Vysoká škola ekonomická v Prahe
Fakulta informatiky a štatistiky

Študijný program: Kvantitatívné metódy v ekonomike

Študijný obor: Štatistika



Aplikácia metódy bootstrap pri posúdení kvality lineárneho regresného modelu

Diplomová práca

Autor diplomovej práce: Bc. Bohuš Nemčovič

Vedúci diplomovej práce: Mgr. Milan Bašta, Ph.D.

Akademický rok 2017/2018

V prvom rade by som chcel poďakovať vedúcemu mojej diplomovej práce
Mgr. Milanovi Baštovi, Ph.D. za všetkú pomoc, cenné rady, profesionálny prístup
a poskytnuté materiály.
Takisto ďakujem svojim rodičom za možnosť študovať a za podporu počas celého
štúdia.
Ďakujem aj svojej priateľke Lenke za podporu a trpezlivosť.

Prehlásenie

Prehlasujem, že som diplomovú prácu zpracoval samostatne a že som uviedol všetky použité pramene a literatúry, z ktorých som čerpal.

V Prahe dňa

Podpis autora .....

Názov práce: Aplikácia metódy bootstrap pri posúdení kvality lineárneho regresného modelu

Autor: Bohuš Nemčovič

Katedra: Katedra štatistiky a pravdepodobnosti

Vedúci diplomovej práce: Mgr. Milan Bašta, Ph.D., Katedra štatistiky a pravdepodobnosti

Abstrakt: Výstavba regresného modelu zahŕňa viacero procedúr ako napríklad identifikácia vzdialených a vplyvných pozorovaní, transformácia premenných alebo výber vhodných premenných do modelu. Väčšinou sa závery odvodzujú z odhadnutého modelu bez toho, aby sa brali do úvahy procedúry, ktoré viedli k takémuto modelu. V tejto práci ukážeme, že to môže viesť k nepresným záverom. Zadefinujeme tri procedúry používané pri výstavbe regresného modelu a posúdime jednotlivý vplyv týchto procedúr na presnosť odhadov odmocniny zo strednej štvorcovej chyby odhadu regresnej funkcie v určitom bode a odhadu regresného parametru. Využijeme metódu bootstrap k spresneniu týchto odhadov. Analýzu predvedieme na viacerých modeloch.

Kľúčové slová: lineárny regresný model, metóda najmenších štvorcov, regresné data analytické funkcie, Monte Carlo, bootstrap, RMSE

Title: Application of bootstrap method during the assessment of the quality of linear regression model

Author: Bohuš Nemčovič

Department: Department of Statistics and Probability

Supervisor: Mgr. Milan Bašta, Ph.D., Department of Statistics and Probability

Abstract: The construction of a regression model consists of many procedures such as identification of outliers and influential points, transformation of variables or variable selection. Usually the inference is made from estimated regression model without conditioning to the procedures which made the model. In this work we will show it can lead to biased inferences. We will define three procedures used in construction of regression model and we will asses individual influence of these procedures on the precision of the estimated root of mean square errors of the estimate of the regression function at certain point and the estimate of the regression parameter. We will use bootstrap method to more accurately approximate these estimates. Also we will perform analysis on multiple models.

Keywords: linear regression model, ordinary least squares, regression data analytical functions, Monte Carlo, bootstrap, RMSE

Obsah

Úvod **2**

1 Regresná analýza **4**
1.1 Jednoduchý regresný model 4
1.1.1 Metóda najmenších štvorcov pre jednoduchý model 5
1.2 Viacnásobný lineárny regresný model 6
1.2.1 Metóda najmenších štvorcov pre viacrozmerný model . . . 8
1.2.2 Vlastnosti odhadov Metódou Najmenších Štvorcov 8
1.3 Regresné data analytické funkcie 12
1.3.1 DAF1 - Odľahlé pozorovania 13
1.3.2 DAF1 - Vplyvné pozorovania 14
1.3.3 DAF2 - Transformácia vysvetľovanej premennej 15
1.3.4 DAF3 - Výber premenných 16

2 Bootstrap **19**
2.1 Základná myšlienka metódy bootstrap 19
2.2 Úloha metódy bootstrap v našej analýze 20
2.2.1 Odhad regresnej funkcie v bode 21
2.2.2 Odhad regresného parametru β_1 22

3 Simulácie **23**
3.1 Odhad regresnej funkcie v bode 26
3.1.1 Vanila model 27
3.1.2 Nelineárny model 30
3.1.3 Hetero model 32
3.1.4 Kolineárny model 34
3.1.5 Nasýtený model 36
3.2 Efekt 38
3.2.1 Vanila model 39
3.2.2 Nelineárny model 40
3.2.3 Hetero model 42
3.2.4 Kolineárny model 43
3.2.5 Nasýtený model 44
3.3 Algoritmus odhadu RMSE odhadu regresnej funkcie 45
3.4 Výpočtová náročnosť 48

Záver **50**

Literatúra **52**

Úvod

Lineárny regresný model sa skladá zo systematickej časti a z náhodnej časti. Systematická časť špecifikuje lineárny vyťah medzi vysvetľujúcimi a vysvetľovanými premennými. Náhodná časť je určená rozdelením chybových zložiek, ktoré sú väčšinou nezávislé a rovnako rozdelené náhodné veličiny. V prípade, že poznáme skutočný model, môžeme priamo postúpiť k záverom, ktoré plynu z takéhoto modelu. V praxi však väčšinou máme k dispozicii iba určité dáta, medzi ktorými sa snažíme nájsť uspokojivý vzťah. Hľadanie uspokojivého modelu väčšinou zahŕňa niekoľko fázi, medzi ne napríklad patrí identifikácia odľahlých a vplyvných pozorovaní, transformácia a výber premenných do modelu. Spomenuté procedúry sú len jedny z mnohých, ktoré slúžia k určeniu finálneho modelu. Tieto procedúry všeobecne nemajú jednoznačne určený postup, pretože na riešenie jedného problému je možné použiť viacero metód. Častokrát po získaní finálneho modelu analytik vyvodzuje závery bez ohľadu na to aké procedúry boli použité k získaniu tohto modelu. Avšak ako ukážeme, tieto závery nemusia byť presné práve preto, že sme nebrali do úvahy procedúry, vďaka ktorým sme nadobudli náš model. Zvyčajne tieto úsudky nadsadzujú spoľahlivosť odhadnutých regresných koeficientov a takisto presnosť odhadu regresnej funkcie v určitom bode [10]. V mnohých odborných knihách je tento problém zdôraznený. V knihe [3] je spomenuté, že ak sme model získali pomocou datovej analýzy, štandardný vzorec na výpočet smerodatnej odchylky odhadu regresnej funkcie v bode môže byť podhodnotený.

Touto problematikou sa zaoberal Julian J. Faraway vo svojej práci [10], na ktorej základoch stojí táto práca. Naprogramoval procedúru, ktorá na základe datovej analýzy určila výsledný regresný model a predviedol ako dokáže metóda bootstrap zlepšiť podhodnotené klasické odhady smerodatnej odchylky odhadnutých regresných koeficientov a odhadnutých regresných funkcii. Na rozdiel od práce [10], kde boli aplikované všetky procedúry slúžiace k výberu modelu na konkrétne dátové súbory, v našej práci aplikujeme osobitne tri procedúry, ktoré sa často používajú pri výstavbe lineárného regresného modelu, a posúdime do akej miery je klasický odhad smerodatnej odchylky podnodnotený pri použití každej z týchto procedúr. Faraway v práci [10] získal konečný(finálny) model postupným aplikovaním všetkých zadefinovaných procedúr na určitý datový súbor, my v našej práci získame viacero konečných(finálnych) modelov aplikovaním procedúr jednotlivo, teda z jedného datového súboru získame tri rôzne finálne modely. Každý z týchto troch modelov vznikol aplikovaním inej procedúry. Následne použijeme metódu bootstrap na zlepšenie klasického odhadu smerodatnej odchylky. Cieľom práce je ukázať, že po aplikovaní procedúr na výstavbu lineárného regresného modelu už klasické odhady smerodatnej odchylky určitých odhadov nie sú presné. Zároveň je naším cieľom posúdiť do akej miery jednotlivé procedúry znižujú presnosť odhadnu smerodatnej odchylky odhadu regresnej funkcie v bode

alebo odhadu regresného parametru a ukázať pomocou metódy bootstrap možnosť opravy.

V prvej kapitole si priblížime základy regresnej analýzy. Definujeme jednoduchý regresný model, viacnásobný regresný model, metódu najmenších štvorcov pre obidva prípady a ich vlastnosti. Ďalej v našej práci predstavíme dátovú analýzu predchádzajúcu výberu modelu a jednotlivé metódy(procedúry) označíme ako data analytické funkcie, skrátene DAF. V druhej kapitole si priblížime fungovanie metódy bootstrap z teoretického aj praktického hľadiska. V poslednej tretej kapitole na simuláciach ukážeme aký dopad majú jednotlivé data analytické funkcie aplikované samostatne na klasický odhad smerodatnej chyby odhadu(odhadu regresnej funkcie v bode a odhadu regresného parametru) a ukážeme do akej miery dokáže bootstrap zlepšiť tento odhad.

