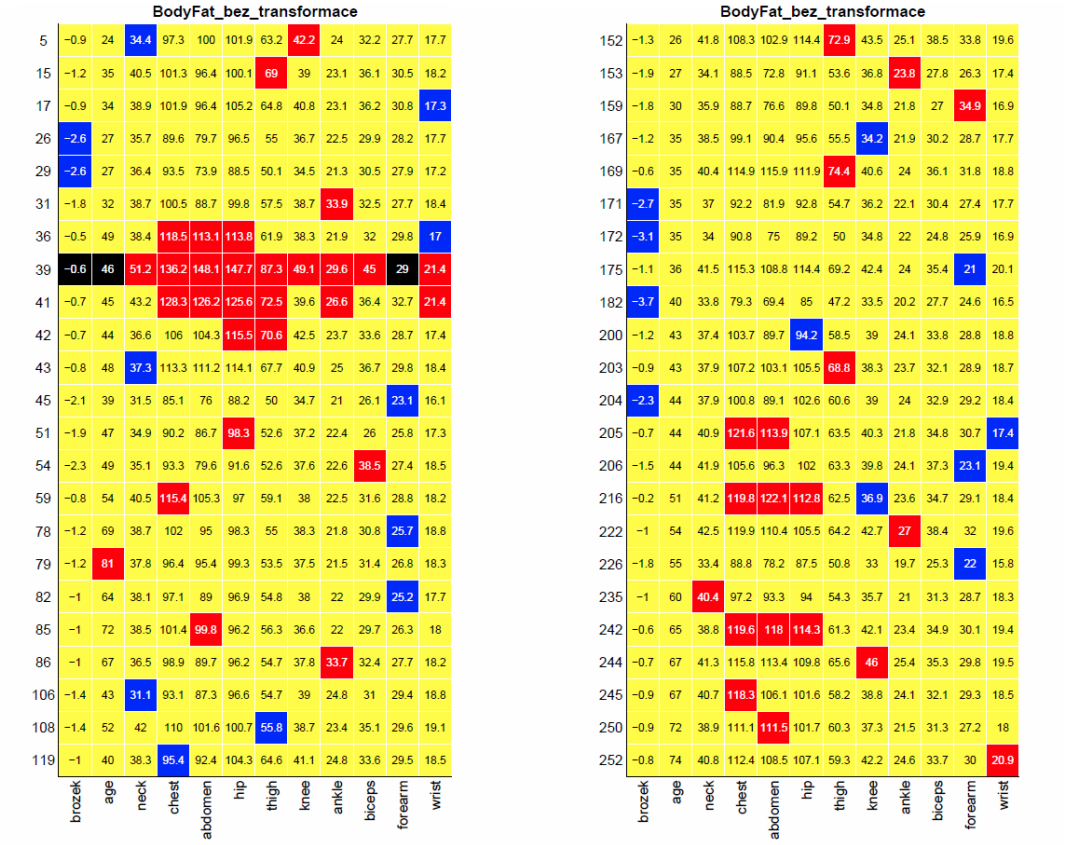
```
> I1<-DDC1$indcells  # indexy prvků, které byly analýzou
označeny jako odlehlé hodnoty
> IR1<-DDC1$indrows   # indexy řádků, které byly označeny jako odlehlé
```

Celkem bylo jako odlehlé hodnoty označeno 71 prvků. Navíc byl označen také jeden celý řádek (jedno pozorování), a to řádek číslo 39. V grafu 4.1 jsou tyto odlehlé hodnoty názorně zobrazeny. Do grafu byla vybrána pouze ta pozorování, která obsahovala alespoň jednu označenou odlehlou hodnotu. Vzorový kód pro vytvoření *cellmap* je následující:

```
> indexDDCcells1 = matrix(FALSE,nrow=n1,ncol=d1)
> indexDDCcells1[DDC1$indcells] = TRUE
> ggpcol = cellMap(D=DDC1$remX,
+                  R=DDC1$stdResid,
+                  indcells=which(indexDDCcells1==TRUE),
+                  indrows=DDC1$indrows,
+                  xlabels=colnames(DDC1$remX),
+                  ylabels=rownames(DDC1$remX),
+                  mTitle="BodyFat_bez_transformace",
+                  yshowindex=yshowindex,
+                  showVals="D",
+                  hjustYlabels=0.5)
> plot(ggpcol)
```

Poznámka 4.1.1. Symbol `n1` značí počet řádků matice `remX`, `d1` počet jejích sloupců. Parametr `showVals` určuje, zda budou v políčkách vypsány hodnoty vstupní matice dat ("D"), nebo rezidua ("R").

Červeně označené políčko znamená, že hodnota na dané pozici byla metodou vyhodnocena jako příliš vysoká. Naopak modře označené políčko je na základě výpočtů metody příliš nízká hodnota. Pokud je celý řádek označen černě, tak je dané pozorování vyhodoceno jako odlehlé. Žluté hodnoty jsou potom ty hodnoty, které nijak nevybočují.



Obrázek 4.1: Cellmap pro datový soubor bodyfat (bez transformace proměnných).

Můžeme si povšimnout, že většina vykreslených pozorování obsahuje pouze jednu označenou odlehlou hodnotu. Výrazněji se projevila pozorování 36, 39 a 41. Tato tři pozorování mají všechna zvýšené hodnoty v proměnných *chest, abdomen* a *hip*. Pozorování 39 se zdá být hodně extrémní, neboť většina jeho hodnot je označena jako zvýšené. Navíc jej metoda označila jako celý řádek. Pozornost si zaslouží i pozorování 41, neboť polovina jeho hodnot byla taktéž označena jako vysoké odlehlé hodnoty. Zajímavá je také podobnost pozorování 216 a 242, jejichž označené zvýšené hodnoty jsou shodně v proměnných *chest, abdomen*a *hip*. Nicméně oproti pozorování 39 a 41 nepůsobí jako výrazné extrémy.

Nyní přistoupíme k výpočtům koeficientů tříkrokové regrese a střílejícího

S-odhadu. Uvažujme model

$$logit(brozek) = \beta_0 + \beta_1 age + \beta_2 neck + \beta_3 chest + \beta_4 abdomen + \beta_5 hip + \beta_6 thigh + \\ + \beta_7 knee + \beta_8 ankle + \beta_9 biceps + \beta_{10} forearm + \beta_{11} wrist + \varepsilon$$

Výpočet regresních koeficientů provedeme pomocí následujících příkazů:

```
# Výpočet koeficientů tříkrokové regrese
> BF.3s<-robreg3S(y=bodyfat$brozek,x=as.matrix(subset(bodyfat,
+ select=-brozek)))
> BF.3s$coef
# Střílející S-odhad užitím funkce Tukey's biweight
> BF.shooting <- shooting(x=as.matrix(subset(bodyfat,select=-brozek)),
+ y=bodyfat$brozek)
> BF.shooting$coef
# Střílející S-odhad užitím funkce skipped Huber
> BF.shooting.Huber<-shooting(x=as.matrix(subset(bodyfat,select=-brozek)),
+ y=bodyfat$brozek, k=2.176922, method='skHuber')
>BF.shooting.Huber$coef
```

Poznámka 4.1.2. Funkce *Tukey's biweight* je pro výpočet nastavená jako výchozí. Chceme-li použít funkci *skipped Huber*, je třeba do příkazu doplnit příslušný parametr.

V tabulce 4.1 jsou uvedeny všechny vypočítané koeficienty a jejich směrodatné odchylky. Pro srovnání jsou uvedeny i hodnoty koeficientů spočítané klasickou metodou nejmenších čtverců.

Poznámka 4.1.3. Sloupec **3S** označuje koeficienty vypočítané pomocí tříkrokové regrese, sloupce **shooting - T.bw.** a **shooting - skH.** střílející S-odhady získané postupně s využitím funkcí *Tukey's biweight* a *skipped Huber* a ve sloupci **LS** jsou koeficienty vypočítané metodou nejmenších čtverců. Pro všechny koeficienty jsou potom ve sloupcích označených jako **sd** uvedeny jejich směrodatné odchylky.

	3S		shooting - T.bw.		shooting - skH.		LS	
Proměnná	**koef.**	**sd**	**koef.**	**sd**	**koef.**	**sd**	**koef.**	**sd**
(Intercept)	-2,6070	0,0186	-2,6207	0,0164	-2,6441	0,0162	-2,6213	0,0182
age	0,0057	0,0015	0,0056	0,0013	0,0049	0,0013	0,0059	0,0014
neck	-0,0311	0,0077	-0,0336	0,0068	-0,0346	0,0067	-0,0422	0,0075
chest	-0,0090	0,0022	-0,0072	0,0019	-0,0068	0,0019	-0,0044	0,0022
abdomen	0,0523	0,0017	0,0490	0,0015	0,0496	0,0015	0,0529	0,0017
hip	-0,0168	0,0026	-0,0115	0,0023	-0,0148	0,0023	-0,0271	0,0025
thigh	0,0225	0,0035	0,0178	0,0031	0,0202	0,0031	0,0263	0,0035
knee	-0,0086	0,0077	-0,0228	0,0068	-0,0202	0,0067	-0,0044	0,0076
ankle	-0,0071	0,0110	0,0103	0,0097	0,0111	0,0096	0,0041	0,0108
biceps	0,0187	0,0062	0,0117	0,0054	0,0081	0,0054	0,0072	0,0060
forearm	0,0053	0,0092	0,0207	0,0081	0,0226	0,0080	0,0291	0,0090
wrist	-0,0941	0,0199	-0,0985	0,0176	-0,0914	0,0173	-0,0971	0,0196

Tabulka 4.1: Odhadnuté regresní koeficienty pro datový soubor bodyfat.

