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

BAKALÁRSKA PRÁCA

Robustná regresná analýza

Vedúci bakalárskej práce:
RNDr. Karel Hron Ph.D.
Rok odovzdania: 2012

Vypracovala:
Zuzana Bednáriková
ME, III. rocník

Prehlásenie

Prehlasujem, že som bakalársku prácu spracovala samostatne pod vedením pána RNDr. Karla Hrona, Ph.D. s použitím uvedenej literatúry.

V Olomouci dňa 7.12.2012

Poďakovanie

Na tomto mieste by som chcela poďakovať predovšetkým svojmu vedúcemu bakalárskej práce pánovi RNDr. Karolovi Hronovi, Ph.D., že mal so mnou dostatok trpezlivosti, aby mi pomohol doviesť túto prácu ku zdarnému koncu. Tiež by som rada poďakovala svojej rodine a priateľom, že ma po celú dobu štúdia podporovali.

Obsah

Úvod

Predmetom mojej bakalárskej práce je robustná regresná analýza. Regresná analýza je v dnešnej dobe populárnym nástrojom na skúmanie závislosti dvoch a viacerých premenných. Čitateľ, ktorý absolvoval aspoň základný kurz matematickej štatistiky by sa mal vedieť v tejto problematike orientovať. Hlavným cieľom bakalárskej práce je zoznámiť čitateľa s regresnou analýzou ako takou a predstaviť mu robustné metódy v teoretickej podobe ale aj aplikované na konkrétnych príkladoch.

Keďže v dnešnej dobe ide veda a technika rýchlo dopredu, nepočítajú sa štatistické výpočty manuálne na papieri a za pomoci kalkulačky, pretože by to bolo nie len časovo veľmi náročné. Aby sme sa nemuseli prácne trápiť s rôznymi výpočtami, bolo vyvinutých množstvo matematických a štatistických softwarov. Prvá kapitola je teda venovaná matematicko - štatistickému softwaru **R**. Jej cieľom je stručne sa zoznámiť s daným softwarom a so základnými funkciami, pretože **R** ponúka množstvo balíkov, knižníc a funkcii, čo by bola téma pre samostatnú prácu.

V druhej kapitole sa venujem lineárnemu regresnému modelu pre jednu a pre viacero vysvetľujúcich premenných. Veľmi dôležitou časťou tejto kapitoly, je metóda najmenších štvorcov, ktorá sa používa pre odhady regresných parametrov.

Tretia kapitola je venovaná jednotlivým robustným metódam, ktoré sa používajú v prípade, že klasická metóda najmenších štvorcov nie je efektívna. To sa deje z dôvodu výskytu odľahlých hodnôt, ktoré klasické odhady skresľujú a znehodnocujú.

V poslednej kapitole aplikujem niektoré metódy a postupy na konkrétnych príkladoch. Výpočty a vizuálne spracovanie bolo vytvorené pomocou softwaru **R** a za použitia knižnice *robustbase*.

Práca je napísaná v typografickom systéme LaTeX, pretože je vhodnejší na písanie matematických vzorcov a dokumentov oproti iným aplikáciám.

1. Štatistický software R

V prvej kapitole sa budeme venovať štatistickému softwaru **R**, pretože s ním budem často pracovať pri tvorbe ostatných kapitol. Je veľmi užitočný pre matematické a štatistické výpočty a grafické zobrazenia. Pri tvorbe tejto kapitoly boli použité zdroje č. [12] a dve príručky [11], [13], ktoré jednoducho popisujú ako v **R** pracovať a sú voľne dostupné z internetovej stránky http://www.karlin.mff.cuni.cz/ kulich/vyuka/Rdoc/index.html. Výhodou tohto softwaru je, že je voľne šíriteľný (na hlavnej webovej stránke http://www.r-project.org/) a funguje pod operačnými systémami ako je Linux, Unix, Microsoft Windows, MacOS a pod.

Pri používaní tohto programu sa môžeme stretnúť s niekoľkými druhmi okien. Ako prvé sa hneď po spustení zobrazí okno s názvom *R Console*, ktoré plní dve funkcie: zadávajú sa doň príkazy (t.j. vstup) a zobrazuje textový výstup. Druhé okno sa zobrazí po zadaní grafického príkazu a ide o okno s názvom *R Graphics*, kde je zobrazený grafický výstup, ktorý sa dá uložiť kamkoľvek na disk. Ďalším oknom je okno nápovedy, ktoré sa zobrazí po zadaní príkazu **help()**. Za každým príkazom nasledujú guľaté zátvorky (), do ktorých sa píšu argumenty. Príkazy sú vkladané vždy na nový riadok. Ak chceme zapísať viac príkazov za seba, je nutné ich oddeliť bodkočiarkou.

Predtým ako začneme v **R** pracovať, mali by sme najskôr:

- vytvoriť *pracovný adresár*, odkiaľ budeme čerpať dáta pri výpočtoch, alebo kde budeme ukladať napr. grafy. Na vytvorenie tohto adresára sa používa príkaz **setwd()**, kde sa ako argument píše do úvodzoviek celá cesta k adresáru. Pre zistenie v akom adresári sa práve nachádzame použijeme príkaz **getwd()**.

- odstrániť všetky premenné (objekty), ktoré boli kedysi vytvorené, použijeme funkciu **rm(list = ls())**.

Veľmi dôležité je, že **R** rozlišuje veľké a malé písmená. Názvy nových objektov a premenných je možné vytvárať len z písmen veľkej a malej abecedy, z číslic

od 0 do 9 a zo symbolov . a _. Názvy nesmú začínať číslicami, musia začínať písmenom a nesmú obsahovať medzeru. Teraz si uvedieme najčastejšie používané príkazové funkcie pri práci s **R**.

- Nápovedu vyvolá funkcia **help()**, resp. hypertextovú nápovedu **help.start()**.
- Objekt vytvoríme priradením pomocou operátora $<-$, resp. $=$ (odporúča sa používať operátor $<-$).
- Pre vloženie komentára na príkazovom riadku použijeme symbol #.
- Pomocou príkazu **ls()** sa vypíšu mená už existujúcich objektov. Ak chceme okrem mena zobraziť aj iné informácie, použijeme **ls.str()**.
- Pri práci v **R** sa všetky vytvorené objekty ukladajú do pamäti a môžeme ich tak kedykoľvek použiť (vytvárajú pracovný priestor, tzv. *workspace*). Ak chceme tieto objekty použiť pri nasledujúcom spustení programu, musíme ich uložiť a vytvoriť súbor s koncovkou .R a to cez *File* $\rightarrow$ *Save workspace*.
- Pre vytvorenie vektora sa používa funkcia **c()**. Funkcia **seq()** vytvorí vektor v tvare aritmetickej postupnosti. Ak použijeme operátor **:**, dostaneme vektor v tvare aritmetickej postupnosti s krokom 1.
- Jednotkovú maticu vytvoríme pomocou príkazu **diag()**.
- Maticu vytvoríme použitím príkazu **matrix**, kde dáta matice sú vyplňované po stĺpcoch. Pre vyplňovanie matice po riadkoch musíme nastaviť jej parameter **byrow** na hodnotu **TRUE**.
- Inverznú maticu získame použitím príkazu **solve()**.

Konkrétne funkcie týkajúce sa regresnej analýzy vždy spomeniem v nasledujúcich kapitolách u príslušnej metódy.

2. Regresná analýza

Regresná analýza je často používaný nástroj v mnohých oboroch (napr. v medicíne, ekonómii a pod.), ktorý sa využíva pri skúmaní závislosti dvoch alebo viacerých kvantitatívnych (číselných) premenných. Pri tvorbe tejto kapitoly som čerpala zo zdrojov [5], [6], [10] a [15]. Ide o model, v ktorom vystupujú dva druhy premenných: *závislá* (vysvetľovaná) premenná (označ. Y) a *nezávislá* (vysvetľujúca) premenná (označ. x). *Závislá premenná (vysvetľovaná)* vystupuje ako výsledok pôsobenia vysvetľujúcej premennej a pokladáme ju za náhodnú veličinu Y, ktorá má pri danej hodnote vysvetľujúcej veličiny x určité rozdelenie pravdepodobnosti. Jej chovanie chceme vysvetliť a popísať regresnou krivkou (priamkou). *Nezávislá premenná (vysvetľujúca)* svojím chovaním vysvetľuje chovanie závislej premennej, je to tzv. príčinná premenná, pretože v dôsledku jej zmeny sa mení aj závislá premenná.

Pre *regresný model* platí predpoklad, že nezávislá premenná x je *nenáhodná* (dopredu zvolená) a závislá premenná Y je *náhodná* (meraná). Regresný model delíme na *lineárny* a *nelineárny*. *Lineárny* je taký, ktorého regresné funkcie sú lineárne z hľadiska parametrov. Naopak *nelineárny* regresný model je taký, ktorý má regresné funkcie nelineárne z hľadiska parametrov. V tejto práci sa budeme venovať lineárnemu regresnému modelu.