Kapitola 1

Regresná analýza

Regresná analýza je súhrn štatistických metód na odhadnutie závislosti medzi jednotlivými premennými. Opovedá na otázky aká je závislosť medzi vysvetľujúcou a jednou alebo viacerými vysvetľovanými premennými. Predpovedá budúce hodnoty vysvetľovanej premennej, zisťuje, ktoré vysvetľujúce premenné sú dôležité a odhaduje vplyv zmeny vysvetľujúcich premenných na hodnotu vysvetľovaných premenných. Regresný model je definovaný ako vzťah medzi vysvetľovanou veličinou $Y = (y_1,...,y_n)$ a vysvetľujúcimi veličinami $X_1,...,X_p$, kde $X_j = (x_{1j},...,x_{nj})$ pre $j = 1,...,p$, n je počet pozorovaní.

1.1 Jednoduchý regresný model

Pri jednoduchom lineárnom regresnom modeli uvažujeme, že máme len jednu vysvetľovanú premennú $Y = (y_1,y_2,...,y_n)$ nadobúdajúcu hodnoty y_i pre i-té pozorovanie a jednu vysvetľujúcu premennú $X = (x_1,...,x_n)$ nadobúdajúcu hodnoty x_i pre i-té pozorovanie. Uvažujme teda na začiatku jednoduchý lineárny regresný model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{pre} \quad i = 1,...,n, \tag{1.1}$$

kde β_0 a β_1 sú regresné parametre, β_0 je stredná hodnota Y pre $X = 0$. β_1 sa dá interpretovať ako zmena strednej hodnoty Y pri zmene X o jednotku. Predpokladáme, že $X = (x_1,...,x_n)$ je nenáhodná spojitá kvantitatívna vysvetľujúce premenná. Chybová zložka modelu, označíme ϵ, spĺňa takzvanú slabú sadu predpokladov. Slabá sada predpokladov zahŕňa:

1. Stredná hodnota chybovej zložky je nulová, $E(\epsilon_i) = 0$, pre $i = 1,...,n$.
2. Konštantný rozptyl chybovej zložky, $Var(\epsilon_i) = \sigma^2 < \infty$, pre $i = 1,...,n$.
3. Nulovú kovarianciu medzi ϵ_i a ϵ_j pre $i = 1,...,n, j = 1,...,n$, kde $i \neq j$. Teda $C(\epsilon_i,\epsilon_j) = 0$ pre $i = 1,...,n, j = 1,...,n$, kde $i \neq j$.

Nech η_i je regresná funkcia. V prípade jednoduchého regresného modelu 1.1 je definovaná ako

$$\eta_i = E(y_i) = \beta_0 + \beta_1 x_i, \quad \text{pre} \quad i = 1,...,n, \tag{1.2}$$

teda ako stredná hodnota pre vysvetľovanú premennú Y rovnú y_i. Označme bodové odhady β_0 a β_1 ako b_0 a b_1. Potom

$$\hat{y}_i = b_0 + b_1 x_i \tag{1.3}$$

je vyrovnaná hodnota pre i-té pozorovanie a

$$e_i = y_i - \hat{y}_i \tag{1.4}$$

je reziduum i-tého pozorovania. Definujeme ďalej $E(y_*) = E(Y|X = x_*)$ ako strednú hodnotu odozvy Y pre X rovné x_*, kde x_* je nová hodnota vysvetľujúcej premennej,

$$\eta_* = E(y_*) = E(Y|X = x_*) = \beta_0 + \beta_1 x_*. \tag{1.5}$$

Nech $\hat{\eta}_*$ je bodový odhad regresnej funkcie v bode x_* daný ako

$$\hat{\eta}_* = b_0 + b_1 x_*. \tag{1.6}$$

Nech $Y = y_*$ je nové pozorovanie prislúchajúce hodnote $X = x_*$. Platí, že

$$y_* = \eta_* + \epsilon_*, \tag{1.7}$$

kde predpokladáme, že $E(\epsilon_*) = 0$, $Var(\epsilon_*) = \sigma^2$ a kovariancia $C(\epsilon_*,\epsilon_i) = 0$ pre $i = 1,...,n$. Bodový odhad nového pozorovania je daný ako

$$\hat{y}_* = \hat{\eta}_* = b_0 + b_1 x_*. \tag{1.8}$$

Chybu predpovedi(Error) definujeme ako

$$y_* - \hat{y}_* = \beta_0 - b_0 + (\beta_1 - b_1)x_* + \epsilon_*. \tag{1.9}$$

1.1.1 Metóda najmenších štvorcov pre jednoduchý model

V tejto sekcii si predstavíme *metódu najmenších štvorcov*. Jedná sa o veľmi rozšírenú metódu, ktorá odhaduje parametre regresného modelu. Pre jednoduchý regresný model 1.1 je b_0 a b_1 odhad regresných koeficientov β_0 a β_1, ktorý minimalizuje výraz $RSS_1(\beta_0,\beta_1)$ ako funkciu β_0 a β_1, kde RSS_1 je definované ako

$$RSS_1(\beta_0,\beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{1.10}$$

Riešenie problému minimalizácie $\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$ získame parciálnym derivovaním funkcie RSS_1 podľa β_0 a β_1. Po krátkych algebraických operáciach získame bodové odhady regresných koeficientov

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1.11}$$

$$b_0 = \bar{y} - b_1\bar{x}, \tag{1.12}$$

kde $\bar{x}$ a $\bar{y}$ je priemer hodnôt vysvetľujúcej premennej $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a priemer hodnôt vysvetľovanej premennej $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Hodnotu $RSS(b_0, b_1)$ voláme reziduálny súčet štvorcov.

1.2 Viacnásobný lineárny regresný model

V našej práci nás však bude zaujímať model s viac ako jednou vysvetľujúcou premennou. V tomto prípade hovoríme o viacnásobnom lineárnom regresnom modeli

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ip} + \epsilon_i, \quad \text{pre} \quad i = 1,...,n, \tag{1.13}$$

kde x_{i1},...,x_{ip} sú hodnoty vysvetľujúcich premenných X_1,...,X_p. Maticovo by sme mohli vzťah 1.13 zaznačiť takto

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{1.14}$$

kde

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \tag{1.15}$$

,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{1.16}$$

$\boldsymbol{y}$ a $\boldsymbol{\epsilon}$ sú náhodné vektory, $\boldsymbol{X}$ je regresná matica a $\boldsymbol{\beta}$ je vektor regresných parametrov. Stredná hodnota vektoru $\boldsymbol{\epsilon}$ je rovná nulovému vektoru $\mathbf{0}$.

$$E(\boldsymbol{\epsilon}) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{1.17}$$

.

Kovariančná matica $\boldsymbol{\epsilon}$ je daná ako

$$C(\boldsymbol{\epsilon}) = E((\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon}))(\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon}))^T) = \sigma^2 \mathbb{1}, \tag{1.18}$$

kde $\mathbb{1}$ je jednotková matica a σ^2 je rozptyl chybovej zložky. Regresná funkcia $\boldsymbol{\eta} = (\eta_1,...,\eta_n)$ je maticovo daná následovne

$$\boldsymbol{\eta} = E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}. \tag{1.19}$$

Vektor vyrovnaných hodnôt je daný ako

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{H}\boldsymbol{y}, \tag{1.20}$$

kde $\boldsymbol{H}$ je takzvaná projekčná matica definovaná ako

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{1.21}$$

a

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \tag{1.22}$$

je vektor odhadnutých regresných koeficientov. Prvky projekčnej matice h_{ij} sú dané vzorcom

$$h_{ij} = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_j = \boldsymbol{x}_j^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i, \tag{1.23}$$

kde $\boldsymbol{x}_i$ je $(p+1)$-členný vektor,ktorý obsahuje hodnoty i-tého riadku regresnej matice $\boldsymbol{X}$.

$$\boldsymbol{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \tag{1.24}$$

Nás budú v nasledujúcej sekcii 1.3 zaujímať diagonálne prvky projekčnej matice h_{ii}. Prvky h_{ii} sa tiež nazývajú *leverage* i-tého pozorovania a merajú ako ďaleko je pozorovanie x_i od ostatných bodov. Bod s vysokou hodnotou h_{ii} je zvyčajne vplyvné pozorovanie [9]. Tento poznatok sa nám zíde pri identifikácii odľahlých a vplyvných pozorovaní.
Vektor reziduí v tomto prípade je určený ako