Pro srovnání odhadů ještě vypočítáme jejich čtvercové normované vzdálenosti [10]

$$n \cdot \sum_{j=1}^{p} (\hat{\beta}_{j,A} - \hat{\beta}_{j,B})^2 \cdot MAD(x_{1j}, \ldots, x_{nj})^2.$$

Jejich přehled je v tabulce 4.2.

Odhad	**3S**	**shooting - T.bw.**	**shooting - skH.**	**LS**
3S	-	1,5001	1,2852	2,4953
shooting - T.bw.		-	0,2079	3,6925
shooting - skH.			-	2,3973
LS				-

Tabulka 4.2: Čtvercové normované vzdálenosti pro dvojice vektorů koeficientů.

Nejprve se budeme věnovat výsledkům v tabulce 4.1. Při pohledu na vypočítané směrodatné odchylky jednotlivých koeficientů si můžeme povšimnout, že koeficienty vypočítané pomocí tříkrokové regrese a metody nejmenších čtverců

mají ve všech případech směrodatnou odchylku nepatrně vyšší než ty, které byly vypočítány pomocí střílejícího S-odhadu (ty se od sebe liší pouze minimálně). Ve všech případech vyšla největší směrodatná odchylka pro proměnnou *wrist*.

Co se týče samotných regresních koeficientů, ve všech případech vyšla nejvyšší hodnota pro proměnnou *abdomen*. Z toho můžeme usuzovat, že se tuk ve většině případů nejvíce ukládá v oblasti břicha, a tudíž se obvod břicha výrazně promítne na konečné hodnotě procent tuku v těle. Podobně se projevily i proměnné *thigh* a *forearm*. Zajímavé ovšem je, že koeficient pro proměnnou *forearm* vyšel výrazně menší při výpočtu tříkrokovou regresí než při výpočtech zbývajícími metodami. Za povšimnutí stojí taktéž rozdíl v koeficientech proměnné *ankle*, neboť pomocí tříkrokové regrese vyšel tento koeficient záporný, ale při výpočtech ostatními metodami vyšel vždy kladný.

Podívejme se nyní ještě na čtvercové normované vzdálenosti pro jednotlivé dvojice vektorů koeficientů. Podle očekávání jsou všechny robustní odhady nejvíce odlišné právě od odhadů pomocí metody nejmenších čtverců. Nejblíže jsou si oba střílející S-odhady, což se dalo očekávat, neboť jde o výpočetně stejné odhady, rozdíl je pouze v použité ztrátové funkci. Největší rozdíl můžeme pozorovat mezi střílejícím S-odhadem s *Tukey's biweight* funkcí a odhadem nejmenších čtverců. Odhady vypočítené pomocí tříkrokové regrese jsou o něco blíž střílejícím S-odhadům s využitím Huberovy funkce než s využitím *Tukey's biweight*. Tento rozdíl nicméně není příliš výrazný.

4.2. Analýza datového souboru po log transformaci

V této části budeme analyzovat opět datový soubor bodyfat, ale s tím rozdílem, že všechny proměnné transformujeme pomocí log-transformace (na proměnnou *brozek* je opět použita logitová transformace).

Tak jako v předchozí podkapitole, opět si nejdříve vykreslíme Q-Q grafy (v příloze C) , abychom ověřili normalitu jednotlivých proměnných a mohli tak na datový soubor aplikovat metodu DDC.

Normalitu tak jako u dat bez log-transformace nezamítáme. Přejdeme tedy

k identifikaci odlehlých hodnot pomocí metody DDC. Kód je analogický jako pro datový soubor bez transformace.

```
> DDC2<-DetectDeviatingCells(bodyfat.T,DDCpars)
```

Opět si zobrazíme, které hodnoty byly metodou označeny jako odlehlé.

```
I2<-DDC2$indcells
IR2<-DDC2$indrows
```

Nyní metoda označila jako odlehlé hodnoty celkem 67 prvků, což je sice méně než v předchozím výpočtu, ale nikterak výrazně. Výraznějším rozdílem je fakt, že byly označeny dohromady dva celé řádky jako odlehlé. K pozorování 39, které bylo takto označeno v předchozím výpočtu, přibylo i pozorování číslo 3. Výsledky si opět graficky zobrazíme na obrázku 4.2.

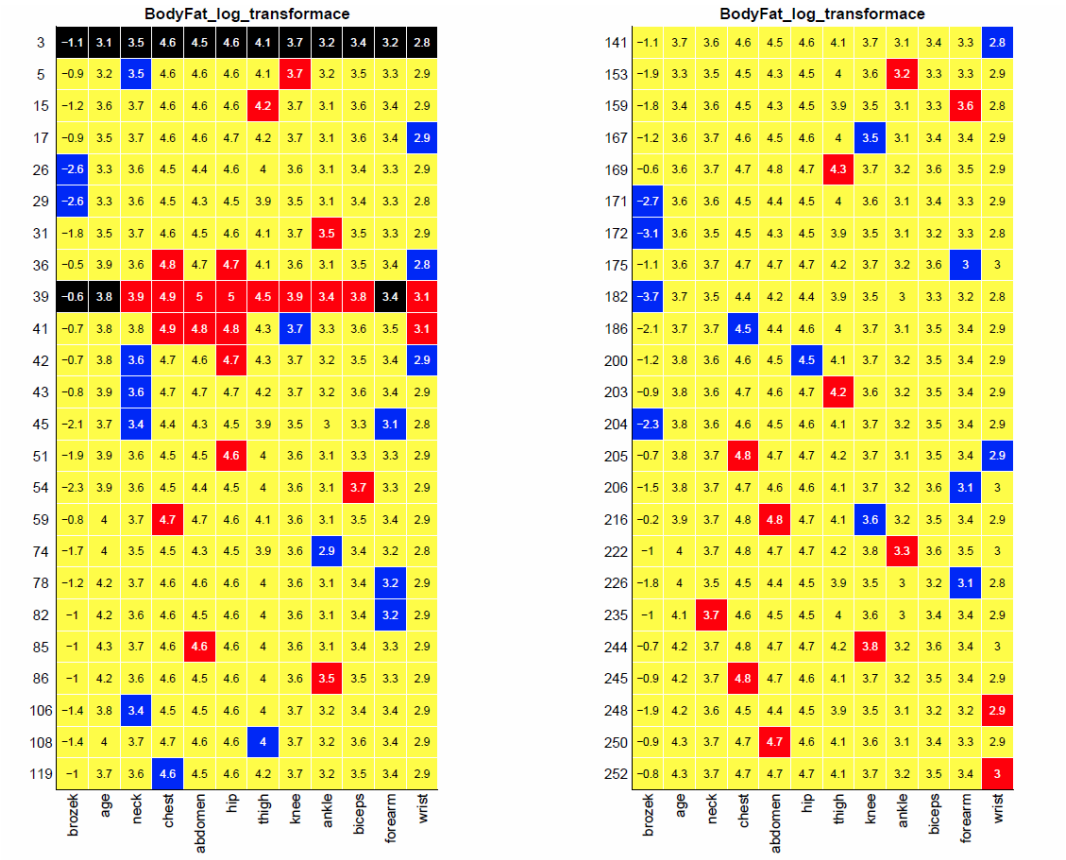
Podobně jako v předchozích výpočtech, i zde většina zobrazených pozorování obsahuje pouze jednu či dvě hodnoty označené jako odlehlé. U pozorování 36 a 41 již nepozorujeme tolik hodnot označených jako výrazně zvýšené, jako tomu bylo v předchozích výpočtech. Dále také vypadla podobnost pozorování 216 a 242, u kterých dříve vyšly tři odlehlé hodnoty ve shodných proměnných. Pro pozorování 242 metoda nyní neoznačila žádnou hodnotu jako odlehlou, a tudíž není ani vykreslené v grafu.