Podkladom pre regresnú analýzu sú dáta, ktoré získame meraním alebo pozorovaním (za predpokladu náhodného výberu) u n jednotiek a považujeme ich za *výberové dáta*. Regresná analýza je teda analýza závislosti závislej premennej na nezávislej premennej.

Regresiu delíme na *jednoduchú* a *viacnásobnú* (mnohonásobnú). O jednoduchej regresii hovoríme vtedy, ak k odhadu závislej premennej Y použijeme len jednu vysvetľujúcu premennú x. To znamená, že v jednom z najjednoduchších prípadov pre pozorované hodnoty $(x_1, Y_1), \ldots, (x_n, Y_n)$ veličín x a Y platí vzťah

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

kde $i = 1, \ldots, n$. Pritom ε_i je náhodná chyba, veličina ktorej rozdelenie pravde-

podobnosti splňuje nasledujúce podmienky: stredná hodnota náhodnej chyby je nulová $\mathsf{E}(\varepsilon_i) = \mathbf{0}$, s konštantným, nezáporným rozptylom $\mathrm{var}(\varepsilon_i) = \sigma^2$ a s nulovou kovarianciou $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \mathbf{0}$, $\forall i \neq j$, $i, j = 1, \ldots n$. Ďalej β_0, β_1 sú neznáme regresné parametre modelu. Vzťah (1) sa nazýva *regresný model* a jeho pravá strana sa nazýva *regresná funkcia*. V tomto konkrétnom prípade tiež často hovoríme o tzv. *regresnej priamke*. Na základe pozorovaní $(x_1, Y_1), \ldots, (x_n, Y_n)$ budeme chcieť v tomto modeli odhadnúť neznáme parametre regresnej funkcie. Ak sú vyššie uvedené podmienky splnené, môžeme regresné koeficienty β_0, β_1 odhadovať pomocou *metódy najmenších štvorcov*, ktorej sa budem venovať neskôr v tejto práci.

V prípade, že použijeme vysvetľujúcich premenných viac, hovoríme o viacnásobnej regresii. Od jednoduchej regresie prechádzame k viacnásobnej preto, aby sa odhady hodnôt vysvetľovanej premennej zlepšili. Neodporúča sa voliť príliš mnoho vysvetľujúcich premenných, pretože sa zvyšuje riziko, že medzi vysvetľujúce premenné zahrnieme aj nepodstatné (nevhodné) faktory. To by analýzu skomplikovalo a ovplyvnilo výsledky, ktoré sa potom ťažšie interpretujú.

2.1. Viacnásobná lineárna regresia

K vytvoreniu tejto podkapitoly boli použité zdroje [4], [5], [6], [15]. Ako som už naznačila na konci predchádzajúceho odseku, viacnásobná lineárna regresia chce pomocou viacerých vysvetľujúcich premenných $x_1, \ldots, x_k$ predpovedať hodnotu závislej premennej Y. Teda hodnota premennej Y závisí na $x_j, j = 1, \ldots, k$ a na náhodnej chybe ε, danej voľbou regresného modelu. O náhodnej chybe hovoríme, že je náhodným členom modelu, pretože zahŕňa chyby meraní, alebo vplyv premenných, ktoré sme do daného modelu nezahrnuli.

Jednoduchý regresný model (1) sa dá analogicky rozšíriť na viacnásobný, ktorý má tvar

$$E(Y|(x_1, \ldots, x_k)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \tag{2}$$

kde $\beta_0, \beta_1, \ldots, \beta_k$ sú neznáme regresné parametre, ktoré určujú funkčný vzťah, na ktorom je model lineárny. Ich hodnoty odhadujeme z n po sebe nasledujúcich

pozorovaní. Teda n nezávislých pozorovaní premennej Y a zvolené vysvetľujúce premenné $x_j, j = 1, \ldots, k$ určujú celý viacnásobný regresný model, kde pre výsledok i-tého pozorovania, $i = 1, \ldots, n$ platí

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i,$$

pritom $x_{1j}, \ldots, x_{nj}$ predstavujú i-té hodnoty k vysvetľujúcich premenných a ε_i je náhodná chyba pri i-tom pozorovaní. Prevedieme tento vzťah do maticového zápisu:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

Prvý stĺpec matice $\mathbf{X}$ obsahuje jednotky, pretože je konštantným násobkom β_0 (β_0 sa nazýva *náhodný člen modelu*). $\mathbf{X}$ je matica typu $n \times (k+1)$.

Rozpísaný maticový zápis (3) vypadá takto,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

kde

$\mathbf{Y}$ je n-rozmerný stĺpcový vektor, ktorý obsahuje namerané hodnoty závislej premennej,

$\mathbf{X}$ je matica rozmeru $n \times (k+1)$, ktorá obsahuje pozorované hodnoty nezávislých premenných,

$\boldsymbol{\beta}$ je stĺpcový vektor rozmeru $k+1$, ktorý obsahuje skutočné neznáme hodnoty regresných parametrov,

$\boldsymbol{\varepsilon}$ je n-rozmerný stĺpcový vektor hodnôt náhodnej zložky.

Už vieme, že $\beta_0, \beta_1 \ldots, \beta_k$ sú neznámymi parametrami regresného modelu. Ich hodnoty sa odhadujú niektorou odhadovou metódou. Budem sa venovať *metóde najmenších štvorcov (MNŠ)* (označovanej aj ako OLS - Ordinary Least Squares). Táto metóda vychádza z niekoľkých predpokladov, ale tomu už bude venovaná nasledujúca podkapitola, v ktorej budem čerpať najmä zo zdrojov [4], [5], [6], [10], [14], [17].

2.2. Metóda najmenších štvorcov

Metóda najmenších štvorcov je najčastejšie používaná numerická vyrovnávacia metóda. Už vieme, že v regresnej analýze vystupujú dva druhy premenných. Vzťah medzi premennými x a Y (resp. medzi $x_1, \ldots, x_k$ a Y) chceme vyjadriť pomocou jednoduchej, ľahko interpretovateľnej funkcie, ktorá bude dobre charakterizovať pozorované hodnoty vysvetľovanej premennej a vysvetľujúcich premenných. K tomu používame také metódy, ktoré sú výpočtovo nenáročné a vykazujú dobré aproximačné vlastnosti.

Nech máme súbor nameraných hodnôt závislej premennej Y. Vieme, že výsledky meraní sú zaťažené chybou merania, musíme preto hľadať takú funkciu, ktorá zachová závislosť medzi veličinami x a Y a zároveň takú, aby chyba aproximácie bola čo najmenšia. Na to nám slúži metóda najmenších štvorcov (ďalej už len MNŠ), ktorá musí vo svojej najjednoduchšej podobe spĺňovať určité predpoklady [17]:

1. Stredná hodnota náhodnej chyby je nulová, takže platí $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ (alebo $\mathsf{E}(\varepsilon_i) = 0$, $i = 1, \ldots, n$), kde $\mathbf{0}$ je n-rozmerný nulový vektor. Podmienka nám hovorí, že náhodná chyba nepôsobí systematicky na hodnoty závislej premennej Y.

2. Rozptyl náhodnej chyby je konštantný , t.j. $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2$.

3. Všetky zložky náhodnej chyby sú nekorelované, teda platí $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \mathbf{0}$, $\forall i \neq j$, $i, j = 1, \ldots, n$.

4. Matica $\mathbf{X}$, ktorej prvky sú nenáhodné hodnoty vysvetľujúcich premenných, má hodnosť rovnú $m = k + 1$ a platí, že $h(\mathbf{X}) = m < n$. Z toho vyplýva, že žiadne dva stĺpce matice $\mathbf{X}$ nie sú kolineárne.

Pri dodržaní daných predpokladov sú odhady parametrov regresnej funkcie nevychýlené, no naďalej zostávajú citlivé na porušenie často užívaného predpokladu normality náhodných chyb, ku ktorému pri práci s reálnymi dátami dochádza.

MNŠ spočíva v hľadaní parametrov regresnej funkcie, pre ktoré je súčet štvorcov odchýlok vyrovnaných hodnôt vysvetľovanej premennej od hodnôt nameraných minimálny, tzn. že je založená na minimalizácii reziduálneho súčtu štvorcov, ktorý sa dá vyjadriť ako

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}, \quad (4)$$

čo sa dá vyjadriť ako $S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2(\boldsymbol{\beta})$.

Po roznásobení rovnice (4) dostaneme

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{\beta}\mathbf{X} = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta},$$

kde sme využili toho, že sčítanie je skalárne a že platí $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} = (\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$.