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}. \tag{1.25}$$

Ďalej stále predpokladáme platnosť slabej sady predpokladov. Podľa [2] definujeme hodnotu regresnej funkcie η_* v bode $\boldsymbol{x_*}$ ako

$$\eta_* = E(\boldsymbol{y}_*) = \boldsymbol{x}_*^T \boldsymbol{\beta}, \tag{1.26}$$

kde

$$\boldsymbol{x_*} = \begin{bmatrix} 1 \\ x_{*1} \\ x_{*2} \\ \vdots \\ x_{*k} \end{bmatrix}$$

a x_{*1},x_{*2},...,x_{*k} predstavuje zvolenú kombináciu hodnôt vysvetľujúcich premenných. Ďalej bodový odhad regresnej funkcie η_* definujeme pre bod $\boldsymbol{x_*}$ ako

$$\hat{\eta}_* = \boldsymbol{x}_*^T \boldsymbol{b}. \tag{1.27}$$

Bodový odhad nového pozorovania zostrojíme za rovnakých predpokladov ako u jednoduchého regresného modelu. Bodový odhad nového pozorovania pre viacrozmerný model je

$$\hat{\boldsymbol{y}}_* = \boldsymbol{x}_*^T \boldsymbol{b} \tag{1.28}$$

a chyba predpovedi(error)

$$\boldsymbol{y}_* - \hat{\boldsymbol{y}}_* = \boldsymbol{x}_*^T(\boldsymbol{\beta} - \boldsymbol{b}) + \epsilon_* \tag{1.29}$$

1.2.1 Metóda najmenších štvorcov pre viacrozmerný model

Pre viacrozmerný model 1.13 je odhad $\boldsymbol{b} = (b_0,b_1,...,b_n)$ vektoru regresných parametrov $\boldsymbol{\beta} = (\beta_0,...,\beta_p)$ taký, ktorý minimalizuje výraz RSS_2 ako funkciu $\boldsymbol{\beta}$, kde RSS_2 je definované ako

$$RSS_2(\boldsymbol{\beta}) = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \\ = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2. \tag{1.30}$$

Parciálne derivácie vyššie uvedeného výrazu podľa $\boldsymbol{\beta}$ vedie na sústavu normálových rovníc

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}. \tag{1.31}$$

Dá sa ukázať, že táto sústava normálnych rovníc je vždy konzistentná(to znamená, že má aspoň jedno riešenie). Ak je štvorcová matica $\boldsymbol{X}^T\boldsymbol{X}$ regulárna, potom má inverziu a riešenie normálových rovnic je práve jedno a to

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.32}$$

Hodnota $RSS_2(\boldsymbol{\beta})$ je reziduálny súčet štvorcov.

1.2.2 Vlastnosti odhadov Metódou Najmenších Štvorcov

Odhad b_1 je takzvaný lineárny odhad, pretože je lineárnou kombináciou náhodných premenných y_i. Môžeme teda písať

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i y_i, \tag{1.33}$$

kde

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1.34}$$

Pri platnosti 1. predpokladu zo slabej sady predpokladov, teda pri platnosti $E(e_i) = 0$ pre $i = 1,...,n$, platí, že odhady získané *metódou najmenších štvorcov*(ďalej ako MNČ) sú nevychýlené

$$E(b_1) = \beta_1, \tag{1.35}$$

$$E(b_0) = \beta_0. \tag{1.36}$$

Za platnosti 2. a 3. predpokladu zo slabej sady predpokladov rozptyl týchto odhadov je

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1.37}$$

$$Var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right). \tag{1.38}$$

Odhadnúť σ^2 v našom jednoduchom modeli môžeme pomocou tzv. *rezidualného rozptylu*, ktorý je definovaný ako

$$s^2(e) = \frac{\sum_{i=1}^{n} e_i^2}{n-2}. \tag{1.39}$$

Odmocnina z rezidualného rozptylu je tzv. *reziduálna smerodatná odchylka* $s(e)$. V prípade splnenia slabej sady predpokladov a pokiaľ je náš model správne špecifikovaný sa dá ukázať, že

$$E(s^2(e)) = \sigma^2. \tag{1.40}$$

Pre *odhady rozptylov odhadov* regresných parametrov platí, že

$$s^2(b_0) = s^2(e) \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right), \tag{1.41}$$

$$s^2(b_1) = \frac{s^2(e)}{\sum_{i=1}^{n}(x_i - \bar{x^2})}. \tag{1.42}$$

Hodnoty $s(b_0)$ a $s(b_1)$(odmocniny $s^2(b_0)$ a $s^2(b_1)$) nazývame *odhady smerodatných chýb odhadov* regresných parametrov.
Pre odhady MNČ platí pre viacrozmerný model

$$E(\boldsymbol{b}) = \boldsymbol{\beta}. \tag{1.43}$$

Odhad vektoru regresných parametrov je nevychýleným odhadom. Ďalej máme

$$Var(\boldsymbol{b}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1}, \tag{1.44}$$

kde $Var(\boldsymbol{b})$ je kovariančná matica, ktorá ma na diagonále rozptyly jednotlivých b_0,b_1,...,b_p. Odhad kovariančnej matice potom môžeme zapísať ako

$$\widehat{Var}(\boldsymbol{b}) = s^2(e) \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1}. \tag{1.45}$$

Odhad smerodatnej odchylky pre parameter b_j je j+1-vý diagonálny prvok matice

$$s(\boldsymbol{b}) = s(e)\sqrt{(\boldsymbol{X}^T \boldsymbol{X})^{-1}}. \tag{1.46}$$

Teda pre $s(b_1)$ platí

$$s(b_1) = \left((s(e)\sqrt{(\boldsymbol{X}^T \boldsymbol{X})^{-1}} \right)_{[2,2]} \tag{1.47}$$

Priblížeme aj vlastnosti vektoru vyrovnaných hodnôt $\hat{\boldsymbol{y}}$

$$E(\hat{\boldsymbol{y}}) = E(\boldsymbol{X}\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta}. \tag{1.48}$$

$$Var(\hat{\boldsymbol{y}}) = Var(\boldsymbol{X}\boldsymbol{b}) = \sigma^2 \boldsymbol{H}. \tag{1.49}$$

Platí, že

$$Var(\hat{y}_i) = \sigma^2 h_{ii} \quad \text{pre} \quad i = 1,...,n, \tag{1.50}$$

kde h_{ii} je i-tý prvok na diagonále matice $\boldsymbol{H}$.

Na ďalšiu vlastnosť odhadov získaných metódou najmenších štvorcov potrebujeme *Gauss-Markov teorém*, ktorý tvrdí, že za splnenia slabej sady predpokladov sú odhady b_0 a b_1 najlepšími odhadmi(v zmysle, že majú najmenší rozptyl) zo všetkých lineárnych nevychýlených odhadov. Hovoríme, že odhad regresných parametrov metódou najmenších štvorcov je tzv. *BLUE*, čo znamená *Best Linear Unbiased estimator*.

V ďalšom texte sa stretneme s *odmocninou zo strednej štvorcovej chyby* (ďalej ako $RMSE$) odhadu. Definujeme *strednú štvorcovú chybu odhadu* b_j parametru β_j. Túto chybu rozložíme na rozptyl odhadu a kvadrát vychýlenia odhadu

$$\begin{aligned} MSE(b_j) &= E((b_j - \beta_j)^2) \\ &= E((b_j - E(b_j) + E(b_j) - \beta_j)^2) = \\ &= Var(b_j) + (B(b_j))^2, \end{aligned} \tag{1.51}$$

kde $B(b_j)$ je vychýlenie odhadu b_j. Pre nevychýlené odhady platí, že

$$MSE(b_j) = Var(b_j). \tag{1.52}$$

$RMSE$ definujeme ako

$$RMSE(b_j) = \sqrt{MSE(b_j)}. \tag{1.53}$$

Ďalej budeme klasické odhady metódou najmenších štvorcov označovať ako „*naivné*“ odhady.