Dále budeme opět pokračovat výpočtem koeficientů tříkrokové regrese a stříleJícího S-odhadu. Uvažujme nyní model

$$\begin{aligned} logit(brozek) = \beta_0 + \beta_1 log(age) + \beta_2 log(neck) + \beta_3 log(chest) + \beta_4 log(abdomen) + \\ + \beta_5 log(hip) + \beta_6 log(thigh) + \beta_7 log(knee) + \beta_8 log(ankle) + \\ + \beta_9 log(biceps) + \beta_{10} log(forearm) + \beta_{11} log(wrist) + \varepsilon \end{aligned}$$

Kód pro výpočty odhadů regresních koeficientů je analogií kódu z předchozí podkapitoly.

```
# Výpočet koeficientů tříkrokové regrese
> BF.T.3s<-robreg3S(y=bodyfat.T$brozek,x=as.matrix(subset(bodyfat.T,
```



Obrázek 4.2: Cellmap pro datový soubor bodyfat (log-transformace proměnných).

```
+ select=-brozek)))
> BF.T.3s$coef
# Střílející S-odhad užitím funkce Tukey's biweight
> BF.T.shooting <- shooting(x=as.matrix(subset(bodyfat.T,select=-brozek)),
+ y=bodyfat.T$brozek)
> BF.T.shooting$coef
# Střílející S-odhad užitím funkce skipped Huber
> BF.T.shooting.Huber<-shooting(x=as.matrix(subset(bodyfat.T,
+ select=-brozek)), y=bodyfat.T$brozek, k=2.176922, method='skHuber')
> BF.T.shooting.Huber$coef
```

Přehled všech vypočítaných koeficientů společně s koeficienty vypočítanými me-

todou nejmenších čtverců je uveden v tabulce 4.3.

	3S		shooting - T.bw.		shooting - skH.		LS	
Proměnná	**koef.**	**sd**	**koef.**	**sd**	**koef.**	**sd**	**koef.**	**sd**
(Intercept)	-9,2461	0,0177	-9,0604	0,0154	9,5801	0,0135	-9,3900	0,0175
age	0,2209	0,0607	0,2213	0,0528	0,1966	0,0463	0,2209	0,0601
neck	-1,1935	0,2802	-1,2838	0,2437	-1,5687	0,2135	-1,5758	0,2771
chest	-0,8575	0,2165	-0,5926	0,1882	-0,4605	0,1649	-0,4394	0,2141
abdomen	4,8529	0,1565	4,5985	0,1361	4,5996	0,1193	5,0054	0,1548
hip	-1,5249	0,2572	-1,2323	0,2236	-1,2424	0,1960	-2,2181	0,2543
thigh	1,0888	0,2054	0,9761	0,1786	1,1370	0,1565	1,4225	0,2031
knee	-0,3541	0,2872	-0,8177	0,2497	-0,9878	0,2189	-0,4444	0,2840
ankle	-0,1471	0,2571	0,1162	0,2235	0,2845	0,1959	0,0737	0,2543
biceps	0,6262	0,1901	0,3578	0,1653	0,3894	0,1449	0,2775	0,1880
forearm	0,0771	0,2480	0,5899	0,2156	0,5778	0,1890	0,6093	0,2453
wrist	-1,6376	0,3483	-1,8879	0,3028	-1,7329	0,2654	-1,7086	0,3444

Tabulka 4.3: Odhadnuté regresní koeficienty pro datový soubor bodyfat po log-transformaci proměnných.

Nejprve se opět podíváme na směrodatné odchylky jednotlivých koeficientů. Tak jako v předchozím případě, i nyní mají odhady spočítané tříkrokovou regresí a metodou nejmenších čtverců větší směrodatné odchylky než ty vypočítané pomocí střílejícího S-odhadu. Nicméně nyní můžeme pozorovat větší rozdíl mezi směrodatnými odchylkami pro odhady spočítané střílejícím S-odhadem než v předchozím případě. Největší směrodatnou odchylku opět pozorujeme u koeficientů proměnné *wrist*.

Podívejme se nyní na samotné hodnoty regresních koeficientů. Také v tomto případě můžeme na základě spočítaných koeficientů usuzovat, že nejvíce se tuk ukládá v oblasti břicha. Podobně je to také s proměnnými *thigh* a *forearm*. Shodně s předchozím případem také pozorujeme, že koeficient proměnné *forearm* opět vyšel výpočtem tříkrokové regrese překvapivě nízký vzhledem k výpočtům pomocí ostatních metod. Taktéž si opět povšimněme rozdílu ve znaménku koeficientů pro proměnnou *ankle*. Tříkrokovou regresí jsme opět získali koeficient záporný, kdežto

ostatní metody jej určily jako kladný.

Odhad	**3S**	**shooting - T.bw.**	**shooting - skH.**	**LS**
3S	-	1,2223	1,6729	1,7921
shooting - T.bw.		-	0,2327	2,0318
shooting - skH.			-	1,8796
LS				-

Tabulka 4.4: Čtvercové normované vzdálenosti pro dvojice vektorů koeficientů.

Oproti výpočtům v předchozí podkapitole si můžeme v tabulce 4.4 povšimnout, že nyní je větší rozdíl mezi koeficienty tříkrokové regrese a střílejícími S-odhady se *skipped Huber* funkcí. Tento rozdíl ale není příliš výrazný. Tak jako v předchozím případě, i nyní je největší rozdíl mezi koeficienty střílejícího S-odhadu s *Tukey's biweight* funkcí a koeficienty odhadnutými pomocí metody nejmenších čtverců.

Závěrem ještě srovnáme výsledky regrese z této a předchozí podkapitoly. Shodně vyšlo, že se tuk patrně nejvíce ukládá v oblasti břicha, stehen a předloktí, přičemž břicho vyšlo v obou případech jako „nejrizikovější partie“. Výraznější rozdíly mezi výsledky obou výpočtů se neobjevily, pouze se mírně lišily efekty některých vysvětlujících proměnných jakož i čtvercové normované vzdálenosti mezi vektory regresních koeficientů.

Závěr

První kapitolu této práce jsem věnovala základům robustní statistiky, aby si i čtenář, který se zatím s touto teorií nesetkal, udělal základní přehled o dané oblasti statistiky. Dále jsem se věnovala problematice odlehlých pozorování a představila dva modely kontaminace dat.

Ve druhé kapitole jsem se věnovala třem robustním metodám pro práci s datovými soubory obsahujícími odlehlé hodnoty. Postupně jsem popsala metodu Detect Deviating Cells, střílející S-odhad a tříkrokovou regresi. Pro každou metodu jsem také popsala její algoritmus. Z těchto tří mi jako nejvíce intuitivní přišel algoritmus metody Detect Deviating Cells.

Ve třetí kapitole jsem se zaměřila na simulační studii. Vytvořila jsem si několik datovým souborů, na které jsem následně aplikovala metody popsané ve druhé kapitole a výsledky zhodnotila. Překvapilo mě, že odhady spočítané pomocí tříkrokové regrese se v simulační studii ukázaly mnohem více náchylné na výskyt odlehlých hodnot v datovém souboru než odhady spočítané pomocí střílejícího S-odhadu. Nečekaným zjištěním byl také náhlý skokový nárůst MSE střílejícího S-odhadu s funkcí *skipped Huber* v datovém souboru o rozměrech 200×20 při nejvyšší použité kontaminaci dat.

V poslední, čtvrté kapitole, jsem se věnovala aplikacím popsaných metod na reálný datový soubor. Provedla jsem dvě analýzy, nejprve pro datový soubor bez transformace proměnných a poté pro datový soubor, jehož proměnné byly transformovány log-transformací. Výsledky se lišily především v počtu řádků, které metoda DDC označila jako odlehlé. Při použití metody DDC na datový soubor bez použití log-transformace se také více pozorování jevilo jako nestan-

dardní.

Tato práce pro mě byla velmi přínosná, neboť jsem prostudovala nové metody, se kterými jsem se dříve nesetkala. Výzvou pro mě byla simulační studie, protože jsem ji doposud neprováděla. Navíc jsem několikrát čelila problému, že vytvořený datový soubor nebyl vhodných rozměrů, což způsobilo výpočetní selhání metod a musela jsem vytvořit datový soubor jiný. Velmi se mi líbil grafický výstup metody DDC, který je názorný a snadno interpretovatelný.