Parciálnou deriváciou podľa parametra $\boldsymbol{\beta}$ z danej rovnice získame $\frac{\partial(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon})}{\partial\boldsymbol{\beta}} = = 0 - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$. Odhad metódou najmenších štvorcov pre $\boldsymbol{\beta}$ dostaneme vyjadrením neznámej z vyššie uvedenej zderivovanej rovnice, teda

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (5)$$

Odhadované (vyrovnané) hodnoty $\hat{\mathbf{y}}$ vysvetľovanej premennej $\mathbf{y}$ vypočítame pomocou vektora $\hat{\boldsymbol{\beta}}$,

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}, \quad (6)$$

kde $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ označujeme ako *hat matrix* (pozri nasledujúcu podkapitolu). Pozorované hodnoty závislej premennej môžeme vyjadriť nasledovne

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

kde $\hat{\boldsymbol{\varepsilon}}$ je vektor, ktorého prvky sú hodnoty reziduálnych odchýlok, vypočítané zo vzťahu

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}}. \quad (7)$$

2.3. Posúdenie kvality regresného modelu (diagnostika)

Pri tvorbe tejto kapitoly som použila zdroje [4] a [6].
Diagnostika tvorí dôležitú časť *regresnej analýzy*. Jej cieľom je overiť, či sme zvolili vhodný regresný model (posudzujeme model ako celok), resp. či sme použili vhodnú metódu pre odhad regresných parametrov (testujeme významnosť parametrov modelu). Chceme tak posúdiť intenzitu závislosti (tesnosti) medzi nezávislou a závislou premennou. Vieme už, že v regresnej analýze ide o to, aby odhadnutá regresná funkcia čo najlepšie kopírovala hodnoty pozorovaných údajov. Teda čím budú realizácie $(x_1, y_1), \ldots, (x_n, y_n)$ pozorovaní bližšie koncentrované okolo odhadnutej regresnej funkcie, tým bude závislosť medzi závislou a nezávislou premennou silnejšia (teda aj regresná funkcia bude lepšia). Naopak, čím viac budú realizácie pozorovaní viac vzdialené od odhadnutej regresnej funkcie, tým bude táto závislosť slabšia (to znamená, že aj regresná funkcia bude menej kvalitná). V tejto práci sa zameriame na diagnostiku odľahlých hodnôt, ktoré ovplyvňujú kvalitu regresného vzťahu.

Najskôr sa budem venovať jednotlivým druhom odľahlých hodnôt, s ktorými sa pri práci s dátami môžeme stretnúť. Definície sú použité zo zdroja [3].

1. *Extrémne odľahlé pozorovania (outliers)* - sú to odľahlé hodnoty pozorovaní vyskytujúce sa u závislej premennej Y, odľahlé hodnoty u vysvetľujúcej premennej sa nazývajú *vybočujúce pozorovania (leverage points)*. Ďalej sa vybočujúce pozorovania delia na:

   - *Dobré vybočujúce pozorovania (good leverage points)* - sú také vybočujúce pozorovania, ktoré nie sú súčasne extrémnymi odľahlými pozorovaniami. Sú relatívne vzdialené od väčšiny pozorovaní, ale ležia blízko regresnej funkcie okolo ktorej je sústredená prevažná časť bodov. Dobré vybočujúce pozorovania majú obmedzený vplyv na kvalitu regresných odhadov.
   - *Zlé vybočujúce pozorovania (bad leverage points)* - sú také vybočujúce pozorovania, ktoré majú hodnoty vysvetľujúcej premennej vzdialené

od väčšiny ostatných pozorovaní (v prípade jednoduchej závislosti ležia ďaleko od regresnej priamky, ktorá charakterizuje priebeh závislosti väčšiny bodov). Ich existencia výrazne znižuje presnosť regresných odhadov a spôsobuje mylnú informáciu o tvare závislosti.

2. *Vplyvné pozorovania (influential points)* - pozorovania, ktorých zaradenie alebo vyradenie z regresného vzťahu spôsobí výrazné zmeny vo vypočítanom modeli (napr. zmena regresných koeficientov, vyrovnaných hodnôt).

3. K posudzovaniu robustnosti odhadov v prípade existencie odľahlých a vybočujúcich pozorovaní bola navrhnutá celá rada meradiel, napr. koncept tzv. *bodov zlyhania (breakdown points = body zvratu)*. Vyjadrujú percentuálne zastúpenie chybných hodnôt, pre ktoré je ešte odhad „správny“ (teda neovplyvnený podstatným výskytom odľahlých hodnôt). Pohybuje sa v rozmedzí $0\% - 50\%$.

Ako diagnostický nástroj v regresnej analýze pre detekciu odľahlých hodnôt slúži hat matrix, označovaná aj ako *projekčná matica*. Zo vzťahu (6) vieme, že $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ a $\hat{\mathbf{y}}$=$\mathbf{H}\mathbf{y}$. To znamená, že $\hat{\mathbf{y}}$ je projekciou vektora $\mathbf{y}$ do priestoru nad stĺpcami matice $\mathbf{X}$. Výskyt vybočujúcich pozorovaní (*Leverage points*) má výrazný vplyv na odhady regresných parametrov. Diagonálne prvky projekčnej matice $\mathbf{H}$ slúžia ako nástroj na vyhľadávanie týchto vybočujúcich pozorovaní. Platí nerovnosť, že $0 \leq h_{ii} \leq 1$, $i = 1, \ldots, n$, pritom stopa matice $\mathbf{H}$ má tvar $tr(\mathbf{H}) = tr[\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] = tr(\mathbf{I}_{k+1}) = k + 1$. Ak $h_{ii} = 0$, nejde o vybočujúce pozorovanie a naopak, ak $h_{ii} = 1$ ide o vybočujúce pozorovanie. Na identifikovanie odľahlých pozorovaní slúži heuristické pravidlo $h_{ii} > \frac{2\cdot(q+1)}{n}$ [4].

Ďalším diagnostickým nástrojom je *regresný diagnostický graf*. O ňom sa však bližšie zmienime až v nasledujúcej kapitole venovanej robustným metódam v regresných modeloch.

3. Robustné metódy

V prípade klasickej lineárnej regresie sme odhadovali parametre pomocou MNŠ. Dôležitú úlohu tu hrá často sa vyskytujúci dodatočný predpoklad normality reziduí, ale aj mnohé iné podmienky (napr. nezávislosť, náhodnosť a pod.). Ak sú tieto podmienky splnené, odhady parametrov regresnej funkcie sú nevychýlené, to znamená, že ich stredné hodnoty sú rovné skutočným (teoretickým) hodnotám regresných parametrov. Navyše sú tieto odhady najlepšie, teda medzi všetkými ostatnými odhadmi majú najmenší rozptyl. Často však sú medzi skutočné dáta zahrnuté aj nezvyčajné (netypické), ktoré tam evidentne nepatria. Výskyt takýchto chýb môže byť spôsobený nesprávnym prepisom dát pri spracovaní (napr. zle opísaná desatinná čiarka), alebo časť dát je zo súboru s iným rozdelením. Môže sa ale stať, že v súbore dát máme pozorovania, ktoré sú správne, ale netypické pre daný model. Chybné pozorovania, ktoré sa nedajú opraviť, bývajú eliminované, no nie vždy je to vhodné riešenie. Takéto vybočujúce pozorovania nám môžu znehodnotiť kvalitu odhadnutých regresných parametrov získaných MNŠ. Problém s vybočujúcimi pozorovaniami prispel k navrhnutiu mnohých metód, ktoré sa snažia obmeniť (modifikovať) klasickú MNŠ tak, aby minimalizovala citlivosť na odľahlé pozorovania. Zároveň tieto metódy musia pri splnení podmienok klasickej metódy poskytovať dobré odhady regresných parametrov. Takéto metódy sa nazývajú robustné. Ich rozvoj je podporovaný rýchlym rozvojom výpočtovej techniky a sú zakomponované v rôznych štatistických softwaroch (napr. SAS, R).

3.1. Základné metódy robustnej regresie

Vieme už, že odľahlé a vybočujúce pozorovania spôsobujú klasickej MNŠ mnohé problémy, preto boli vyvinuté robustné metódy, ktoré citlivosť na tieto pozorovania eliminujú. Zároveň, v prípade zachovania podmienok MNŠ a v prípade nekontaminavaných dát, musia robustné metódy poskytnúť rovnaké odhady regresných funkcií, ako poskytuje klasická regresia. Toto je základný princíp robustných metód. Teda robustné metódy potlačajú vplyv odľahlých hodnôt na dátovom súbore. Teraz prejdem k jednotlivým robustným metódam a popíšem

možnosti ich použitia. Informácie čerpám zo zdrojov č. [3], [4], [7], [8], [9].