Stále predpokladáme platnosť slabej sady predpokladov. Pre bodový odhad regresnej funkcie η_*, viď 1.27, platia podľa [2] nasledujúce vlastnosti. Pre strednú hodnotu odhadu môžeme písať

$$E(\hat{\eta}_*) = E(b_0 + b_1 x_*) = \beta_0 + \beta_1 x_* = \eta_*. \tag{1.54}$$

Pre rozptyl odhadu regresnej funkcie platí

$$Var(\hat{\eta}_*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x}^2)} \right). \tag{1.55}$$

Pre odhad rozptylu odhadu regresnej funkcie platí

$$s^2(\hat{\eta}_*) = s^2(e) \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x}^2)} \right) \tag{1.56}$$

a pre odhad smerodatnej chyby odhadu regresnej funkcie platí

$$s(\hat{\eta}_*) = s(e) \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x}^2)}}. \tag{1.57}$$

Pre viacrozmerný model sme si definovali regresnú funkciu $\boldsymbol{\eta}_*$ v bode $\boldsymbol{x}_*$, viď 1.26. Za platnosti slabej sady predpokladov pre strednú hodnotu odhadu regresnej funkcie v bode $\boldsymbol{x}_*$ platí podľa [2], že

$$E(\hat{\boldsymbol{\eta}}_*) = E(\boldsymbol{x}_*^T \boldsymbol{b}) = \boldsymbol{x}_*^T \boldsymbol{\beta}. \tag{1.58}$$

Pre rozptyl odhadu regresnej funkcie v bode $\boldsymbol{x}_*$ platí

$$Var(\hat{\boldsymbol{\eta}}_*) = Var(\boldsymbol{x}_* \boldsymbol{b}) = \boldsymbol{x}_*^T Var(\boldsymbol{b}) \boldsymbol{x}_* = \sigma^2 \boldsymbol{x}_*^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_*. \tag{1.59}$$

Pre nevychýlený odhad rozptylu odhadu regresnej funkcie v bode $\boldsymbol{x}_*$ platí

$$s^2(\hat{\boldsymbol{\eta}}_*) = s^2(e) \boldsymbol{x}_*^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_*. \tag{1.60}$$

Ďalej dá sa ukázať, že pre odhad smerodatnej chyby odhadu regresnej funkcie v bode $\boldsymbol{x}_*$ platí

$$s(\hat{\boldsymbol{\eta}}_*) = s(\boldsymbol{x}_*^T \boldsymbol{b}) = s(e) \sqrt{\boldsymbol{x}_*^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_*}. \tag{1.61}$$

Keď budeme robiť simulácie, tento klasický odhad 1.61 označíme ako *naivný*. Na začiatku sme si definovali určitý, jednoznačne definovaný model 1.13. V bežnej praxi bohužial často nie je známy skutočný model a je nutné tento vzťah vyhľadať pomocou dát, ktoré máme k dispozícii. Avšak dáta, ktoré máme k dispozícii nemusia byť ani zďaleka optimálne. Ako to už býva zvykom, analytik sa často stretáva s

- chýbajúcimi dátami,
- pozorovaniami, ktoré majú oproti väčšine iných pozorovaní výrazne odlišnú hodnotu, voláme ich odľahlé pozorovania prípadne outliery,
- vplyvnými pozorovaniami, to sú buď jednotlivé pozorovania alebo malá skupinka pozorovaní, ktoré majú silný vplyv na konečnú podobu odhadu regresnej funkcie,
- dátami, ktoré zjavne nie sú homoskedastické, teda nemajú konštantný rozptyl,
- dátami so zjavne nelineárnym priebehom strednej hodnoty odozvy.

Regresná analýza pri výstavbe regresného modelu preto v sebe zahŕňa postupy, ktoré upravujú dostupné dáta, tak aby výsledný model bol čo najjednoduchší a čo najviac zobrazoval vzťah medzi pozorovaniami. Medzi najčastejšie operácie na dátach by sme mali zmieniť identifikáciu a odstránenie odľahlých a vplyvných pozorovaní. Pri výstavbe regresného modelu je takisto nutné určiť, ktoré premenné zaradiť do modelu a ktoré nie. Respektíve zistiť, ktoré premenné sú významné. Tieto operácie súhrnne pomenujeme ako *data analytické funkcie*, skrátene DAF. Regresná analýza pozostáva z mnohých analytických postupov a procedúr. Je nutné poznamenať, že sa jedná o nejednoznačnú analýzu, pretože na danú úlohu je možné použiť viacero metód, ktoré môžu odhadnúť rozdielny model analyzovaných dát. Tým pádom DAF môže byť veľké množstvo. My sme zmienili iba zopár, ktoré sa najčastejšie používajú. Vytvoríme DAF, ktoré identifikujú a

odstránia vzdialené a vplyvné pozorovania, ktoré určia vhodnú transformáciu a transformujú vysvetľovanú premennú. Pozrieme sa taktiež na DAF, ktorá určí vhodné premenné do odhadnutého modelu. Na to, aby sme mohli posúdiť vplyv danej metódy na dátach budeme potrebovať vytvoriť také procedúry v programe, ktoré nám ich umožnia použiť viackrát za sebou. Regresná analýza je neexaktná procedúra, ale v našom prípade jednotlivé DAF budú musieť byť jednoznačne definované tak, aby subjektívny postup vôbec nezasahoval do analýzy. Našu regresnú analýzu teda obmedzíme na procedúry, ktoré sa dajú presne určiť pomocou výpočetných metód. Regresná analýza môže byť vnímaná ako proces, pri ktorom aplikujeme určité za sebou nasledujúce data analytické funkcie, ktoré nám určujú ako má vyzerať náš odhadnutý model. Naším cieľom bude otestovať aký vplyv na presnosť výsledného odhadnutého modelu(presnosť odhadu odmocniny zo strednej štvorcovej chyby odhadu regresnej funkcie a presnosť odhadu odmocniny zo strednej štvorcovej chyby odhadu regresných parametrov) majú zvolené data analytické funkcie.

Pri našej analýze predpokladáme, že máme dáta (x_{i1},x_{i2},...,x_{ip},y_i), kde x_{ij} je i-tá hodnota j-tej vysvetľujúcej premennej a y_i je i-tá hodnota vysvetľovanej premennej pre $i = 1,...,n$, n je počet pozorovaní a $j = 1,...,p$. Našim cieľom je odhadnúť model tvaru

$$g(y_i) = \beta_0 + \sum_{j=1}^{p'} \beta_j f_j(x_{ij}) + w_i \epsilon_i, \tag{1.62}$$

kde p' môže byť väčšie ako p v prípade, že model obsahuje kvadratické alebo interakčné členy. Rovné alebo menšie ako p, ak sú niektoré vysvetľujúce premenné vynechané. Náhodná zložka ϵ_i má normálne rozdelenie so strednou hodnotou $E(\epsilon_i) = 0$ a rozptylom $Var(\epsilon_i) = 1$. Funkcie g a f_j sú transformácie a w_i sú váhy. Jedným z dôsledkov w_i je, že ovplyvňujú rozptyl celého modelu.

1.3 Regresné data analytické funkcie

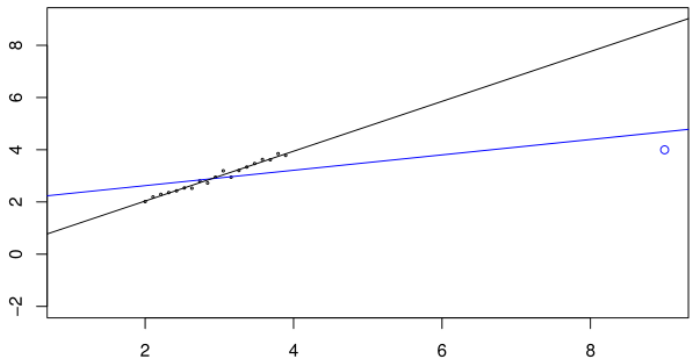
Teoreticky sme predstavili ako bude vyzerať náš odhadnutý model. Východzí model, ktorý odhadujeme na začiatku analýzy ešte pred použitím jednotlivých DAF bude obsahovať všetky dostupne vysvetľujúce premenné X_1,X_2,...,X_p s odhadnutými regresnými koeficientami b_0,b_1,b_2,...,b_p vypočítanými podľa známej metódy najmenších štvorcov, viď 1.2.1. Procedúry, ktoré aplikujeme na prvotný model zhrnieme v následujúcich sekciách. Pre lepšiu zrozumiteľnosť prikladáme aj obrázok schémy 1.1. Prvotný dátový súbor sú dáta, ktoré máme k dispozícii, a ktoré chceme odhadnúť regresným modelom. Východzí, respektíve prvotný model je taký, ktorý vznikne po aplikovaní metódy najmenších štvorcov na dáta, ktoré neboli nijako pozmenené. Potom nasleduje aplikovanie jednej z troch data analytických funkcii. Po aplikovaní DAF1 na identifikáciu a odstránenie odľahlých a vplyvnývh pozorovaní na prvotný datový súbor dostaneme sekundárny dátový súbor, ktorý už tieto pozorovania neobsahuje. Odhadneme model zo sekundárného dátového súboru a ďalej posudzujeme kvalitu takto získaného finálného modelu. Pri aplikovaní DAF2, teda určenie najvhodnejšej transformácie vysvetľovanej premennej odhadneme modely pre každú transformáciu zvlášť a porovnáme, ktorý z týchto transformovaných modelov(aj identitu v tomto prípade považujeme za

transformáciu) je najuspokojivejší. Takýto model označíme za finálny a postúpime k posúdeniu kvality takéhoto modelu. Pri aplikovaní DAF3 na prvotný model, pomocou predprogramovanej funkcie, dôjde na základe AIC kritéria, viď. 1.3.4, k výberu premenných do finálného modelu, na ktorom následne je posúdime kvalitu regresného modelu.