Příloha A

Datový soubot *bodyfat* bez transformace proměnných (mimo *brozek*).

brozek	**age**	**N**	**CH**	**ABD**	**H**	**T**	**K**	**A**	**B**	**F**	**W**
-1,7768	23	36,2	93,1	85,2	94,5	59	37,3	21,9	32	27,4	17,1
-2,3069	22	38,5	93,6	83	98,7	58,7	37,3	23,4	30,5	28,9	18,2
-1,0527	22	34	95,8	87,9	99,2	59,6	38,9	24	28,8	25,2	16,6
-1,9138	26	37,4	101,8	86,4	101,2	60,1	37,3	22,8	32,4	29,4	18,2
-0,8998	24	34,4	97,3	100	101,9	63,2	42,2	24	32,2	27,7	17,7
-1,2616	24	39	104,5	94,4	107,8	66	42	25,6	35,7	30,6	18,8
-1,3523	26	36,4	105,1	90,7	100,3	58,4	38,3	22,9	31,9	27,8	17,7
-1,7615	25	37,8	99,6	88,5	97,1	60	39,4	23,2	30,5	29	18,8
-2,5349	25	38,1	100,9	82,5	99,9	62,9	38,3	23,8	35,9	31,1	18,2
-1,8236	23	42,1	99,6	88,6	104,1	63,1	41,7	25	35,6	30	19,2
-2,2396	26	38,5	101,5	83,6	98,2	59,7	39,7	25,2	32,8	29,4	18,5
-2,1349	27	39,4	103,6	90,9	107,7	66,2	39,2	25,9	37,2	30,2	19
-1,2671	32	38,4	102	91,6	103,9	63,4	38,3	21,5	32,5	28,6	17,7
-1,2506	30	39,4	104,1	101,8	108,6	66	41,5	23,7	36,9	31,6	18,8
-1,2018	35	40,5	101,3	96,4	100,1	69	39	23,1	36,1	30,5	18,2
-1,2671	35	36,4	99,1	92,8	99,2	63,1	38,7	21,7	31,1	26,4	16,9
-0,8859	34	38,9	101,9	96,4	105,2	64,8	40,8	23,1	36,2	30,8	17,3
-1,1648	32	42,1	107,6	97,5	107	66,9	40	24,4	38,2	31,6	19,3
-1,5303	28	38	106,8	89,6	102,4	64,2	38,7	22,9	37,2	30,5	18,5
-1,5045	33	40	106,2	100,5	109	65,8	40,6	24	37,1	30,1	18,2
-1,3523	28	39,1	103,3	95,9	104,9	63,5	38	22,1	32,5	30,3	18,4
-1,5831	28	41,3	111,4	98,8	104,8	63,4	40,6	24,6	33	32,8	19,9
-1,5565	31	33,9	86	76,4	94,6	57,4	35,3	22,2	27,9	25,9	16,7
-1,4358	32	35,5	86,7	80	93,4	54,9	36,2	22,1	29,8	26,7	17,1

brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-1,659	28	34,5	90,2	76,3	95,8	58,4	35,5	22,9	31,1	28	17,6
-2,6068	27	35,7	89,6	79,7	96,5	55	36,7	22,5	29,9	28,2	17,7
-2,1349	34	36,2	88,6	74,6	85,3	51,7	34,7	21,4	28,7	27	16,5
-1,1648	31	38,8	97,4	88,7	94,7	57,5	36	21	29,2	26,6	17
-2,5921	27	36,4	93,5	73,9	88,5	50,1	34,5	21,3	30,5	27,9	17,2
-2,0476	29	36,7	97,4	83,5	98,7	58,9	35,3	22,6	30,1	26,7	17,6
-1,8	32	38,7	100,5	88,7	99,8	57,5	38,7	33,9	32,5	27,7	18,4
-2,354	29	37,3	93,5	84,5	100,6	58,5	38,8	21,5	30,1	26,4	17,9
-1,7167	27	38,1	93	79,1	94,5	57,3	36,2	24,5	29	30	18,8
-1,2451	41	39,8	111,7	100,5	108,3	67,1	44,2	25,2	37,5	31,5	18,7
-0,7517	41	42,1	117	115,6	116,1	71,2	43,3	26,3	37,3	31,7	19,7
-0,4561	49	38,4	118,5	113,1	113,8	61,9	38,3	21,9	32	29,8	17
-1,1029	40	38,5	106,5	100,9	106,2	63,5	39,9	22,6	35,1	30,6	19
-0,9137	50	42,1	105,6	98,8	104,8	66	41,5	24,7	33,2	30,5	19,4
-0,6362	46	51,2	136,2	148,1	147,7	87,3	49,1	29,6	45	29	21,4
-0,743	50	40,2	114,8	108,1	102,5	61,3	41,1	24,7	34,1	31	18,3
-0,6657	45	43,2	128,3	126,2	125,6	72,5	39,6	26,6	36,4	32,7	21,4
-0,7257	44	36,6	106	104,3	115,5	70,6	42,5	23,7	33,6	28,7	17,4
-0,7824	48	37,3	113,3	111,2	114,1	67,7	40,9	25	36,7	29,8	18,4
-0,7648	41	41,5	106,6	104,3	106	65	40,2	23	35,8	31,5	18,8
-2,145	39	31,5	85,1	76	88,2	50	34,7	21	26,1	23,1	16,1
-1,6661	43	35,7	96,6	81,5	97,2	58,4	38,2	23,4	29,7	27,4	18,3
-1,8886	40	33,6	88,2	73,7	88,5	53,3	34,5	22,5	27,9	26,2	17,3
-2,366	39	34,6	89,8	79,5	92,7	52,7	37,5	21,9	28,8	26,8	17,9
-1,7167	45	32,8	92,3	83,4	90,4	52	35,8	20,6	28,8	25,5	16,3
-2,5489	47	34	83,4	70,4	87,2	50,6	34,4	21,9	26,8	25,8	16,8
-1,9309	47	34,9	90,2	86,7	98,3	52,6	37,2	22,4	26	25,8	17,3
-2,2506	40	34,3	89,2	77,9	91	51,4	34,9	21	26,7	26,1	17,2
-2,115	51	36,5	89,7	82	89,1	49,3	33,7	21,4	29,6	26	16,9
-2,2841	49	35,1	93,3	79,6	91,6	52,6	37,6	22,6	38,5	27,4	18,5
-2,5631	42	37,8	87,6	77,6	88,6	51,9	34,9	22,5	27,7	27,5	18,5
-1,1753	54	39,9	107,6	100	99,6	57,2	38	22	35,9	30,2	18,9

brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-1,2894	58	39,1	100	99,8	102,5	62,1	39,6	22,5	33,1	28,3	18,5
-0,9323	62	40,5	111,5	104,2	105,8	61,8	39,8	22,7	37,7	30,9	19,2
-0,7824	54	40,5	115,4	105,3	97	59,1	38	22,5	31,6	28,8	18,2
-1,0827	61	38,4	104,8	98,3	99,6	60,6	37,7	22,9	34,5	29,6	18,5
-1,0135	62	41,4	112,3	104,8	103,1	61,6	40,9	23,1	36,2	31,8	20,2
-0,854	56	35,6	102,9	94,7	100,8	60,9	38	22,1	32,5	29,8	18,3
-0,8179	54	38	107,6	102,4	99,4	61	39,4	23,6	32,7	29,9	19,1
-1,0281	61	37,4	105,3	99,7	99,7	60,8	40,1	22,7	33,6	29	18,8
-0,7561	57	40,1	105,3	105,5	108,3	65	41,2	24,7	35,3	31,1	18,4
-0,8494	55	40,9	103	100,3	104,2	64,8	40,2	22,7	34,8	30,1	18,7
-1,2342	54	35,6	90	83,9	93,9	55	36,1	21,7	29,6	27,4	17,4
-1,6732	55	36,9	95,4	86,6	91,8	54,3	35,4	21,5	32,8	27,4	18,7
-2,2841	54	37,5	89,3	78,4	96,1	56	37,4	22,4	32,6	28,1	18,1
-1,7315	55	36,3	94,4	84,6	94,3	51,2	37,4	21,6	27,3	27,1	17,3
-1,0978	62	35,5	97,6	91,5	98,5	56,6	38,6	22,4	31,5	27,3	18,6
-2,0476	55	38,7	88,5	82,8	95,5	58,9	37,6	21,6	30,3	27,3	18,3
-2,076	56	36,4	93,6	82,9	96,3	52,9	37,5	23,1	29,7	27,3	18,2
-1,6948	55	33,2	87,7	76	88,6	50,9	35,4	19,1	29,3	25,7	16,9
-1,8236	61	36,5	93,4	83,3	93	55,5	35,2	20,9	29,4	27	16,8
-1,3935	61	36	91,6	81,8	94,8	54,5	37	21,4	29,3	27	18,3
-2,0665	57	38,7	91,6	78,8	94,3	56,7	39,7	24,2	30,2	29,2	18,1
-1,2018	69	38,7	102	95	98,3	55	38,3	21,8	30,8	25,7	18,8
-1,2342	81	37,8	96,4	95,4	99,3	53,5	37,5	21,5	31,4	26,8	18,3
-1,3757	66	37,4	102,7	98,6	100,2	56,5	39,3	22,7	30,3	28,7	19
-0,7913	67	38,4	97,7	95,8	97,1	54,8	38,2	23,7	29,4	27,2	19
-0,9845	64	38,1	97,1	89	96,9	54,8	38	22	29,9	25,2	17,7
-1,3995	64	39,3	103,1	97,8	99,6	58,9	39	23	34,3	29,6	19
-0,9749	70	38,7	101,8	94,9	95	56	36,5	24,1	31,2	27,3	19,2
-0,9797	72	38,5	101,4	99,8	96,2	56,3	36,6	22	29,7	26,3	18
-0,9941	67	36,5	98,9	89,7	96,2	54,7	37,8	33,7	32,4	27,7	18,2
-1,6034	72	37,7	97,5	88,1	96,9	57,2	37,7	21,8	32,6	28	18,8
-1,1543	64	36,5	104,3	90,9	93,8	57,8	39,5	23,3	29,2	28,4	18,1

brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-2,1051	46	38	97,3	86	99,3	61	38,4	23,8	30,2	29,3	18,8
-1,6519	48	36,7	96,7	86,5	98,3	60,4	39,9	24,4	28,8	29,6	18,7
-1,2838	46	37,2	99,7	95,6	102,2	58,3	38,2	22,5	29,1	27,7	17,7
-1,4055	44	39,2	101,9	93,2	100,6	58,9	39,7	23,1	31,4	28,4	18,8
-2,0665	47	37,5	97,2	83,1	95,4	56,9	38,3	22,1	30,1	28,2	18,4
-1,0726	46	38	106,6	97,5	100,6	58,9	40,5	24,5	33,3	29,6	19,1
-2,0289	47	37,3	99,6	88,8	101,4	57,4	39,6	24,6	30,3	27,9	17,8
-1,4542	53	41,1	113,2	99,2	107,5	61,7	42,3	23,2	32,9	30,8	20,4
-1,9835	38	37,5	99,1	91,6	102,4	60,6	39,4	22,9	31,6	30,1	18,5
-1,897	50	38,7	99,4	86,7	96,2	62,1	39,3	23,3	30,6	27,8	18,2
-1,4297	46	35,9	95,1	88,2	92,8	54,7	37,3	21,9	31,6	27,5	18,2
-1,2018	47	40	107,5	94	103,7	62,7	39	22,3	35,3	30,9	18,3
-1,2506	49	40,1	106,5	95	101,7	59	39,4	22,3	32,2	31	18,6
-1,2894	48	37	99,1	92	98,3	59,3	38,4	22,4	27,9	26,2	17
-1,3063	41	36,3	96,7	89,2	98,3	60	38,4	23,2	31	29,2	18,4
-1,1912	49	40,7	103,5	95,5	101,6	59,1	39,8	25,4	31	30,3	19,7
-1,0478	43	39,6	104	98,6	99,5	59,5	36,1	22	30,1	27,2	17,7
-1,4236	43	31,1	93,1	87,3	96,6	54,7	39	24,8	31	29,4	18,8
-1,3465	43	38,6	105,2	102,8	103,6	61,2	39,3	23,5	30,5	28,5	18,1
-1,3995	52	42	110	101,6	100,7	55,8	38,7	23,4	35,1	29,6	19,1
-1,4604	43	38,5	110,1	88,7	102,1	57,5	40	24,8	35,1	30,7	19,2
-1,2396	40	34,2	97,8	92,3	100,6	57,5	36,8	22,8	32,1	26	17,3
-1,3234	43	37,2	96,3	90,6	99,3	61,9	38	22,3	33,3	28,2	18,1
-0,9323	43	37,1	108	105	103	63,7	40	23,6	33,5	27,8	17,4
-1,2072	47	40,2	99,7	95	98,6	62,3	38,1	23,9	35,3	31,1	19,8
-1,2451	42	35,3	93,5	89,6	99,8	61,5	37,8	21,9	30,7	27,6	17,4
-0,9893	48	38	100,7	92,4	97,5	59,3	38,1	21,8	31,8	27,3	17,5
-1,4918	40	36,3	97	86,6	92,6	55,9	36,3	22,1	29,8	26,3	17,3
-1,3063	48	36,8	96	90	99,7	58,8	38,4	22,8	29,9	28	18,1
-1,6661	51	41	99,2	90	96,4	56,8	38,8	23,3	33,4	29,8	19,5
-1,0281	40	38,3	95,4	92,4	104,3	64,6	41,1	24,8	33,6	29,5	18,5
-1,4175	44	38	101,8	87,5	101	58,5	39,2	24,5	32,1	28,6	18

brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-0,937	52	40,8	104,3	99,2	104,1	58,5	39,3	24,6	33,9	31,2	19,5
-1,0527	44	39,5	99,2	98,1	101,4	57,1	40,5	23,2	33	29,6	18,4
-1,6171	40	36,9	99,3	83,3	97,5	60,5	38,7	22,6	34,4	28	17,6
-1,5368	47	36,9	94	86,1	95,2	58,1	36,5	22,1	30,6	27,5	17,6
-1,6732	50	37,7	98,9	84,1	94	58,5	36,6	23,5	34,4	29,2	18
-1,4481	46	36,6	101	89,9	100	60,7	36	21,9	35,6	30,2	17,6
-0,9654	42	38,9	98,7	92,1	98,5	60,7	36,8	22,2	33,8	30,3	17,2
-1,4481	43	37,5	95,9	78	93,2	53,5	35,8	20,8	33,9	28,2	17,4
-1,2727	40	39,8	103,9	93,5	99,5	61,7	39	21,8	33,3	29,6	18,1
-1,6034	42	38,3	96,2	87	97,8	57,4	36,9	22,2	31,6	27,8	17,7
-1,4115	49	35,5	97,8	90,1	95,8	57	38,7	23,2	27,5	26,5	17,6
-1,1753	40	36,3	94,6	90,3	99,1	60,3	38,5	23	31,2	28,4	17,1
-1,1284	47	37,8	103,6	99,8	103,2	61,2	38,1	22,6	33,5	28,6	17,9
-1,0183	50	37,8	100,4	89,4	92,3	56,1	35,6	20,5	33,6	29,3	17,3
-1,0928	41	36,5	98,4	87,2	98,4	56	36,9	23	34	29,8	18,1
-0,9702	44	37,8	104,6	101,1	102,1	58,9	37,9	22,7	30,9	28,8	17,6
-1,2179	39	37	92,9	86,1	95,6	58,8	36,1	22,4	32,7	28,3	17,1
-0,8722	43	37,7	97,8	98,6	100,6	63,6	39,2	23,8	34,3	28,4	17,7
-1,1965	40	34,3	98,3	88,5	98,3	58,1	38,4	22,5	31,7	27,4	17,6
-1,2894	49	40,8	104,7	106,6	107,7	66,5	42,5	24,5	35,5	29,8	18,7
-1,0677	40	37,4	98,6	93,1	101,6	59,1	39,6	21,6	30,8	27,9	16,6
-1,4055	40	36,5	99,5	93	99,3	60,4	38,2	22	32	28,5	17,8
-1,1491	52	37,5	102,7	91	98,9	57,1	36,7	22,3	31,6	27,5	17,9
-2,0015	23	35,5	92,1	77,1	93,9	56,1	36,1	22,7	30,5	27,2	18,2
-1,9223	23	38	96,6	85,3	102,5	59,1	37,6	23,2	31,8	29,7	18,3
-1,6449	24	35,7	92,7	81,9	95,3	56,4	36,5	22	33,5	28,3	17,3
-1,3523	24	39,2	102	99,1	110,1	71,2	43,5	25,2	36,1	30,3	18,7
-0,8631	25	40,9	110,9	100,5	106,2	68,4	40,8	24,6	33,3	29,7	18,4
-2,4029	25	35,2	92,3	76,5	92,1	51,9	35,7	22	25,8	25,2	16,9
-1,0577	26	40,6	114,1	106,8	113,9	67,6	42,7	24,7	36	30,4	18,4

brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-2,0015	26	35,4	92,9	77,6	93,5	56,9	35,9	20,4	31,6	29	17,8
-1,3465	26	41,8	108,3	102,9	114,4	72,9	43,5	25,1	38,5	33,8	19,6
-1,9395	27	34,1	88,5	72,8	91,1	53,6	36,8	23,8	27,8	26,3	17,4
-1,5045	27	37,9	94	88,2	95,2	56,8	37,4	22,8	30,6	28,3	17,9
-1,2671	27	38,2	101,1	100,1	105	62,1	40	24,9	33,7	29,2	19,4
-1,4604	28	35,6	92,1	83,5	98,3	57,3	37,8	21,7	32,2	27,7	17,7
-0,7957	28	38,5	105,6	105	106,4	68,6	40	25,2	35,2	30,7	19,1
-1,9482	28	37	98,5	90,8	102,5	60,8	38,5	25	31,6	28	18,6
-1,7615	30	35,9	88,7	76,6	89,8	50,1	34,8	21,8	27	34,9	16,9
-1,1859	31	36,2	101,1	92,4	99,3	59,4	39	24,6	30,1	28,2	18,2
-2,0015	31	35	94	81,2	91,5	52,5	36,6	21	27	26,3	16,5
-1,6171	33	38,5	103,8	95,6	105,1	61,4	40,6	25	31,3	29,2	19,1
-1,7241	33	40,7	98,9	92,1	103,5	64	37,3	23,5	33,5	30,6	19,7
-1,5899	34	36	89,2	83,4	89,6	52,4	35,6	20,4	28,3	26,2	16,5
-0,9606	34	39,5	111,4	106	108,8	63,8	42	23,4	34	31,2	18,5
-1,3523	35	40,5	107,5	95,1	104,5	64,8	41,3	25,6	36,4	33,7	19,4
-1,2179	35	38,5	99,1	90,4	95,6	55,5	34,2	21,9	30,2	28,7	17,7
-1,295	35	43,9	108,2	100,4	106,8	63,3	41,7	24,6	37,2	33,1	19,8
-0,5987	35	40,4	114,9	115,9	111,9	74,4	40,6	24	36,1	31,8	18,8
-1,5045	35	37,6	99,1	90,8	98,1	60,1	39,1	23,4	32,5	29,8	17,4
-2,6836	35	37	92,2	81,9	92,8	54,7	36,2	22,1	30,4	27,4	17,7
-3,1014	35	34	90,8	75	89,2	50	34,8	22	24,8	25,9	16,9
-1,2838	35	38,4	100,5	90,3	98,7	57,8	37,3	22,4	31	28,7	17,7
-1,4855	36	38,7	98,2	90,3	99,9	59,2	37,7	21,5	32,4	28,4	17,8
-1,0527	36	41,5	115,3	108,8	114,4	69,2	42,4	24	35,4	21	20,1
-1,9569	37	36	96,8	79,4	89,2	50,3	34,8	22,2	31	26,9	16,9
-1,7167	37	35,3	92,6	83,2	96,4	60	38,1	22	31,5	26,6	16,7
-0,854	37	42,1	119,2	110,3	113,9	69,8	42,6	24,8	34,4	29,5	18,4
-1,1859	38	38	102,7	92,7	101,9	64,7	39,5	24,7	34,8	30,3	18,1
-1,4855	39	42,8	109,5	104,5	109,9	69,5	43,1	25,8	39,1	32,5	19,9
-0,9941	39	40	108,5	104,6	109,8	68,1	42,8	24,1	35,6	29	19
-3,6636	40	33,8	79,3	69,4	85	47,2	33,5	20,2	27,7	24,6	16,5