3.2. Robustná lineárna regresia

3.2.1. LS-regresia (Least Squares (LS))

V regresnej analýze máme n pozorovaní závislej premennej Y, ktoré odpovedajú hodnotám nezávislých premenných $x_1, \ldots, x_k$, viď regresný model (3). Rozdiel medzi hodnotami závislej premennej a vyrovnanými hodnotami nazývame reziduálny. Základným princípom metódy najmenších štvorcov je minimalizovanie súčtu štvorcov reziduí, čo máme vyjadrené vzťahom (4). Symbolom $\hat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$ označujeme vektor, ktorý tento vzťah minimalizuje a platí, že je odhadom regresných parametrov $\boldsymbol{\beta}$. Teda

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{8}$$

podobne vypočítame aj vyrovnané hodnoty $\hat{\mathbf{y}}$ z $\mathbf{y}$,

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}, \tag{9}$$

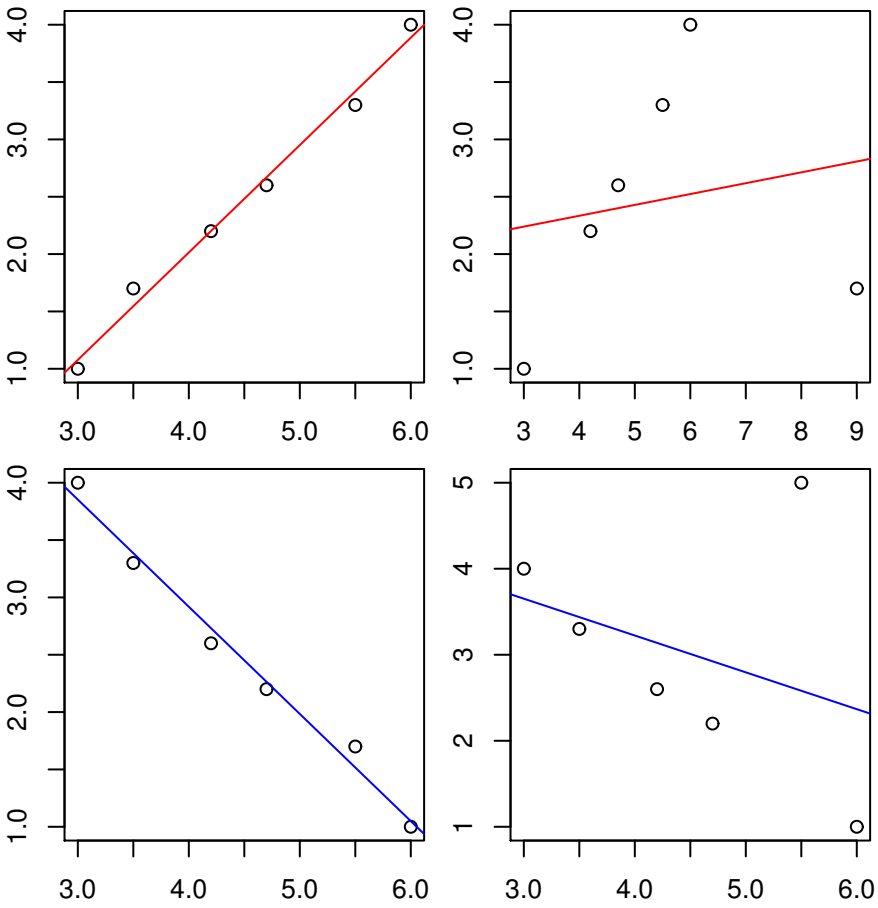
kde $\mathbf{H}$ je hat matrix (v slovenskej a českej literatúre označovaná ako tzv. klobúková matica alebo projekčná matica) a platí $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Pre odhadované reziduá platí

$$\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{LS}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \tag{10}$$

V LS-regresii sa k odhadom parametrov pomocou MNŠ používajú druhé mocniny reziduí. To vedie k tomu, že hodnoty ktoré sa neriadia lineárnym trendom, výrazne ovplyvnia odhad.

V nasledujúcom grafe máme zakreslené body, ktoré sa riadia lineárnym trendom. Preložíme ich LS-regresnou priamkou (grafy na ľavej strane). V prípade, že sa nejaký z týchto bodov vychýli v y-vom smere, regresné parametre tým budú silno ovplyvnené a tým sa posunie aj LS-regresná priamka (graf vpravo dole). Ak sa však vychýli nejaký bod v x-vom smere, má to na LS-regresiu ešte horší dopad. LS-regresia teda nie je najvhodnejšou metódou, pretože berie do úvahy štvorce reziduí, a preto aj jedna odľahlá hodnota môže znehodnotiť odhady parametrov.



Obrázek 1: Bodové grafy preložené LS-regresnou priamkou. Hore: zobrazenie odľahlej hodnoty v x-vom smere. Dole: zobrazenie odľahlej honoty v y-vom smere.

Tomu hovoríme, že má *bod zlyhania* rovný nule, pretože táto metóda nepripúšťa žiadne odľahlé hodnoty.

Graf sme vytvorili v softwari **R**, kde sme si ľubovoľne zvolili hodnoty vysvetľujúcej a vysvetľovanej premennej. V ľavom hornom grafe sú hodnoty závislej premennej zapísané ako vektor $\mathbf{y} = (1, 1.7, 2.2, 2.6, 3.3, 4)^T$ a vektor nezávislej premennej $\mathbf{x} = (3, 3.5, 4.2, 4.7, 5.5, 6)^T$. V pravo hore je odľahlá hodnota v x-vom smere, teda sme zmenili druhú hodnotu vo vektore $\mathbf{x}$ z hodnoty 3.5 na 9, teda $\mathbf{x}_1 = (3, 9, 4.2, 4.7, 5.5, 6)^T$. Vľavo dole máme graf, ktorého vektor nezávislej premennej má hodnoty $\mathbf{x} = (3, 3.5, 4.2, 4.7, 5.5, 6)^T$ a vektor závislej premennej $\mathbf{y} = (4, 3.3, 2.6, 2.2, 1.7, 1)^T$. Vektor $\mathbf{y}$ pre graf s odľahlou hodnotou v y-vom

smere sa zmenil v predposlednej hodnote z 1.7 na hodnotu 5. Vidíme, že zvlášť pri výskyte odľahlých hodnôt v x -vom smere dochádza k dramatickej zmene regresnej priamky.

3.2.2. L_1 odhady

Vyššie sme sa oboznámili s LS-regresiou a zdôvodnili, že je citlivá na veľké hodnoty reziduí, pretože z nich počítame druhé mocniny. Problému tohto druhu môžeme predísť tak, že štvorce reziduí nahradíme absolútnou hodnotou reziduí. To znamená, že L_1-regresia nám minimalizuje súčet absolútnych hodnôt reziduí, čo sa dá zapísať ako

$$\sum_{i=1}^{n} |\varepsilon_i(\boldsymbol{\beta})|. \tag{11}$$

Tak dostaneme odhady regresných koeficientov $\hat{\boldsymbol{\beta}}_{\boldsymbol{L_1}}$. Pre ich výpočet musíme použiť iteratívny algoritmus, pretože neexistuje explicitný vzorec ako v prípade MNŠ. *Bod zlyhania* je tu opäť rovný nule, pretože už jedná odľahlá hodnota môže mať vážny dopad na regresné odhady.

Podobne ako u LS-regresie, ak na grafe vybočuje nejaká hodnota v y-vom smere, L_1-regresiu to neovplyvní. Ak však vybočuje v x-vom smere, spôsobí to preklopenie regresnej priamky. Hovoríme, že L_1-regresia je robustná voči odľahlým pozorovaniam v y-vom smere, ale nie voči vybočujúcim pozorovaniam a vplyvným bodom.

3.3. Odhady s vysokým bodom zlyhania

LS-regresia aj L_1-regresia berú do úvahy všetky hodnoty pozorovaní vysvetľujúcich premenných a vysvetľovanej premennej, teda celý súbor o rozsahu n. Tak ako LS-regresia, ani L_1-regresia nám neposkytuje požadované výsledky, pretože je citlivá na vybočujúce pozorovania a vplyvné body. Robustné metódy chcú prispôsobiť model len na určitú časť dát a nie na celý súbor. Geometricky to znamená, že ide o nájdenie najtesnejšieho pásu, ktorý pokryje určitú časť pozorovaní.

3.3.1. LMS-regresia (Least Median of Squares(LMS))

Jednou z metód s vysokým bodom zlyhania je *LMS-regresia*, ktorá je daná vzťahom

$$median_i\, {\varepsilon_i}^2(\boldsymbol{\beta}) \tag{12}$$

to znamená, že minimalizujeme medián štvorcov reziduí. Ide o rovnaký vzťah ako u LS-regresie, viď (4), akurát sumu sme nahradili mediánom. K výpočtu regresných koeficientov $\hat{\boldsymbol{\beta}}_{\boldsymbol{LMS}}$, ako u L_1-regresie, neexistuje žiaden explicitný vzorec, preto je potrebné použiť nejaký iteratívny aproximačný algoritmus.

Bod zlyhania je 50%, teda až 50% dátových bodov môžeme presunúť bez toho, aby to výrazne ovplyvnilo regresné odhady. To sa síce môže zdať na jednu stranu veľmi výhodné, no v praxi väčšinou nepresahuje počet odľahlých hodnôt v dátovom súbore 20-30%. Preto hľadáme metódu, kde bude možné nastaviť veľkosť bodu zlyhania ako parameter.