My sme v našej práci použili 3 data analytické funkcie ale pri výstavbe regresného modelu sa môže používať viac analytických procedúr. V práci [10] boli použité DAF, ktoré napríklad upravujú váhu chybovej zložky v prípade, že chybová zložka je heteroskedastická, alebo aplikujú Box-Cox transformáciu na vysvetľovanú premennú.

1.3.1 DAF1 - Odľahlé pozorovania

Pri niektorých pozorovaniach sa stáva, že svojou hodnotou nekorešpondujú do celkovej väčšiny pozorovaní. Pri preložení dát priamkou väčšina pozorovaní leží blízko regresnej priamky ale zopár pozorovaní leží mimo. Pozorovania, ktoré nenasledujú rovnaký model ako ostatné pozorovania voláme *odľahlé pozorovania* [2]. Prvou data analytickou funkciou nutnou na korektnú konštrukciu regresného modelu bude správne určenie a odstránenie odľahlých pozorovaní. Odľahlé pozorovania totiž môžu negatívne vychýliť sklon regresnej funkcie.[1]



Obr. 1.2: Ukážka ako dokáže odľahlé pozorovanie(biela bodka) ovplyvniť smer regresnej funkcie(modrá priamka). Regresná funkcia(čierna priamka) vznikne po odstránení odľahlého pozorovania z dátového súboru.

Keďže potrebujeme, aby táto operácia bola vykonaná objektívne, určíme odľahlé pozorovania pomocou výpočtu externe študentizovaných reziduí pre každé pozorovanie a následne tie, ktoré prekročia kritickú hodnotu vypočítanú pomocou Bonferroniho korekcie, vylúčime z analýzy. Predpokladajme model

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i, \tag{1.63}$$

ktorý odhadneme z dát pomocou metódy najmenších štvorcov. Ďalej pracujeme s takýmto modelom. Na výpočet externe študentizovaných reziduí použijeme vzorec

[1] Funkcia v R-ku je naprogramovaná pod názvom *outlier_fun*. Okrem odľahlých pozorovaní je v nej zahrnuté aj vyhľadávanie vplyvných pozorovaní.

$$t_i = \frac{y_i - \hat{y}_{i(-i)}}{s_{(-i)}(e)\sqrt{(\mathbf{x_i}^T(\mathbf{X_{(-i)}}^T\mathbf{X_{(-i)}})^{-1}\mathbf{x_i}) + 1}}, \tag{1.64}$$

kde $s_{-i}(e)$ je reziduálna smerodatná odchylka z odhadu modelu, kde bolo i-té pozorovanie odobraté. Ďalej y_i je i-té pozorovanie a $\hat{y}_{i(-i)}$ je vyrovnaná hodnota získaná na základe odhadu regresnej funkcie pri odobraní i-tého pozorovania. Ďalej $\mathbf{X_{(-i)}}$ značí regresnú maticu $\mathbf{X}$, kde bol vypustený i-tý riadok a $\mathbf{x_i}$ je i-tý stĺpec matice $\mathbf{X}$. Ak by sme dopredu predpokladali, že i-té pozorovanie je odľahlé, mohli by sme priamo porovnať, či t_i je v absolútnej hodnote väčšie ako kritická hodnota $t_{0.975}[n-p-1]$ teda väčšie ako 97,5% kvantil študentovho rozdelenia s $n-p-1$ stupňami voľnosti, kde n je počet dát a p je počet parametrov. Avšak my chceme, aby tento proces prebiehal automaticky bez subjektívneho rozhodovania. Preto pre všetkých n dát musíme testovať, či testová štatistika 1.65 spadá do kritického oboru. Dá sa ukázať, že t_i sa dá vypočítať ako

$$t_i = \frac{e_i}{s_{(-i)}(e)\sqrt{1-h_{ii}}}, \tag{1.65}$$

kde $s_{(-i)}(e)$ sa vypočíta ako

$$s_{(-i)}(e) = \sqrt{\frac{(n-p)s^2(e) - e_i^2/(1-h_{ii})}{n-p-1}}. \tag{1.66}$$

Zo vzorca 1.66 vyplýva, že nie je nutné odstraňovať z dát i-té pozorovania, ale dajú sa externe študentizované reziduá vypočítať pomocou klasických reziduí e_i, výberového rozptylu reziduí $s^2(e)$ a projekčnej matice $\mathbf{H}$ podľa vzorca 1.65.

Ak by sme robili n testov signifikantnosti, pravdepodobnosť, že test s najväčšou hodnotou t štatistiky presiahne hodnotu 2 je oveľa väčšia ako 0.05. Z toho vyplýva, že je nutné použiť inú kritickú hodnotu pre test založený na maxime viacerých testov. My použijeme techniku *Bonferroniho korekcie* k výpočtu kritických hodnôt [2]. Táto technika tvrdí, že pre n testov, každý na hladine α, pravdepodobnosť chyby prvého druhu nie je väčšia ako $n\alpha$. V našom prípade pre $n = 50$ a $\alpha = 0.05$ by bola chyba 1. druhu väčšia ako 1. Ak ale zvolíme kritickú hodnotu $(\frac{\alpha}{n}) \times 100\%$ nebude hladina signifikantnosti väčšia ako $n(\frac{\alpha}{n}) = \alpha$.

1.3.2 DAF1 - Vplyvné pozorovania

Následne naprogramujeme data analytickú funkciu, ktorá zistí, aké pozorovania výrazne ovplyvňujú výslednú podobu regresného modelu. Takéto pozorovania voláme *vplyvné pozorovania*. Pozorovania budú považované za vplyvné, ak po vymazaní, spôsobia výraznu zmenu v odhadnutej regresnej funkcii. Základom je vypočítať Cookovú vzdialenosť. Táto štatistika zachycuje vplyv vypustenia i-tého pozorovania na zmenu odhadu regresných parametrov. Je zadefinovaná podľa [2] ako

$$D_i = \frac{(\boldsymbol{b}_{(-i)} - \boldsymbol{b})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{b}_{(-i)} - \boldsymbol{b})}{p\, s^2(e)}, \tag{1.67}$$

kde $\boldsymbol{b}_{(-i)}$ predstavuje vektor odhadu regresných parametrov pri vypustení i-tého pozorovania, $\boldsymbol{b}$ je odhad regresných parametrov celého datového súboru, p je počet

parametrov odhadovaného regresného modelu, $\mathbf{X}$ je regresná matica a $s(e)$ je reziduálna smerodatná odchylka. Na náš výpočet Cookovej vzdialenosti použijeme však iný vzorec. Dá sa ukázať, že vzorec 1.67 môžeme prepísať ako

$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}}, \tag{1.68}$$

kde

$$r_i = \frac{e_i}{s(e)\sqrt{1 - h_{ii}}}. \tag{1.69}$$

Vo finále máme teda vzorec, podľa ktorého vypočítame Cookovu vzdialenosť.

$$D_i = \frac{e_i^2 \, h_{ii}}{p(1 - h_{ii})^2 s^2(e)} \tag{1.70}$$

Pozorovanie, pre ktoré je D_i veľké, respektíve väčšie ako 1, budeme označovať ako vplyvné pozorovanie. Znamená to, že tieto pozorovania majú značný vplyv na odhad regresných parametrov a tým pádom aj na vyrovnané hodnoty. Odstránenie týchto hodnôt môže mať výrazný dopad na výsledný regresný model. V našej analýze odľahlé a vplyvné pozorovania v rámci data analytickej funkcii vymažeme a odhadneme regresný model z dátového súboru, ktorý neobsahuje tieto pozorovania.[2]

1.3.3 DAF2 - Transformácia vysvetľovanej premennej

Pri hľadaní vhodného modelu je niekedy pre dosiahnutie približnej lineárnej závislosti, homogenného rozptylu alebo približne normálneho rozdelenia chybovej zložky vhodné transformovať vysvetľované premenné prípadne vysvetľujúce premenné. Nehomogénne dáta sú také, ktoré nemajú rovnaký rozptyl pre všetky hodnoty vysvetľujúcich premenných. V prípade heteroskedasticity[3], odhady získané metódou najmenších štvorcov už nie sú najlepšie nestranné lineárne odhady, pretože nie sú splnené predpoklady Gauss-Markovova teorému 1.2.2. Takisto nie sú splnené predpoklady pre testovanie významnosti regresných koeficientov pomocou *t-testu* a *F-testu* a odhady smerodatných odchyliek sú nepresné [5]. V našej práci budeme transformovať jedine vysvetľované premenné za účelom zaistenia normálneho rozdelenia chybovej zložky. Budeme vyberať z 3 základných transformácii vysvetľovanej premennej y a to

- $\log y$,
- $\sqrt{y}$,
- identita.