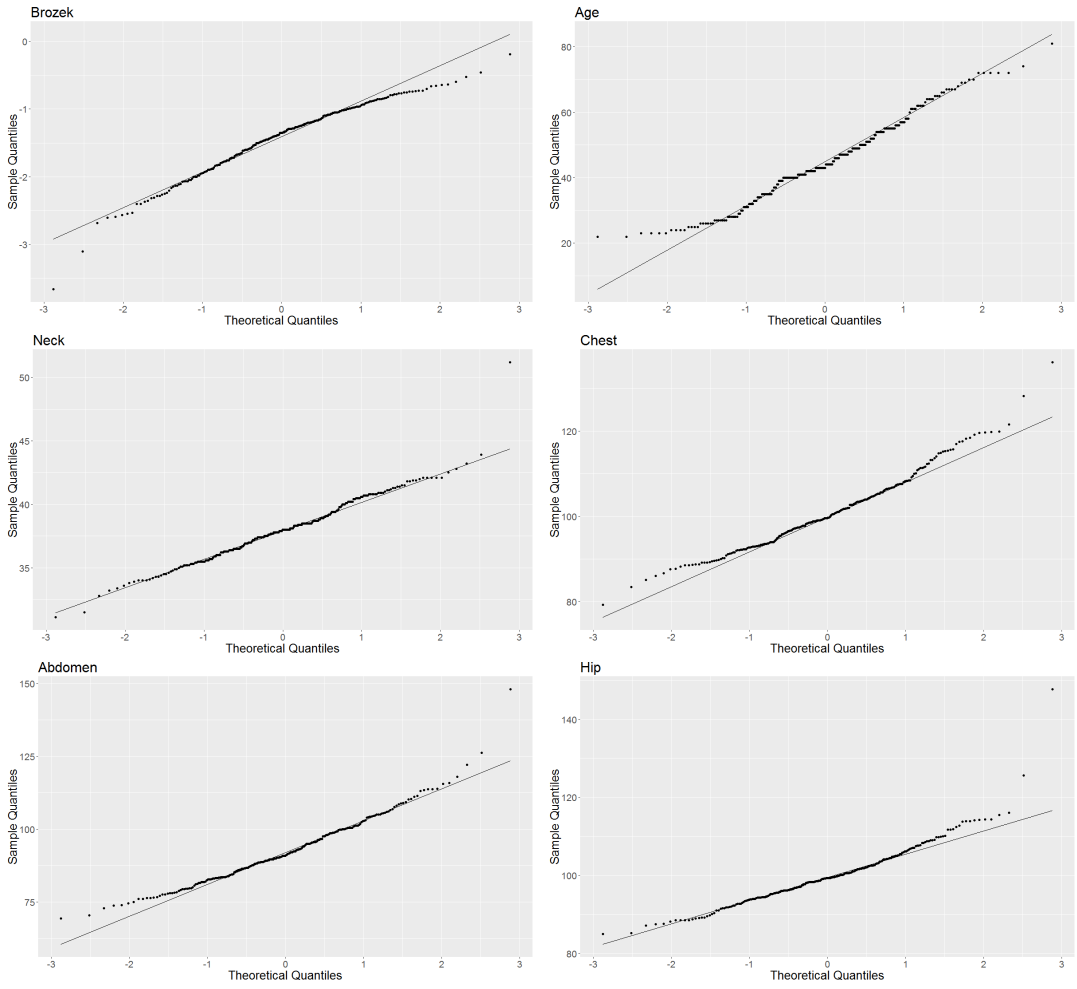
brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-1,8316	40	35,5	95,5	83,6	91,6	54,1	36,2	21,8	31,4	28,3	17,2
-1,7922	40	35,3	92,3	86,8	96,1	58	39,4	22,7	30	26,4	17,4
-1,4481	40	37,7	98,9	90,4	95,5	55,4	38,9	22,4	30,5	28,9	17,7
-2,0665	40	39,4	89,5	83,7	98,1	57,3	39,7	22,6	32,9	29,3	18,2
-1,1336	41	41,9	117,5	109,3	108,8	67,7	41,3	24,7	37,2	31,8	20
-1,2894	41	38,5	107,4	98,9	104,1	63,5	39,8	23,5	36,4	30,4	19,1
-1,2838	41	40,8	109,2	98	101,8	62,8	41,3	24,8	36,6	32,4	18,8
-1,0928	41	38	103,4	101,2	103,1	61,5	40,4	22,9	33,4	29,2	18,5
-1,8396	41	36,4	91,4	80,6	92,3	54,3	36,3	21,8	29,6	27,3	17,9
-0,5247	42	41,8	115,2	113,7	112,4	68,5	45	25,5	37,1	31,2	19,9
-1,5368	42	40,7	104,9	94,1	102,7	60,6	38,6	24,7	34	30,1	18,7
-1,0827	42	38,5	106,7	105,7	111,8	65,3	43,3	26	33,7	29,9	18,5
-1,17	42	35,4	92,2	85,6	96,5	60,2	38,9	22,4	31,7	27,1	17,1
-1,0429	42	38,5	101,6	96,6	100,6	61,1	38,4	24,1	32,9	29,8	18,8
-1,2126	42	35,5	97,8	86	96,2	57,7	38,6	24	31,2	27,3	17,4
-1,4358	42	36,5	92	89,7	101	62,3	38	22,3	30,8	27,8	16,9
-2,2616	42	37,6	94	78	99	57,5	40	22,5	30,6	30	18,5
-1,1543	43	37,4	103,7	89,7	94,2	58,5	39	24,1	33,8	28,8	18,8
-1,7845	43	37,8	102,7	89,2	99,2	60,2	39,2	23,8	31,7	28,4	18,6
-1,2018	43	35,2	91,1	85,7	96,9	55,5	35,7	22	29,4	26,6	17,4
-0,9044	43	37,9	107,2	103,1	105,5	68,8	38,3	23,7	32,1	28,9	18,7
-2,3185	44	37,9	100,8	89,1	102,6	60,6	39	24	32,9	29,2	18,4
-0,6531	44	40,9	121,6	113,9	107,1	63,5	40,3	21,8	34,8	30,7	17,4
-1,4982	44	41,9	105,6	96,3	102	63,3	39,8	24,1	37,3	23,1	19,4
-0,7257	44	39,1	100,6	93,9	100,1	58,9	37,6	21,4	33,1	29,5	17,3
-0,7343	47	40,2	102,7	101,3	101,7	60,7	39,4	23,3	36,7	31,6	18,4
-1,9835	47	36	99,8	83,9	91,8	53	36,2	22,5	31,4	27,5	17,7
-1,8803	47	34,5	92,9	84,4	94	56	38,2	22,6	29	26,2	17,6
-2,2073	49	35,8	91,2	79,4	89	51,1	35	21,7	30,9	28,8	17,4
-0,9654	49	40,2	115,6	104	109	63,7	40,3	23,2	36,8	31	18,9
-1,3349	49	38,3	98,3	89,7	99,1	56,3	38,8	23	29,5	27,9	18,6
-1,3816	50	39	103,7	97,6	104,2	60	40,9	25,5	32,7	30	19