3.3.2. LTS-regresia (Least Trimmed Squares(LTS))

LTS-regresia nazývaná ako metóda najmenších useknutých štvorcov je daná minimalizáciou výrazu

$$\sum_{i=1}^{h} (\varepsilon^2(\boldsymbol{\beta}))_i, \tag{13}$$

pričom platí, že $(\varepsilon^2(\boldsymbol{\beta}))_1 \leq (\varepsilon^2(\boldsymbol{\beta}))_2 \leq \ldots \leq (\varepsilon^2(\boldsymbol{\beta}))_h \leq \ldots \leq (\varepsilon^2(\boldsymbol{\beta}))_n$, je usporiadanie reziduí. To znamená, že minimalizujeme súčet h najmenších usporiadaných reziduí. Aj v prípade *LTS*-regresie nemáme pre odhady regresných parametrov $\hat{\boldsymbol{\beta}}_{\boldsymbol{LTS}}$ explicitný vzťah, ako u MNŠ, je potreba použiť aproximačné metódy [16]. *Bod zlyhania* v tomto prípade je 50% pre $h \approx \frac{n}{2}$ a pre väčšie h sa *bod zlyhania* znižuje na hodnotu $\frac{n-h}{n}$. Vďaka svojím vlastnostiam *LTS*-regresia predtsavuje v súčasnej dobe jednu z najpoužívanejších robustných regresných metód.

Existuje niekoľko algoritmov pre aproximáciu odhadov regresných koeficientov pomocou LTS regresie a jeden z nich je uvedený v ([8], str. 23).

3.3.3. RLS odhady (Reweighted Least Squares (RLS))

Nakoniec pre úplnosť spomenieme tiež *RLS odhady* - odhady metódou vážených štvorcov. RLS slúži na zvýšenie účinnosti odhadov klasickou MNŠ. Vážená MNŠ minimalizuje súčet reziduí, kedy najskôr odhady parametrov vypočítame nejakou robustnou metódou a štvorce reziduí vynásobíme váhami w_i, ktoré nadobúdajú hodnotu 0 alebo 1 (napr. na základe robustnej diagnostiky o ktorej sa zmienime ďalej). Hodnotu 0 priraďujeme odľahlým pozorovaniam a ostatným dáme váhu 1. Vážená MNŠ je vlastne klasická MNŠ, ktorú aplikujeme na takto zredukovaný súbor pozorovaní s cieľom použiť ďalej všetky vlastnosti odhadov, ktoré táto metóda poskytuje.

3.4. Robustné mnohorozmerné odhady a regresná diagnostika

V tejto kapitole sa budem venovať robustným odhadom *mnohorozmernej polohy a kovariancie*. Mnohorozmerná poloha a kovariancia sú veľmi dôležité pre mnohé štatistické metódy. V klasickom prípade je na základe náhodného výberu $\mathbf{x_1}, \ldots, \mathbf{x_n}$ z nejakého p-rozmerného rozdelenia takto odhadovaná stredná hodnota $\boldsymbol{\mu}$ s aritmetickým priemerom $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}$ a variančnú maticu $\boldsymbol{\Sigma}$ s výberovou kovariančnou maticou $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})^T$. Tieto odhady sú citlivé na odľahlé hodnoty, a preto pristupujeme k robustifikácii. Ďalej sa zoznámime s odhadovou metódou *MCD*, ktorá sa v praxi používa zrejme najčastejšie.

3.4.1. MCD (Minimum Covariance Determinant) odhady

V tomto prípade budeme používať namiesto klasických výberovych odhadov aritmetického priemeru $\bar{\mathbf{x}}$ a kovariančnej matice $\mathbf{S}$, robustné MCD-odhady pre strednú hodnotu $\boldsymbol{\mu}$ a variančnú maticu $\boldsymbol{\Sigma}$. Už z názvu tohto odhadu plynie, že budeme hľadať odhad, ktorý má minimalný determinant kovariančnej matice. Pri tvorbe podkapitoly boli použité zdroje [4], [7], [8] a [9].

Skôr ako sa začnem venovať MCD odhadu, je potrebné vedieť, že MCD odhad polohy, $\hat{\boldsymbol{\mu}}$, a kovariancie, $\hat{\boldsymbol{\Sigma}}$, je affinne ekvivariantný, t.j. pri affinnej transformácii

dát sa budú tieto odhady chovať analogicky ako vyššie spomenuté „klasické“ odhady. To znamená že platí

$$\hat{\boldsymbol{\mu}}(\mathbf{A}\mathbf{x_1}+\mathbf{b},\ldots,\mathbf{A}\mathbf{x_n}+\mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\mu}}(\mathbf{x_1},\ldots,\mathbf{x_n})+\mathbf{b} \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}(\mathbf{A}\mathbf{x_1}+\mathbf{b},\ldots,\mathbf{A}\mathbf{x_n}+\mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\Sigma}}(\mathbf{x_1},\ldots,\mathbf{x_n})\mathbf{A}^T, \tag{15}$$

kde $\mathbf{A}$ je *non-singulárna* matica rádu p a $\mathbf{b} \in \mathbf{R}^p$.

MCD metóda hľadá h pozorovaní z celkových n pozorovaní (rozsah výberu), pre ktoré platí, že ich výberová kovariančná matica má najmenší možný determinant. Platí, že MCD odhad polohy $\hat{\boldsymbol{\mu}}$ je centrum elipsoidy, ktorej determinant výberovej kovariančnej matice (obsahuje najmenej h pozorovaní) je minimálny a počíta sa ako aritmetický (výberový) priemer daných h pozorovaní. MCD odhad kovariancie, $\hat{\boldsymbol{\Sigma}}$, je prenásobený konštantou pre konzistenciu pri normálnom rozdelení. MCD odhady strednej hodnoty $\boldsymbol{\mu}$ a variančnej matice $\boldsymbol{\Sigma}$ poskytujú odhady s bodom zlyhania $\approx (n-h)/n$. Pre maximálny bod zlyhania 50% sa volí h=$[(n+p+1)/2]$.

Pretože presný výpočet MCD by bol numericky veľmi náročný, počítajú sa MCD odhady pomocou vhodných približných algoritmov. Existuje niekoľko takýchto postupov, my si uvedieme jeden z nich a to *FAST MCD algoritmus*. Hlavnou zložkou tohto algoritmu je tzv. *C-krok*.

Veta 1. *([12], str. 23) Vezmeme* $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ *a nech* $H_1 \subset \{1,\ldots,n\}$ *je taká* podmnožina, *že* $|H_1| = h$ *(teda počet jej prvkov je rovný* h*). Vezmeme* $\hat{\boldsymbol{\mu}}_{\mathbf{1}} := \frac{1}{h}\sum_{i\in H_1}\mathbf{x}_i$ *a* $\hat{\boldsymbol{\Sigma}}_{\mathbf{1}} := \frac{1}{h}\sum_{i\in H_1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{\mathbf{1}})(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{\mathbf{1}})^T$. *Ak* $det(\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}) \neq 0$, *určíme Mahalanobisové vzdialenosti* $d_1(i) := \sqrt{(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{\mathbf{1}})^T\hat{\boldsymbol{\Sigma}}_1^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{\mathbf{1}})}$, *pre* $i = 1,\ldots,n$. *Teraz vezmeme* H_2 *takú množinu indexov, že* $\{d_1(i); i \in H_2\} := \{(d_1)_{1:n},\ldots,(d_1)_{h:n}\}$, *kde* $(d_1)_{1:n} \leq (d_1)_{2:n} \leq,\ldots,(d_1)_{h:n}$ *sú usporiadané vzdialenosti a počítame* $\hat{\boldsymbol{\mu}}_2$ *a* $\hat{\boldsymbol{\Sigma}}_2$ *založené na* H_2. *Potom* $det(\hat{\boldsymbol{\Sigma}}_2) \leq det(\hat{\boldsymbol{\Sigma}}_1)$, *kde rovnosť platí vtedy a len vtedy, ak* $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ *a* $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{\Sigma}}_1$.

Ak vyjde $det(\hat{\boldsymbol{\Sigma}}_1) > 0$ (podmienka, že $det(\hat{\boldsymbol{\Sigma}}_1) \neq 0$ nie je skutočným obmedzením, pretože determinant nemôže nadobúdať zápornú hodnotu a tak minimálnu objektívnu hodnotu), potom C-krokom vyjde $det(\hat{\boldsymbol{\Sigma}}_2)$ také, pre ktoré platí

$det(\hat{\mathbf{\Sigma}}_2) \leq det(\hat{\mathbf{\Sigma}}_1)$. To znamená, že $det(\hat{\mathbf{\Sigma}}_2)$ má menší determinant. C-krok sa teda počíta následovne, vezmeme úvodné odhady $(\hat{\boldsymbol{\mu}}_{old}, \hat{\mathbf{\Sigma}}_{old})$

- vypočítame vzdialenosti $d_{old}(i)$, pre $i = 1, \ldots, n$;
- usporiadáme vzdialenosti do permutácie π pre každé $d_{old}(\pi(1)) \leq d_{old}(\pi(2)) \leq \ldots \leq d_{old}(\pi(n))$;
- vytvoríme $H_{new} := \{\pi(1), \pi(2), \ldots, \pi(n)\}$;
- vypočítame príslušné $\hat{\boldsymbol{\mu}}_{new}$ ako výberový priemer pozorovaní s indexmi z H_{new} a $\hat{\mathbf{\Sigma}}_{new}$ odpovedajúce výberovej kovariančnej matici.