Vyberáme vhodnú transformáciu na základe šikmosti reziduí. Model s najmenšou šikmosťou reziduí určuje vhodnú transformáciu vysvetľovanej premennej. Pri použití transformácie na vysvetľovanú premennú a následnom odhade regresnej funkcie v určitom bode, prípadne odhadu regresného koeficientu, bude nutné tieto hodnoty spätne transformovať za účelom možného porovnania týchto hodnôt.

[2] V R-ku máme naprogramovanú funkciu, ktorá odstráni vplyvné pozorovania z dátového súboru, je spojená s funkciou, ktorá odstraňuje odlahlé pozorovania a volá sa $outlier_fun$.

[3] Reziduálne zložky nemajú konštantný rozptyl.

1.3.4 DAF3 - Výber premenných

Na získanie správneho modelu je nutné určiť, ktoré vysvetľujúce premenné sú významné, a ktoré nie. Existujú objektívne kritéria a automatické procedúry pre výber modelu(predovšetkým pre stanovenie optimálneho výberu a počtu vysvetľujúcich premenných), kde každý zvolený výber je možné kvantitatívne ohodnotiť a hlavne objektívne porovnať s iným výberom [6]. Jedna z najznámejších kritérii sa okamžite ponúka *Akaikeho Informačné kritérium*, v skratke *AIC*.

Cieľom výberu premenných je rozdeliť regresnú maticu $\boldsymbol{X}$ na množinu aktívnych premenných $\boldsymbol{X}_A$ a množinu neaktívnych premenných $\boldsymbol{X}_N$. Predpokladajme, že máme určitú podmožinu potencionálnych kandidátov na premenné do modelu $\boldsymbol{X}_C$. *AIC* je dané vzorcom z [2].

$$AIC = n\,log(\frac{RSS_C}{n}) + p_C, \tag{1.71}$$

kde n je počet dát a p_C je počet parametrov v modeli. RSS_C je reziduálny súčet štvorcov podľa vzorca 1.30. Hľadáme model s najnižšou hodnotou AIC. S rastúcim počtom parametrov p_C nám prvý člen 1.71 klesá a druhý člen stúpa. Minimalizujeme *AIC*, teda hľadáme model, ktorý má nízku hodnotu RSS_C a zároveň nízky počet parametrov p_C. V našej práci používame spätne eliminačnú metódu. V prípade, že táto metóda zlyhá použijeme metódu postupného výberu premenných.

Spätne eliminačná metóda:

1. Začneme s modelom, ktorý obsahuje všetky premenné vrátane absolútneho členu.

2. Postupne odoberieme zvlášť každý člen a pre každý takto získaný model vypočítame hodnotu *AIC*. Pokračujeme s modelom, ktorý po odstránení premennej má najnižšiu hodnotu *AIC*.

3. Opakujeme bod 2 dokým po odstránení hociktorej premennej z modelu sa hodnota *AIC* už iJba zväčší, alebo dokým v modeli nezostane iba absolútny člen.

Metóda postupného výberu premenných:

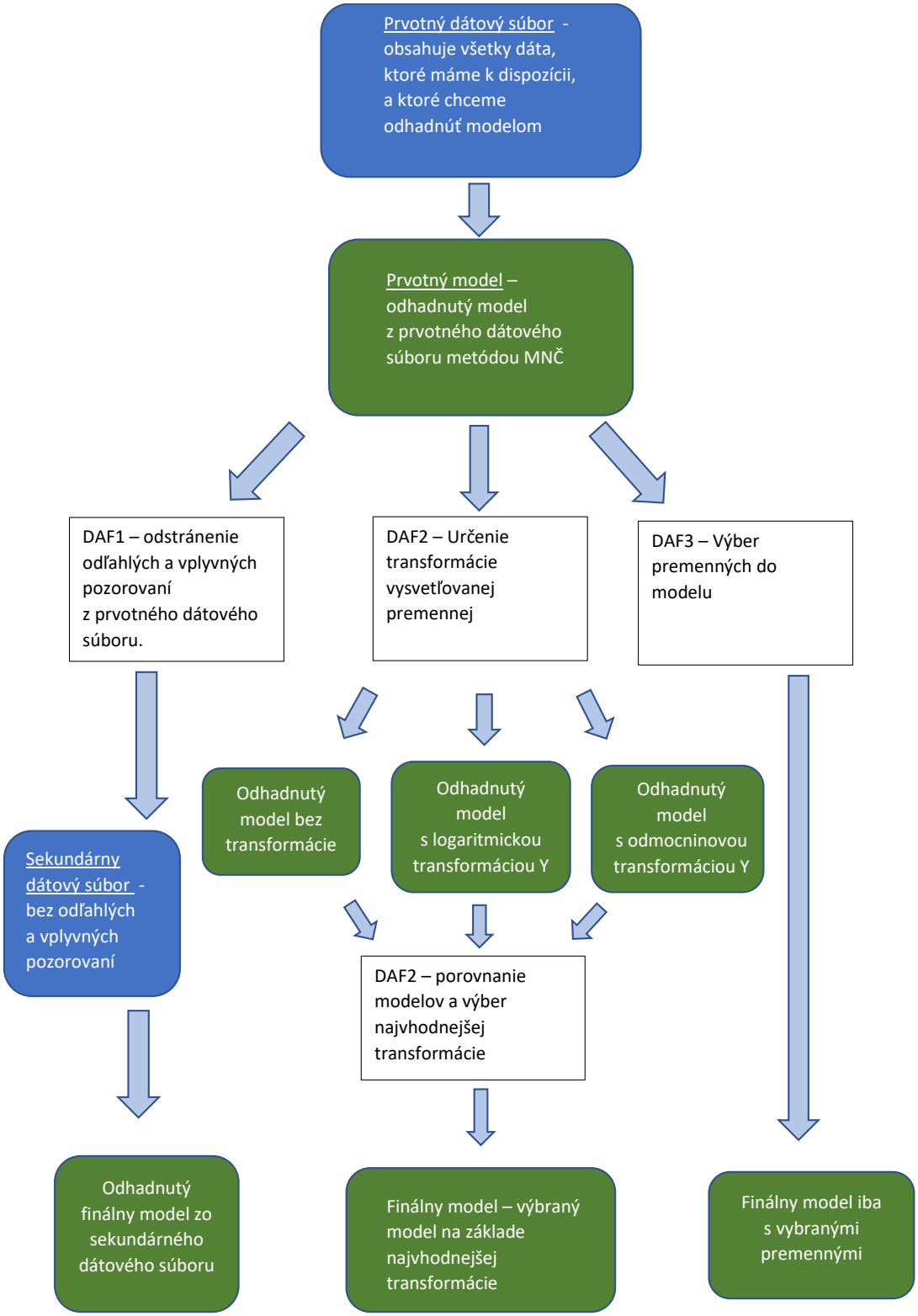
1. Začneme bez akýchkoľvek vysvetľujúcich premenných v modeli. Model obsahuje iba absolútny člen.

2. Vyberáme vysvetľujúce premenné na základe hodnoty Akaikeho informačného kritéria tak, že pridáme vysvetľujúcu premennú, ktorá najviac zníži hodnotu AIC kritéria.

3. Takto postupujeme až dovtedy kým sa ďalším pridaním vysvetľujúcej premennej do modelu už nezníži hodnota AIC, alebo dokým nepridáme všetky vysvetľujúce premenné do modelu.

Tento postup je pomerne šikovne implemetovaný v balíčku *olsrr* v programe R, preto ho nebudeme samostatne programovať.[4] Po úspešnom aplikovaní regresný analytických funkcí získame finálny model metódou najmenších štvorcov.

[4]V balíčku *olsrr* na spätne eliminačnú metódu používame funkciu *ols_stepaic_backward*, na metóda postupného výberu premenných používame funkciu *ols_stepaic_forward*.



Obr. 1.1: Schéma použitia jednotlivých data analytických funkcí(DAF) pri výstavbe regresného modelu.

Kapitola 2

Bootstrap

2.1 Základná myšlienka metódy bootstrap

Metóda bootstrap patrí medzi intenzívne počítačové metódy pre štatistickú analýzu dát. Predovšetkým sa používa ku zkúmaniu vlastností odhadov parametrov. Umožňuje zistiť rôzne štatistiky výberových odhadov(vychýlenie, rozptyl, konfidenčné intervaly, chybu predikcie a iné) [1]. Prvý článok o bootstrape [4] vydal americký štatistik Bradley Efron vďaka, ktorému sa stal známy. Po vydaní tohto článku nasledovala celá rada teoreticky i prakticky rozširujúcich článkov [1][11][7] a iné, ktoré skúmali vlastnosti a použiteľnosť tejto metódy. My sa v následujúcich riadkoch zameriame na stručnú charakteristiku metódy bootstrap a vysvetlíme ako pomocou bootstrapu zlepšíme odhad strednej štvorcovej chybu odhadu.