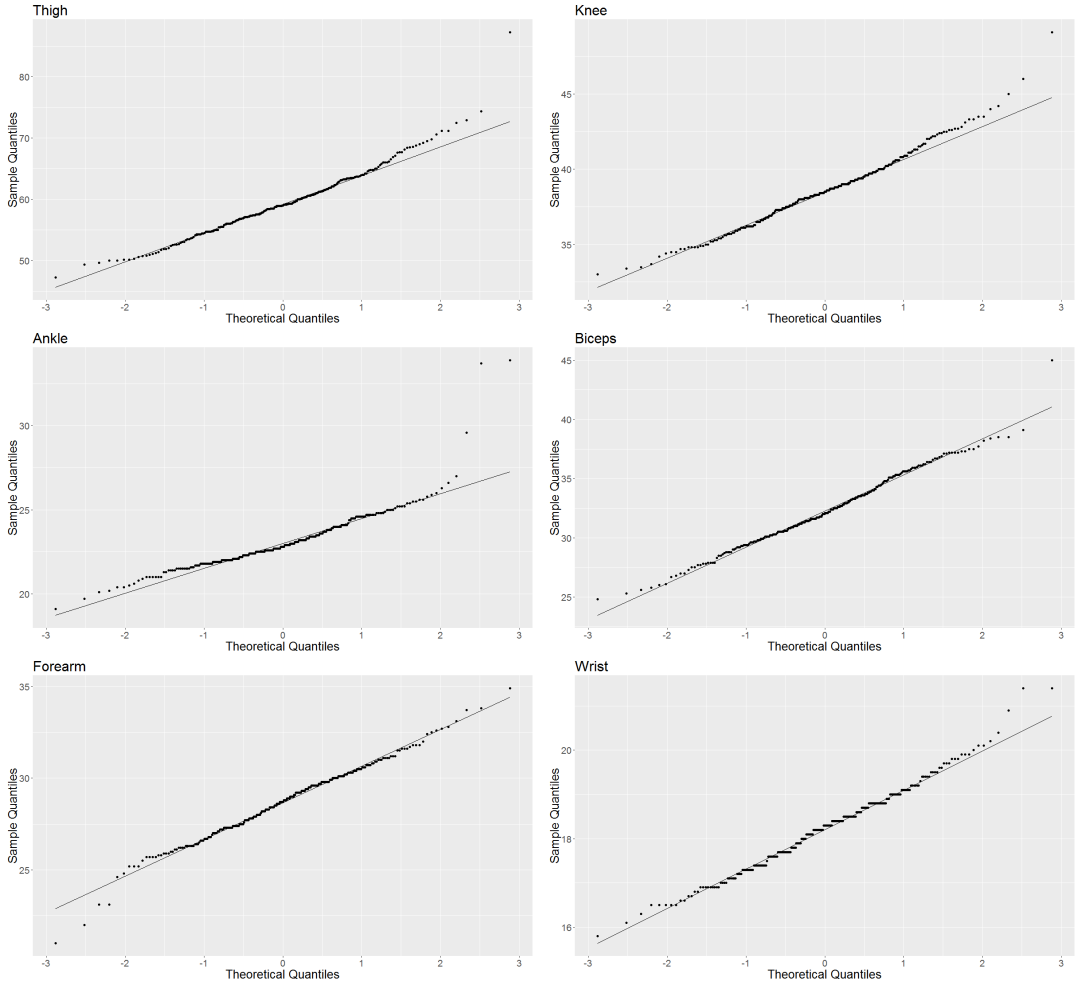
brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-1,3349	50	37,4	98,7	87,6	96,1	57,1	38,1	21,8	28,6	26,7	18
-0,1867	51	41,2	119,8	122,1	112,8	62,5	36,9	23,6	34,7	29,1	18,4
-1,6875	51	34,8	92,8	81,1	96,3	53,8	36,5	21,5	31,3	26,3	17,8
-2,1654	51	36,9	93,3	81,5	94,4	54,7	39	22,6	27,5	25,9	18,6
-1,0877	52	39,4	106,8	100	105	63,9	39,2	22,9	35,7	30,4	19,2
-1,5967	53	37,6	93,9	88,7	94,5	53,7	36,2	22	28,5	25,7	17,1
-1,7691	54	38,5	99	91,8	96,2	57,7	38,1	23,9	31,4	29,9	18,9
-1,0183	54	42,5	119,9	110,4	105,5	64,2	42,7	27	38,4	32	19,6
-1,8316	54	37,4	94,2	87,6	95,6	59,7	40,2	23,4	27,9	27	17,8
-2,4029	55	35,2	92,7	82,8	91,9	54,4	35,2	22,5	29,4	26,8	17
-1,8803	55	41,1	106,9	95,3	98,2	57,4	37,1	21,8	34,1	31,1	19,2
-1,7615	55	33,4	88,8	78,2	87,5	50,8	33	19,7	25,3	22	15,8
-1,6103	55	37,2	101,7	91,1	97,1	56,6	38,5	22,6	33,4	29,3	18,8
-1,0577	55	38,3	105,3	96,7	106,6	64	42,6	23,4	33,2	30	18,4
-1,6034	56	38,1	104	89,4	98,4	58,4	37,4	22,5	34,6	30,1	18,8
-1,4792	56	37,4	98,6	93	97	55,4	38,8	23,2	32,4	29,7	19
-1,897	57	35,2	99,6	86,4	90,1	53	35	21,3	31,7	27,3	16,9
-1,5303	57	39,4	103,4	96,7	100,7	59,3	38,6	22,8	31,8	29,1	19
-1,5698	58	38	100,2	88,1	97,8	57,1	38,9	23,6	30,9	29,6	18
-0,9893	58	35,1	94,9	94,9	100,2	56,8	35,9	21	27,8	26,1	17,6
-1,0086	60	40,4	97,2	93,3	94	54,3	35,7	21	31,3	28,7	18,3
-1,3875	62	38,3	104,7	95,6	93,7	54,4	37,1	22,7	30,3	26,3	18,3
-1,0827	62	40,6	104	98,2	101,1	59,3	40,3	23	32,6	28,5	19
-0,9654	63	40,2	117,6	113,8	111,8	63,4	41,1	22,3	35,1	29,6	18,5
-1,7691	64	37,9	95,8	82,8	94,5	61,2	39,1	22,3	29,8	28,9	18,3
-0,854	65	40,8	106,4	100,5	100,5	59,2	38,1	24	35,9	30,5	19,1
-1,4729	65	34,7	93	79,7	87,6	50,7	33,4	20,1	28,5	24,8	16,5
-0,6446	65	38,8	119,6	118	114,3	61,3	42,1	23,4	34,9	30,1	19,4
-0,8314	66	41,4	119,7	109	109,1	63,7	42,4	24,6	35,6	30,7	19,5
-0,7386	67	41,3	115,8	113,4	109,8	65,6	46	25,4	35,3	29,8	19,5
-0,8859	67	40,7	118,3	106,1	101,6	58,2	38,8	24,1	32,1	29,3	18,5
-1,5831	68	36,3	97,4	84,3	94,4	54,3	37,5	22,6	29,2	27,3	18,5

brozek	age	N	CH	ABD	H	T	K	A	B	F	W
-0,8404	69	40,8	113,7	107,6	110	63,3	44	22,6	37,5	32,6	18,8
-1,8639	70	34,9	89,2	83,6	88,8	49,6	34,8	21,5	25,6	25,7	18,5
-0,6998	72	40,9	108,5	105	104,5	59,6	40,8	23,2	35,2	28,6	20,1
-0,8768	72	38,9	111,1	111,5	101,7	60,3	37,3	21,5	31,3	27,2	18
-1,0183	72	38,9	108,3	101,3	97,8	56	41,6	22,7	30,5	29,4	19,8
-0,7692	74	40,8	112,4	108,5	107,1	59,3	42,2	24,6	33,7	30	20,9