C-krok môžeme opakovať a to až dovtedy, kým sa $det(\hat{\mathbf{\Sigma}}_{new}) = 0$ alebo $det(\hat{\mathbf{\Sigma}}_{new}) = det(\hat{\mathbf{\Sigma}}_{old})$. Pretože existuje len konečne veľa h-prvkových podmnožin, postupnosť takto získaných determinantov musí konvergovať ku konečnému číslu použitých krokov (čo nezaručuje, že $det(\hat{\mathbf{\Sigma}}_{new})$ je globálnym minimom podľa definície MCD odhadu).

V inom predvolenom nastavení FAST-MCD zasa počítame odhady pomocou *jednokrokových vážených odhadov*, ktorý má rovnakú hodnotu bodu zlyhania ako základný MCD odhad. Na ich výpočet sa používajú vzorce:

$$\hat{\boldsymbol{\mu}}_1 = \left(\sum_{i=1}^{n} w_i \mathbf{x}_i \right) / \left(\sum_{i=1}^{n} w_i \right), \tag{16}$$

$$\hat{\mathbf{\Sigma}}_1 = d_{h,n} \left(\sum_{i=1}^{n} w_i (\mathbf{x}_i - \hat{\mu}_1)(\mathbf{x}_i - \hat{\mu}_1)^T \right) \left(\sum_{i=1}^{n} w_i \right)^{-1}, \tag{17}$$

kde

- $w_i = 1$ pre $d_{(\hat{\boldsymbol{\mu}}_{MCD}, \mathbf{\Sigma}_{MCD})}(i) \leq \sqrt{\chi^2_{p,0.975}}$;
- $w_i = 0$ inak;
- $d_{h,n}$ je korekčný faktor, v prípade dátových súborov s normálnym rozdelením pravdepodobnosti sa používa preto, aby sme získali eficientné odhady;

- $(\hat{\boldsymbol{\mu}}_{MCD}, \hat{\boldsymbol{\Sigma}}_{MCD})$ sú východiskové (predvolené) MCD odhady.

Tieto jednokrokové vážené odhady majú oproti vyššie uvedným MCD odhadom tú výhodu, že sú štatisticky viac eficientné. Preto sa výsledky iteratívneho algoritmu často používajú práve ako východiskové odhady v jednokrokovej metóde.

MCD odhady mnohorozmernej polohy a kovariancie, $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\Sigma}}$, sa používajú v hodnotení kvality regresného modelu ku konštrukcii regresného diagnostického grafu. Za týmto účelom nahradíme klasické odhady $\bar{\mathbf{x}}$ a $\mathbf{S}$ v Mahalanobisovej vzdialenosti pozorovaní $\mathbf{x_i}$ od centra distribúcie $\boldsymbol{\mu}$ vzhľadom ku kovariančnej štruktúre, danej kovariančnou maticou $\boldsymbol{\Sigma}$,

$$d(i) = \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right]^{1/2}, \tag{18}$$

kde $i = 1, \ldots, n$. Robustnými odhadmi, $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\Sigma}}$, dostaneme robustné Mahalanobisové vzdialenosti (hovoríme tiež len ako o robustných vzdialenostiach),

$$d_R(i) = \left[(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})\right]^{1/2}. \tag{19}$$

Ak náhodný výber $\mathbf{x}_i$, $i = 1, \ldots, n$ pochádza z p-rozmerného normálneho rozdelenia, riadia sa štvorce Mahalanobisových vzdialeností (klasických aj robustných) približne χ^2-rozdelením o p stupňoch voľnosti. Ak $d_R(i)$ je väčšia ako daná hraničná hodnota (väčšinou v tomto ohľade berieme odmocninu z 0,975-kvantilu χ^2-rozdelenia o p stupňoch voľnosti, $\sqrt{\chi^2_{p;0,975}}$) považujeme dané pozorovanie $\mathbf{x}_i$ za odľahlé.

Ako ďalší diagnostický nástroj na zistenie odľahlých hodnôt nám slúži *diagnostický graf*, kde sú zobrazené štandardizované LTS reziduá a robustné vzdialenosti počítané MCD metódou. Odľahlé hodnoty sa vyznačujú vysokou hodnotou štandardizovaných reziduí, to znamená, že sú vykreslené mimo rozsah ± 2.5, [4]. Dobré odľahlé pozorovania sú také, ktoré majú veľkú robustnú vzdialenosť a zároveň ležia v páse ± 2.5. Zlé odľahlé pozorovania ležia mimo daný pás a majú vysokú hodnotu robustnej vzdialenosti, vertikálne odľahlé hodnoty majú vysokú

hodnotu štandardizovaných reziduí a ležia mimo pás ± 2.5. Diagnostický graf si ukážeme v nasledujúcej kapitole na konkrétnych príkladoch.

4. Príklady

Na dvoch konkrétnych príkladoch budem aplikovať teoretické poznatky o jednotlivých metódach. Oba príklady sú zpracované v štatistickom softwari **R** a dátové súbory pre obidva príklady čerpám z balíka `robustbase`, ktorý načítame pomocou príkazu `library(robustbase)`.

4.1. Príklad 1: Hertzsprung-Russell (H-R) diagram hviezdokopov Cyg OB1

V prvom príklade budem pracovať s údajmi o hviezdach pochadzajúce z tzv. *Hertzsprung-Russell star dátového súboru*, ktorý sa používa pri vytvorení Hertzsprung-Russell (H-R) diagramu. Ide o 47 pozorovaní, kde každé pozorovanie prislúcha jednej hviezde. Po načítaní balíka `robustbase` je nutné načítať dátový súbor pomocou príkazu `data(starsCYG)`. Nezávislou premennou x je *logaritmus teploty na povrchu hviezdy* (T_e), závislou premennou Y je *logaritmus intenzity svetla hviezdy* (L/L_0). Konkrétne hodnoty jednotlivých pozorovaní sú uvedené v prílohe č.1, a získala som ich použitím príkazu `starsCYG`.

Na vytvorenie grafov bola použitá funkcia `plot`. Dátový súbor vykreslím pomocou bodového grafu, viď ľavý graf na Obrázku 1. Vidíme, že v ľavom hornom rohu sú štyri pozorovania, ktoré sú vzdialené od hlavného trendu dát. V pravom grafe je modrou priamkou vykresná LS-regresná priamka. Vidíme, že má klesajúcu tendenciu, dôkazom čoho je aj záporná hodnota odhadu regresného parametra $\hat{\beta}_1$, viď tabuľka 1. LTS-regresná priamka je červená a má rastúci charakter, pretože jej odhad regresného parametera $\hat{\beta}_1$ vyšiel kladný. Je zrejmé, že na rozdiel od regresnej priamky získanej pomocou MNŠ, charakterizuje robustná regresná priamka trend väčšiny pozorovaní.

Takto by sme určili testovaciu štatistiku T v prípade klasickej regresie. V prípade MNŠ môžeme urobiť testy významnosti jedného regresného parametra podľa následujúcej vety.

Veta 1. *([1], str. 84) Označme v_{ij} prvky matice $(\mathbf{X}^T\mathbf{X})^{-1}$ a nech*

$$T_j = \frac{\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j}{\sqrt{s^2 v_{jj}}}.$$

Potom pre každé $j = 0, \ldots, k$ platí $T_j \sim t_{n-(k+1)}$, teda pre testovaciu štatistiku T_j platí, že má Študentovo t*-rozdelenie o $n - k - 1$ stupňoch voľnosti.*

V prípade robustnej regresie by sme rozdelenie, resp. príslušnú p-hodnotu, uvedených testovacích štatistik T_j získaných pomocou príslušných robustných odhadov, určili pomocou metódy tzv. *cross validation* [15]. Takto nakoniec pracujú aj ostatné príslušné funkcie v štatistickom softwari **R**.