Uvažujme nezávislé rovnako rozdelené náhodné veličiny X_1,...,X_n, ktorých distribučná funkcia F nie je bližšie špecifikovaná. Označme $\theta = \theta(F)$ ako určitú charakteristiku(parameter) rozdelenia. θ chceme odhadnúť za pomocou náhodného výberu. Nech $T_n = T_n(X_1,...,X_n)$ je štatistika pre odhad parametru θ a nech $R_n = R_n(X_1,...,X_n)$ je ich vhodne štandardizovaná funkcia, napríklad $R_n = \sqrt{n}(T_n - \theta)$. Označme

$$H_n(x) = P[R_n(X_1,...,X_n,F) \leq x] \tag{2.1}$$

ako distribučnú funkciu štatistiky R_n. V prípade, že je distribučná funkcia F známa vieme postupovať metódou Monte Carlo a generovať zo známej distribučnej funkcie sériu nezávislých náhodných výberov - bootstrapových vzoriek. Ďalej pre každú realizáciu náhodného výberu spočítam hodnotu príslušnej štatistiky a jej skutočné rozdelenie aproximujeme empirickým rozdelením získaným pomocou náhodných výberov.

Ak je skutočná distribučná funkcia F neznáma, čo sa stáva oveľa častejšie, je možné aproximovať H_n asymptotickým rozdelením, ktoré sa dá odvodiť na základe limitných viet teórie pravdepodobnosti. Presnosť takejto aproximácie je avšak ovplyvnená a obmedzená počtom pozorovaní, ktoré sú skutočne k dispozícii[8]. Metóda *bootstrap* navrhuje iné riešenie. Táto metóda kombinuje substitučný princíp a metódu *Monte Carlo*.

Najprv priblížime substitučný princíp. Označme si $F_n(x)$ odhad distribučnej funkcie. Zvyčajne sa jedná o empirickú distribučnú funkciu založenú na náhodnom výbere X_1,...,X_n, teda

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x], \tag{2.2}$$

kde n je počet výberov a $I[A]$ značí indikátor množiny A. Nech X_1^*,...,X_n^* je nezávislý náhodný výber z F_n, čo znamená, že pri daných pozorovaniach X_1,...,X_n sú X_1^*,...,X_n^* nezávislé, rovnako rozdelené náhodné veličiny, z ktorých každá nadobúda hodnoty X_1,...,X_n s pravdepodobnosťou $\frac{1}{n}$. Súbor X_1^*,...,X_n^* nazývame *bootstrapový výber*, prípadne bootstrapová vzorka. Pôvodný výber nahradíme bootstrapovým výberom a neznámu distribučnou funkciou F nahradíme známou distribučnou funkciou F_n. Dostaneme parameter $\theta^* = \theta(F_n)$ a štatistiky $T_n^* = T_n(X_1^*,...,X_n^*)$ a $R_n^* = R_n(X_1^*,...,X_n^*,F_n)$. Najčastejšie sa na bootstrapový výber X_1^*,...,X_n^* a známu empirickú distribučnú funkciu F_n aplikuje metóda Monte Carlo, kedy sa veľakrát(v rádoch tisícoch až stotisícoch) generuje nezávislý náhodný výber z rozdelenia F_n, pri každom opakovaní sa spočítajú hodnoty T_n^*, R_n* a z nich sa stanoví aritmetický priemer prípadne iná požadovaná štatistika. Ak chceme zistiť distribúciu komplexných odhadov, je možné použiť metódu bootstrap.

2.2 Úloha metódy bootstrap v našej analýze

Hlavnou úlohou našej analýzy je posúdiť vplyv použitých analytických procedúr na presnosť odhadu $RMSE$ odhadu regresnej funkcie v určitom bode a presnosť odhadu $RMSE$ odhadu regresných parametrov modelu. Naivný odhad $RMSE$ je odhadovaný pomocou klasického odhadu smerodatnej chyby z metódy najmenších štvorcov, viď. 1.61 , neberie do úvahy použité data analytické funkcie. Ukážeme ako naivný odhad $RMSE$ v kontexte použitých data analytických funkcii zlyháva a predvedieme ako dokáže bootstrap zlepšiť tento odhad. V prípade dát, ktoré máme k dispozícii, máme na výber z dvoch metód získania bootstrapovho výberu: Preusporiadať reziduá, ktoré sú závislé na modeli alebo preusporiadať dáta podľa riadkov, čož je nezávislé na modeli [10]. Mohli by sme spraviť kompletnú regresnú analýzu, aby sme získali model a následne prevzorkovať reziduá tohto modelu čím vygenerujeme bootstrapov výber. Keby sme robili výber z reziduí znamenalo by to, že implicitne predpokladáme, že prvotný model je správny [10]. My však nechceme predpokladať akúkoľvek správnosť modelu. Preto zvolíme nepodmienené vzorkovanie cez riadky (X_i,Y_i). Výhoda tohto postupu je jednoduchosť, pretože nemusíme pred analýzou definovať akýkoľvek model. Nevýhoda tohto postupu je, že bootstrapové odhady rozptylu sú vychýlené [10].

Náš postup spočíva vo výbere z pôvodných dát s vracaním a následne aplikovanie regresnej analýzy na každý takýto výber, tak ako to je opísané v predchádzajúcej kapitole. Teda aplikujeme regresné analytické funkcíe nielen na pôvodné dáta ale aj na dáta získané bootstrapovým výberom z pôvodných dát. Keďže naše pôvodné dáta pochádzajú z modelu so známou teoretickou distribučnou funkciou, vieme použiť štatistiky, ktoré nám priblížia kvalitu odhadu $RMSE$ odhadu regresnej funkcie v určitom bode a odhad $RMSE$ odhadu regresných parametrov.

Na porovnanie kvality odhadu smerodatnej odchylky odhadu regresnej funkcie v určitom bode $\hat{\boldsymbol{\eta}}_*$ a odhadu smerodatnej odchylky odhadu regresných koeficien-

tov **b** použijeme *odmocninu zo strednej štvorcovej chyby*($RMSE$). Definujeme ďalej štatistiku $RMSE$. Hodnota $RMSE$ pre estimátor $\hat{\theta}$ je daná ako

$$RMSE(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}, \tag{2.3}$$

kde $\hat{\theta}$ je odhad teoretickej hodnoty θ. Na každý bootstrapový výber aplikujeme data analytické funkcie a pomocou odhadu štatistiky $RMSE$ sa pokúsime posúdiť vychýlenie klasického odhadu štandardnej chyby odhadu regresnej funkcie v bode a odhadu regresných parametrov. Klasický(naivný) odhad RMSE budeme porovnávať s hodnotou RMSE získanou pomocou metódy Monte Carlo, viď 3.3, a takisto s RMSE získaným pomocou metódy bootstrap, viď 3.4.

2.2.1 Odhad regresnej funkcie v bode

Predpokladajme, že chceme rozdelenie pre odhad regresnej funkcie v bode $\boldsymbol{x}_0$. Pre model m je táto hodnota rovná $\hat{\eta}(m_{\boldsymbol{x}_0})$. Priblížime si náš celkový postup pozmenený od článku [10].

1. Najprv na dátach urobíme analýzu vo forme presne špecifikovaných data analytických funkcí definovaných v kapitole 2. Označíme jednotlivé aplikované data analytické funkcie ako a_1,a_2,a_3. Potom odhad regresnej funkcie v bode $\boldsymbol{x}_0$ označíme po aplikovaní jednotlivých DAF ako

$$\hat{\eta}_1 = \hat{\eta}((a_1(m_{initial}))_{\boldsymbol{x}_0}), \tag{2.4}$$

$$\hat{\eta}_2 = \hat{\eta}((a_2(m_{initial}))_{\boldsymbol{x}_0}), \tag{2.5}$$

$$\hat{\eta}_3 = \hat{\eta}((a_3(m_{initial}))_{\boldsymbol{x}_0}), \tag{2.6}$$

   kde $m_{initial}$ je prvotný model, ktorý získam aplikovaním metódy najmenších štvorcov na prvotný dátový súbor zahrnutím všetkých premenných.

2. Následne vygenerujeme bootstrapové vzorky (X_i^{boot},Y_i^{boot}) vzorkovaním z originálnych dát a vytvoríme $m_{initial}^{boot}$. Teraz aplikujeme rovnako ako v bode 1 data analytické funkcie. Dostaneme model, ktorý sa môže líšiť vo veľkosti parametrov, počtu prediktorov alebo použitých transformácii. Bootstrapový odhad regresnej funkcie v bode $\boldsymbol{x}_0$ po aplikovaní jednotlivých DAF označíme

$$\hat{\eta}_1^{boot} = \hat{\eta}((a_1(m_{initial}^{boot}))_{\boldsymbol{x}_0}), \tag{2.7}$$

$$\hat{\eta}_2^{boot} = \hat{\eta}((a_2(m_{initial}^{boot}))_{\boldsymbol{x}_0}), \tag{2.8}$$

$$\hat{\eta}_3^{boot} = \hat{\eta}((a_3(m_{initial}^{boot}))_{\boldsymbol{x}_0}), \tag{2.9}$$

3. Mnohonásobným opakovaním vytvoríme rozdelenie bootstrapových odhadov regresnej funkcie $\hat{\eta}_1^{boot}$, $\hat{\eta}_2^{boot}$, $\hat{\eta}_3^{boot}$ v danom bode $\boldsymbol{x_0}$.