Příloha B

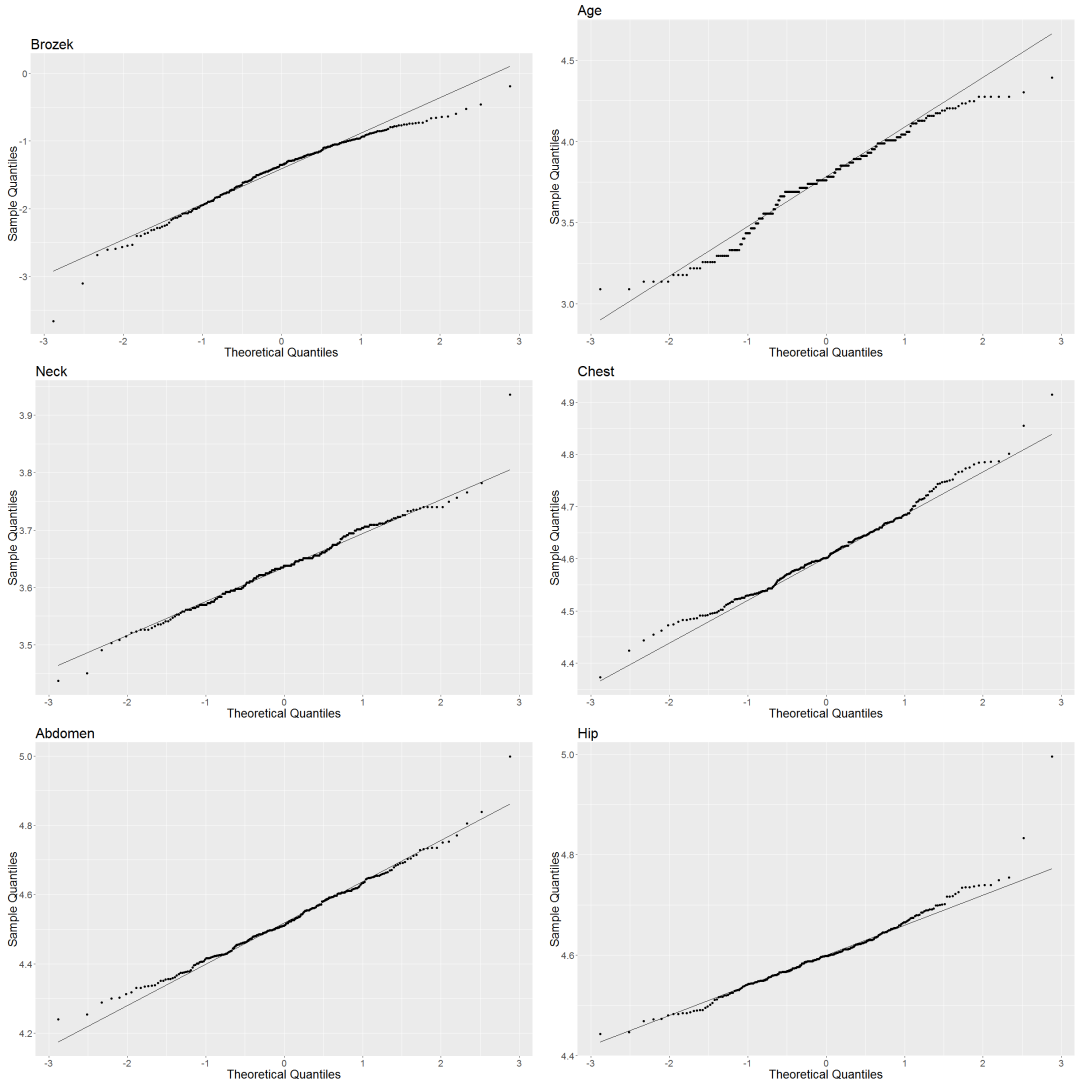
Q-Q grafy pro jednotlivé proměnné bez log-transformace (proměnná *brozek* po logitové transformaci).

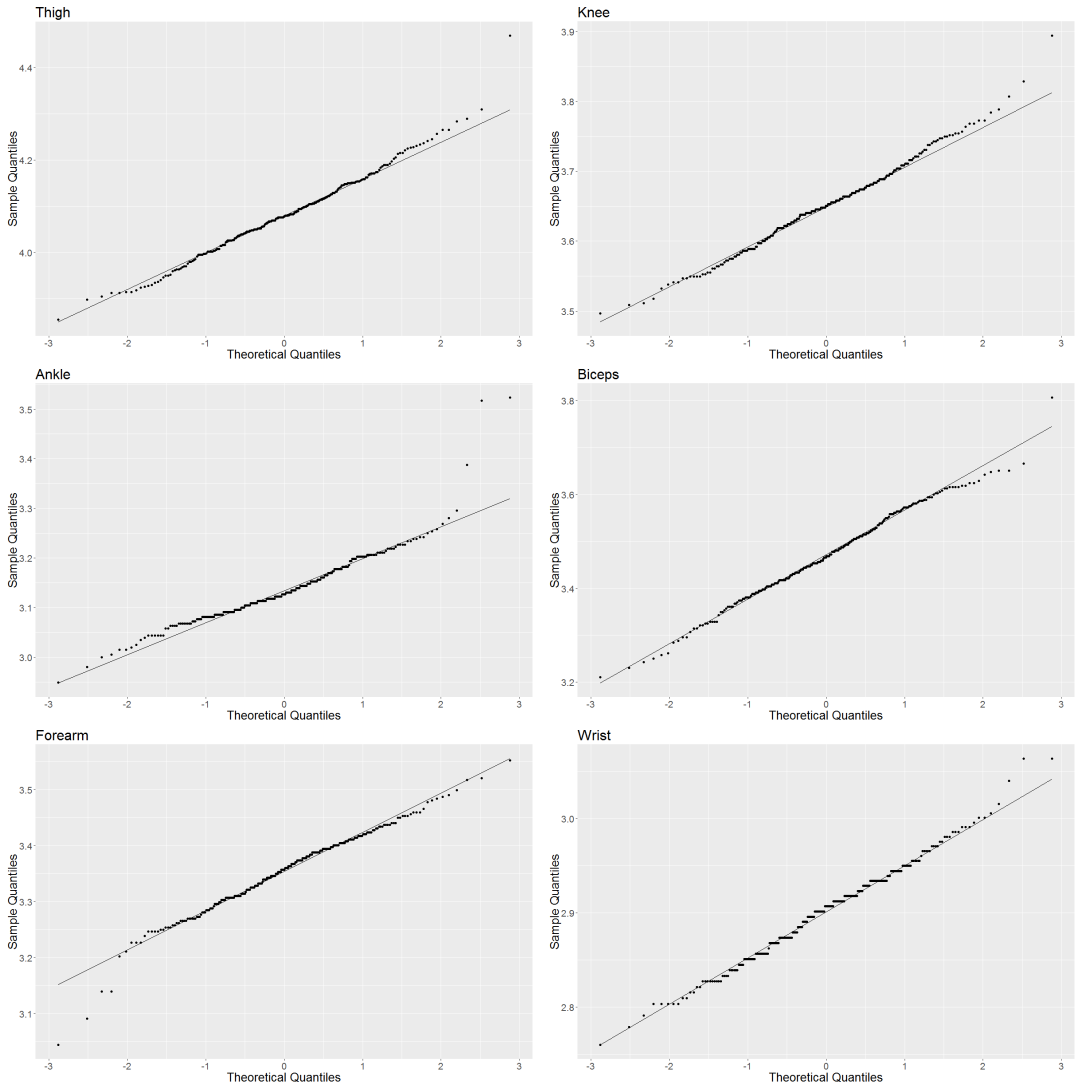



Thigh
Knee
Ankle
Biceps
Forearm
Wrist
Sample Quantiles
Theoretical Quantiles

Příloha C

Q-Q grafy pro jednotlivé proměnné po log-transformaci (proměnná *brozek* po logitové transformaci).




Thigh
Knee
Ankle
Biceps
Forearm
Wrist
Sample Quantiles
Theoretical Quantiles

Literatura

[1] Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H.: *Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination*, TEST, 2015, **24**(3), 441–461. ISSN: 1133-0686.

[2] Alqallaf, F., van Aelst, S., Yohai, V.J., Zamar, R.H.: *Propagation of outliers in multivariate data*, The Annals of Statistics, 2009, **37**(1), 311-331. ISSN: 0090-5364.

[3] Benjamini, Y., Hochberg, Y.: *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B: Methodological, 1995, **57**(1), 289-300. ISSN: 0035-9246 .

[4] Fawcett, T.: *An introduction to ROC analysis*, Pattern Recognition Letters, 2006, **27**(8), 861-874. ISSN: 0167-8655.

[5] Filzmoser, P.: *Robust statistics: Theoretical and practical considerations*, prezentace, TU Wien, 2013.

[6] Hron, K., Kunderová, P.: *Základy počtu pravděpodobnosti a matematické statistiky*, Univerzita Palackého v Olomouci, Olomouc, 2015.

[7] Huber, P.J.: *Robust Statistics*, John Wiley & Sons, Ltd, 1981.

[8] Koller, M., Stahel, W. A.: *Sharpening Wald-type inference in robust regression for small samples*, Computational Statistics and Data Analysis, 2011, **55**(8), 2504-2515. ISSN: 0167-9473.

[9] Leung, A., Zhang, H., Zamar, R.: *Package 'robreg3S'* [online]. 2015, [cit. 2018-04-08]. Dostupné z: https://cran.r-project.org/web/packages/robreg3S/robreg3S.pdf.

[10] Leung, A., Zhang, H., Zamar, R.: *Robust regeression estimation and inference in the presence of cellwise and casewise contamination*, Computational Statistics and Data Analysis, 2016, **99**, 1-11. ISSN: 0167-9473.

[11] Maronna, R.A., Martin, R.D., Yohai, V.J.: *Robust Statistics: Theory and Methods*, John Wiley & Sons, Ltd, 2006.

[12] Öllerer, V.: *Function 'The shooting S-estimator'* [online]. 2015, [cit. 2018-03-20]. Dostupné z: https://feb.kuleuven.be/viktoria.oellerer/software.

[13] Öllerer, V., Alfons, A., Croux, C.: *The shooting S-estimator for robust regression*, Computational Statistics, 2016, **31**(3), 829–844. ISSN: 0943-4062.

[14] Raymaekers, J., Rousseeuw, P.J., Van den Bossche, W., Hubert, M.: *Package 'cellWise'* [online]. 2018, [cit. 2018-03-05]. Dostupné z: https://cran.r-project.org/web/packages/cellWise/cellWise.pdf.

[15] Rousseeuw, P.J., Van Den Bossche, W.: *Detecting Deviating Data Cells*, Technometrics, 2018, **60**(2), 135-145. ISSN: 0040-1706.