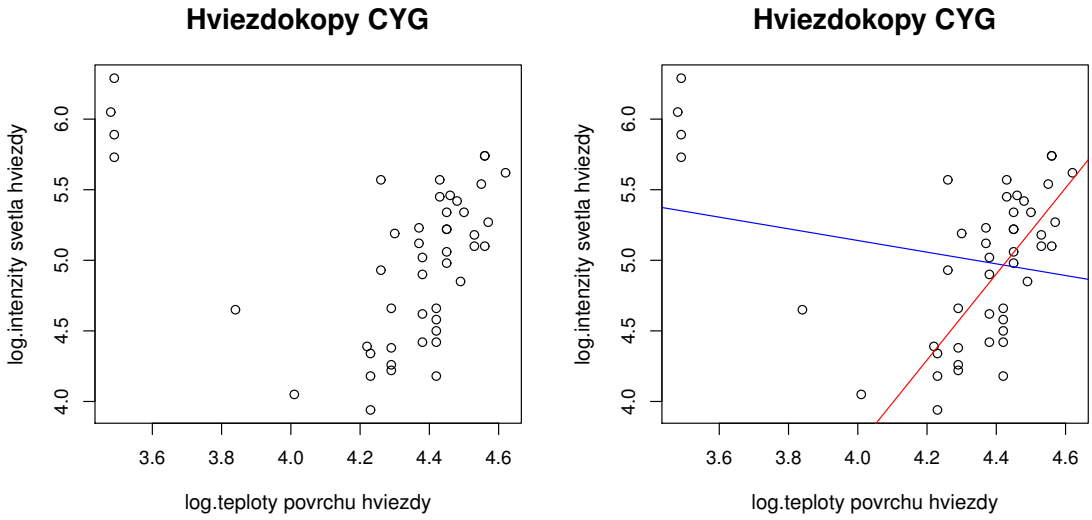
Použitím funkcie `summary` sme okrem jednotlivých odhadov regresných parametrov získali výsledky testovania, ich štatistické významnosti pomocou (štandardnej) T-štatistiky (vrátane príslušných p-hodnôt), čo je zaznačené v tabuľkách 1 a 2. Z týchto tabuliek je tiež vidieť, že smerodajné odchýlky sa pre regresné parametre pre LS a LTS regresiu výrazne nelíšia.

	Odhad parametra	**Smerodajná odchýlka**	**Hodnota test. štatistiky T**	**p-hodnota**
β_0	6.7935	1.2365	5.494	1.75×10^{-6}
β_1	-0.4133	0.2863	-1.444	0.156

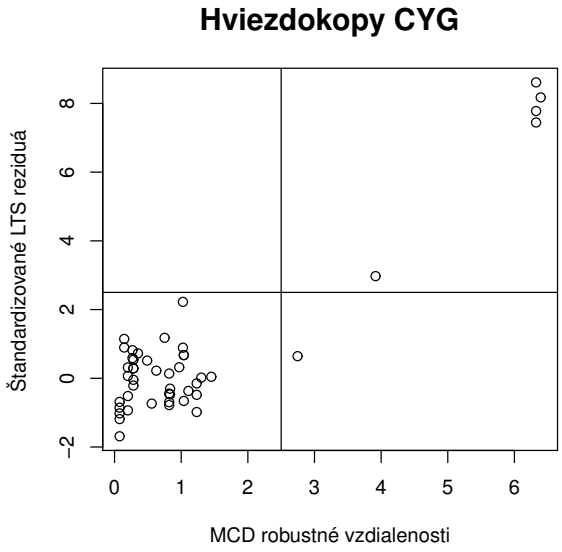
Tabuľka 1: Výsledky testovania pre LS regresiu

	Odhad parametra	**Smerodajná odchýlka**	**Hodnota test. štatistiky T**	**p-hodnota**
β_0	-8.5001	1.9263	-4.413	7.83×10^{-5}
β_1	3.0462	0.4373	6.965	2.39×10^{-8}

Tabuľka 2: Výsledky testovania pre LTS regresiu



Obr. 2: Vľavo: Vykreslenie dátového súboru. Vpravo: Dáta preložené LS a LTS rgresnou priamkou.



Obr. 3: MCD robustné vzdialenosti vs. štandardizované LTS reziduá.

Na obrázku, je diagnostický graf, kde je v páse ±2.5 zobrazené dobré odľahlé pozorovanie s vysokou hodnotou robustnej MCD vzdialenosti. V pravom hornom rohu nad pásom ±2.5 sú vplyvné pozorovania, o ktorých môžeme povedať, že zodpovedajú odľahlým pozorovaniam v ľavom hornom rohu na obrázku 2.

Tabuľka 3 obsahuje údaje o Mahalanobisových vzdialenostiach. Vzdialenosti, ktoré pri danej hladine $\alpha = 0.025$ prekročili hodnotu odmocniny kvantilu

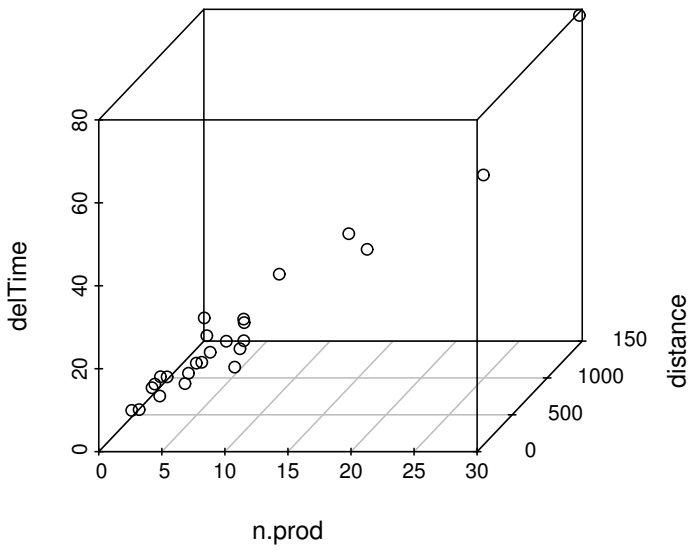
$\chi^2_{1;0.975} = 5.02390$ sú vyznačené tučným písmom.

Index	**MD**	**Index**	**MD**	**Index**	**MD**
1	0.2686	17	1.2321	33	0.2820
2	1.0391	18	0.0755	34	**6.3251**
3	1.0256	19	1.2321	35	1.2321
4	1.0391	20	**6.3251**	36	1.4520
5	0.7503	21	0.8192	37	0.8326
6	0.3508	22	0.8192	38	0.2820
7	3.9162	23	0.0755	39	0.8326
8	1.1079	24	0.5573	40	0.1444
9	1.0256	25	0.1998	41	0.1998
10	0.2686	26	0.0755	42	0.2820
11	**6.3251**	27	0.8192	43	0.6261
12	0.1444	28	0.1998	44	0.2820
13	0.4885	29	1.3009	45	0.9702
14	2.7462	30	**6.3939**	46	0.2820
15	0.8192	31	0.1998	47	0.0755
16	0.0755	32	1.0392		

Tabuľka 3: Mahalanobisové vzdialenosti

4.2. Príklad 2: Doba obsluhy predajného automatu

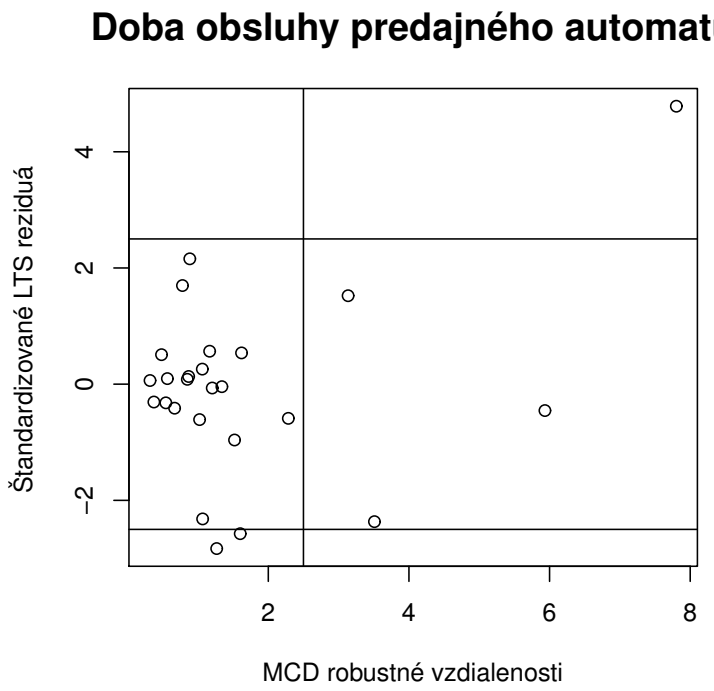
V ďalšom príklade budem pracovať s dátami o dobe obsluhy predajného automatu. Dáta sa nazývajú Delivery Time Data a pochádzajú z Montgomery and Peck (1982), majú 3 premenné a 25 pozorovaní. Úlohou je vysvetliť, čas potrebný na obsluhu predajného automatu pomocou počtu ponúkaných produktov automatom a prejdenou trasou (vzdialenosťou) obsluhujúceho personálu. Vystupujú tu dve nezávislé premenné, kde x_1 je počet produktov, x_2 vzdialenosť a nezávislá premenná Y ako doba obsluhy. Lineárny model bude mať regresnú funkciu v tvare: $\mathsf{E}(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Keďže v tomto príklade ide o viacnásobnú regresiu, dáta som musela vykresliť pomocou 3D grafu použitím príkazu `scatterplot3d()`, pozri obrázok 3.



Obr. 4: 3D graf: Bodové vykreslenie dátového súboru o čase doručenia.

Opäť prostredníctvom funkcie `summary()` som zistila informácie o testovaním a sú v tabuľkách 4 a 5. Odhady regresných parametrov vyšli pre LS aj pre LTS regresiu kladné, čo znamená že regresná primka bude mať rastúci charakter.

Smerodajné odchýlky sa opäť výrazne od seba nelíšia.



Obr. 5: MCD robustné vzdialenosti vs. štandardizované LTS reziduá.

Na obrázku 5 je opäť diagnostický graf, z ktorého je vidieť, že v páse ± 2.5 sú zobrazené tri dobre odľahlé pozorovania a nad týmto pásom v pravom hornom rohu je vplyvné pozorovanie. Vľavo dole sú dve vertikálne odľahlé pozorovania.

	Odhad parametra	**Smerodajná odchýlka**	**Hodnota test. štatistiky T**	**p-hodnota**
β_0	2.3412	1.0967	2.135	0.0442
β_1	1.6159	0.1707	9.464	3.25×10^{-9}
β_2	0.0144	0.0036	3.981	0.0006

Tabuľka 4: Výsledky testovania pre LS regresiu.

Tabuľka 6 obsahuje údaje o Mahalanobisových vzdialenostiach. Vzdialenosti, ktoré pri danej hladine $\alpha = 0.025$ prekročili hodnotu odmocniny kvantilu $\chi^2_{2;0.975} = 7.37778$ sú vyznačené tučným písmom.

	Odhad parametra	**Smerodajná odchýlka**	**Hodnota test. štatistiky T**	***p*-hodnota**
β_0	3.7196	0.9101	4.087	0.0006
β_1	1.4058	0.1313	10.708	1.73×10^{-9}
β_2	0.0162	0.0030	5.402	3.27×10^{-5}

Tabuľka 5: Výsledky testovania pre LTS regresiu.

i	$d_R(i)$	i	$d_R(i)$
1	1.2661	14	0.8675
2	0.8485	15	1.0615
3	1.0231	16	2.2850
4	0.8834	17	0.3179
5	0.5431	18	0.7794
6	0.3730	19	1.1661
7	1.2027	20	3.5094
8	0.4817	21	1.5198
9	**7.8030**	22	5.9339
10	1.6197	23	1.0661
11	3.1348	24	1.6008
12	1.3397	25	0.6676
13	0.5645		

Tabuľka 6: Mahalanobisové vzdialenosti

Záver

V teoretickej časti tejto práce som zhrnula informácie o regresnej analýze či už pre jednu alebo viacero vysvetľujúcich premenných. Pri odhade regresných parametrov sa štandardne používa metóda najmenších štvorcov, ktorá minimalizuje súčet štvorcov odchýlok hodnôt vysvetľovanej premennej od hodnôt vyrovnaných regresnou funkciou (reziduí). Je samozrejmé, že s meraním realizícii pozorovaní často dochádza k chybám, a preto sa v dátovych súboroch vyskytujú pozorovania, ktoré nenasledujú dátový trend. Na tieto pozorovania je metóda najmenších štvorcov citlivá, preto som sa v ďalšej teoretickej časti venovala robustným metódam. Tie sú už menej citlivé na výskyt odľahlých hodnôt, a je možné ich použitím takéto pozorovania odhaliť. Praktická časť bola venovaná dvom konkrétnym príkladom, na ktorých som aplikovala niektoré popísané metódy a postupy. Všetko bolo počítané a vykreslené v štatistickom softwari **R**, ktorému som venovala priestor hneď na úvod tejto práce a stručne som ho predstavila.

Práca mi prieniesla širší prehľad o tom, ako sa dá s dátovými súbormi pracovať a ako jednotlivé robustné metódy pomáhajú pri detekcii odľahlých pozorovaní. Doplnila som si tak znalosti, ktoré sa na základnom kurze štatistiky nevyučujú a zároveň som sa naučila pracovať s dvoma, pre mňa celkom novými, softwarmi ako je **R** a LaTeX. Dúfam, že táto práca bude slúžiť čitateľovi ako náhľad do danej problematiky a po jej prečítaní sa v nej bude vedieť orientovať.

Prílohy

Príloha 1

Index	x	y	Index	x	y
1	4.37	5.23	26	4.42	4.66
2	4.56	5.74	27	4.29	4.66
3	4.26	4.93	28	4.38	4.90
4	4.56	5.74	29	4.22	4.39
5	4.30	5.19	30	3.48	6.05
6	4.46	5.46	31	4.38	4.42
7	3.84	4.65	32	4.56	5.10
8	4.57	5.27	33	4.45	5.22
9	4.26	5.57	34	3.49	6.29
10	4.37	5.12	35	4.23	4.34
11	3.49	5.73	36	4.62	5.62
12	4.43	5.45	37	4.53	5.10
13	4.48	5.42	38	4.45	5.22
14	4.01	4.05	39	4.53	5.18
15	4.29	4.26	40	4.43	5.57
16	4.42	4.58	41	4.38	4.62
17	4.23	3.94	42	4.45	5.06
18	4.42	4.18	43	4.50	5.34
19	4.23	4.18	44	4.45	5.34
20	3.49	5.89	45	4.55	5.54
21	4.29	4.38	46	4.45	4.98
22	4.29	4.22	47	4.42	4.50
23	4.42	4.42			
24	4.49	4.85			
25	4.38	5.02			

Tabuľka 7: CYG OB1 Star Cluster Data

Príloha 2

Index	Počet produktov	Vzdialenosť	Doba obsluhy
1	7	560	16.68
2	3	220	11.50
3	3	340	12.03
4	4	80	14.88
5	6	150	13.75
6	7	330	18.11
7	2	110	8.00
8	7	210	17.83
9	30	1460	79.24
10	5	605	21.50
11	16	688	40.33
12	10	215	21.00
13	4	255	13.50
14	6	462	19.75
15	9	448	24.00
16	10	776	29.00
17	6	200	15.35
18	7	132	19.00
19	3	36	9.50
20	17	770	35.10
21	10	140	17.90
22	26	810	52.32
23	9	450	18.75
24	8	635	19.83
25	4	150	10.75

Tabuľka 8: Delivery Data

Literatúra

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

[2] Alfons, A.: *Robust statistics in practice*. Summer school, , Leuven, 2011.

[3] Blatná, D.: Robustní přístup v lineární regresi [online], dostupné z http://www.panda.hyperlink.cz/cestapdf/pdf08c3/blatna.pdf, [citované 14. 2. 2012].

[4] Filzmoser, P.: *Multivariate statistik, Institut für Statistik und Wahrscheinlichkeitstheorie*. Wien, 2007.

[5] Hindls, R., Hronová, S., Novák, I.: *Metody statistické analýzy pro ekonomy*, 2. přepracované vydání. Management press,´Praha, 2000.

[6] Hron, K., Kunderová P.: *Regresní analýza - podpůrný materiál*. KMA, PřF UP, Olomouc, 2011.

[7] Hubert, M.: *Robust Multivariate Statistics*. IASC Summer school, Leuven, 2011.

[8] Hubert, M., Rousseeuw, P. J., Van Aelst, S.: High-Breakdown Robust Multivariate Methods [online], dostupné z http://arxiv.org/pdf/0808.0657.pdf, [citované 3.4.2012].

[9] Cheng, T., Victoria-Feser, M. P.: High Breakdown Estimation of Multivariate Mean and Covariance With Missing Observations [online], dostupné z $http://www.hec.unige.ch/www/dms/hec_en/victoriafeser/recherche/paper-ertbs.pdf$, [citované 4.6.2012].

[10] Jednoduchá regrese [online], dostupné z $http://homel.vsb.cz/\ dom033/predmety/statistika/ucebni_text/15Regrese.pdf$, [citované 28.2.2012].

[11] Komárek, A.: *Hrátky s R*. Praha, Katedra pravděpodobnosti a matematické statistiky, Matematicko - fyzikální fakulta UK, Praha, 2009.

[12] Konečná, K.: *Výuka jazyka R*, Bakalářská práce, PřF, Masarykova univerzita, Brno, 2010 [online], dostupné z $http://is.muni.cz/th/270073/prif_b/Bakalarska_prace.pdf$, [citované 29.4.2012].

[13] Kulich, M.: *Stručný úvod do R* [online], dostupné z http://www.karlin.mff.cuni.cz/ kulich/vyuka/Rdoc/uvodrfpm.pdf, [citované 16.3.2012].

[14] Lineární regresní modely, [online], dostupné z $http://meloun.upce.cz/docs/research/chemometrics/methodology/6jedmetody.pdf$, [citované 7.3.2012].

[15] Marček, M.: *Viacnásobná štatistická analýza dát a modely časových radov v ekonómii*. FPF SU, Opava, 2009.

[16] Maronna, R., Martin, R.D., Yohai, V.J.: *Robust Statistics: Theory and methods*. Wiley, New York, 2006.

[17] Metoda nejmenších čtverců [online], dostupné z $http://herodes.feld.cvut.cz/mereni/mnc/mnc.php$, [citované 7.3.2012].

[18] Zvára, K.: *Regrese*. Matfyzpress, Praha, 2008.