Je nutné dodať, že odhad regresnej funkcie je výsledok použitia data analytických funkcí v bode $\boldsymbol{x_0}$. Dôležité je ale urobiť bootstrapový výber z originálnych dát a nie na datách, na ktoré už boli aplikované regresné funkcie, pretože inak sa môže stať, že bootstrapové dáta nezobrazia celkový vplyv regresnej analýzy. Pre jednotlivé vygenerované bootstrapové dátové súbory odhadneme metódou najmenších štvorcov prvotný lineárny model a použijem data analytické funkcie k vytvoreniu finálnych modelov.

2.2.2 Odhad regresného parametru β_1

Postupujeme podobne ako v 2.2.1 s tým, že predpokladáme, že chceme rozdelenie pre odhad regresného parametru β_1 po aplikovaní jednotlivých DAF osobitne.

1. Odhad regresného parametru β_1 označíme pre jednotlivé DAF ako

$$b_{1_1} = b_1((a_1(m_{initial}))), \tag{2.10}$$

$$b_{1_2} = b_1((a_2(m_{initial}))), \tag{2.11}$$

$$b_{1_3} = b_1((a_3(m_{initial}))), \tag{2.12}$$

2. Znova vygenerujeme bootstrapové vzorky tak isto ako v 2.2.1. Bootstrapové odhady regresného parametru β_1 po aplikovaní jednotlivých DAF označíme

$$b_{1_1}^{boot} = b_1(a_1(m_{initial}^{boot})), \tag{2.13}$$

$$b_{1_2}^{boot} = b_1(a_2(m_{initial}^{boot})), \tag{2.14}$$

$$b_{1_3}^{boot} = b_1(a_3(m_{initial}^{boot})), \tag{2.15}$$

3. Mnohonásobným opakovaním rovnako ako v 2.2.1 vytvoríme rozdelenie bootstrapových odhadov regresného parametru β_1 a to $b_{1_1}^{boot}$, $b_{1_2}^{boot}$, $b_{1_3}^{boot}$.

Kapitola 3

Simulácie

Našim cieľom v tejto kapitole je posúdiť funkčnosť metódy prevzatej z [10] pri vygenerovaných dátach podľa nami zvoleného teoretického modelu. Pripomíname, že si definujeme všeobecný model z druhej kapitoly 1.62

$$g(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(x_{ij}) + w_i \epsilon_i \quad i = 1,...,n. \tag{3.1}$$

kde $(x_{i1},...,x_{ip})$ sú hodnoty vysvetľovaných premenných $X_1,...,X_p$, kde p je počet premenných. Všetky premenné sú nekorelované a majú distribučnú funkciu štandardné normálne rozdelenie $F_x \sim N(0,1)$. Ďalej ϵ_i má takisto distribúčnú funkciu štandardné normálne rozdelenie $F_\epsilon \sim N(0,1)$ a regresné koeficienty ak nie sú definované inak sú $\beta_0 = 0$ a $\beta_i = 1, i \neq 0$, kde n je počet pozorovaní. g symbolizuje funkciu aplikovanú na hodnoty vysvetľovanej premennej. f_j symbolizuje j-tú funkciu aplikovanú na hodnoty vysvetľujúcej premennej X_j, teda funkciu aplikovanú na hodnoty $x_{1j},x_{2j}...,x_{nj}$. Bod na odhad regresnej funkcie zvolíme $\boldsymbol{x}_0 = (0.2,0.2,...,0.2)$. Naše východzie dátové súbory vygenerujeme metódou Monte Carlo z rôznych variánt modelu 3.1, ktoré si zadefinujeme v tabuľke 3.1. Každý z 5 modelov je niečím špecifický, vanila model, viď. 3.1.1, obsahuje iba predvolené hodnoty, ktoré sú $n = 50, p = 5, g$ a f_j sú funkcie identity pre $j = 1,...,p$, F_x a F_ϵ majú štandardné normálne rozdelenie $N(0,1)$ a váhy $w_i = 1$ pre $i = 1,...,n$. Nelineárny model, viď. 3.1.2, má dáta generované z rovnomerného rozdelenia U v intervale (0,0.2) a funkcia g respektíve g^{-1} nie je tentokrát identita. Hetero model, viď. 3.1.3, ako už názov napovedá, má heteroskedastický rozptyl spôsobený váhami $w_i = \sum_{j=1}^{p} \beta_j f_j(x_{ij})$ chybovej zložky. Náhodné veličiny X_j pre $j = 1,...,5$ v prípade hetero modelu majú distribučnú funkciu z rovnomerného rozdelenia z intervalu (0,0.2). Kolineárny model, viď. 3.1.4, obsahuje iba 3 vysvetľujúce premenné, ktoré sú lineárne závislé. A nakoniec cieľom nasýteného modelu, viď. 3.1.5, je zistiť presnosť jednotlivých odhadov $RMSE$ v prípade, že máme k dispozícii nízky počet pozorovaní $n = 30$ a vysoký počet premenných v modeli.

Metódou *Monte Carlo* postupujeme tak, že pre každý model vygenerujeme 500 datasetov, každý o veľkosti $n = 50$(u nasýteného modelu je výnimka $n = 30$). Počet simulácii rovný 500 sme zvolili kvôli výpočtovej náročnosti celého procesu viď. sekciu 3.4. Pre každý dataset odhadneme prvotný model pomocou metódy najmenších štvorcov. Vyzerá následovne

Názov modelu	Hodnoty
Vanila	Predvolené hodnoty
Nelineárny	$F_x \sim U(0,0.2)$, $g^{-1}(x) = e^{\frac{x}{5}}$
Hetero	$F_x \sim U(0,0.2), w_i = \sum_{j=1}^{p} \beta_j f_j(x_{ij})$
Kolineár	$X_1 \sim N(0,1), X_2, X_3 \sim N(0,0.01) - X_1$
Nasýtený	$n = 30, p = 15$ a $n = 50, p = 15$

Tabuľka 3.1: Modely použité na generovanie dát.

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad i = 1,...,n. \tag{3.2}$$

Prvotný model 3.2 je celkom lineárny, bez transformácii premenných, s jednotkovými váhami $w_i = 1$ pre $i = 1,...,n$ a obsahuje všetkých p vysvetľujúcich premenných. Určím si, ktorý model budem simulovať a akú DAF použijem. Potom na každý dataset aplikujeme túto data analytickú funkciu zadefinovanú v kapitole 1 a následne odhadnem finálny model. Ďalej vypočítame požadovaný odhad z finálného modelu(v našom prípade to bude buď odhad regresnej funkcie v bode $\boldsymbol{x}_0 = (0.2,0.2,...,0.2)$ alebo odhad regresného parametru β_1) a označíme ho θ_i. Vieme z akého modelu boli generované dáta a preto vieme určiť aká by mala byť skutočná hodnota, túto skutočnú hodnotu označíme θ.

Odhad RMSE teda odhad odmocniny zo strednej štvorcovej chyby odhadu na základe metódy Monte Carlo vypočítame nasledovne:

$$\widehat{RMSE}^{MC}(\hat{\theta}) = \sqrt{\frac{1}{500} \sum_{i=1}^{500} (\theta_i - \theta)^2}, \tag{3.3}$$

kde $\hat{\theta}$ je odhad θ z primárného dátového súboru. Ďalej v analýze budeme odhad $RMSE$ na základe Monte Carlo metódy považovať za presný odhad skutočného $RMSE$. V kapitole 1 sme ukázali rozklad MSE estimátora 1.51, v našom prípade sa MSE odhadu regresnej funkcie skladá z dvoch časti a to z kvadrátu vychýlenia odhadu regresnej funkcie a rozptylu odhadu regresnej funkcie v bode $\boldsymbol{x}_0$. Odhad rozptyl odhadu regresnej funkcie môžeme takisto získať z Monte Carlo simulácii, tak že po vygenerovaní jednotlivých dátových súborov a odhadnutí príslušného modelu si uložíme hodnotu odhadu regresnej funkcie v bode $\boldsymbol{x}_0$. Tento proces treba opakovať pre každý dátový súbor získaný pomocou Monte Carlo, dostaneme tým rozdelenie odhadov regresnej funkcie v bode $\boldsymbol{x}_0$. Keby sme vypočítali výberový rozptyl tohto rozdelenia dostaneme odhad rozptylu odhadu regresnej funkcie. Vychýlenie nie je potreba počítať, pretože pri odhade $RMSE$ a odhade rozptylu si viem podľa vzorca 1.51 odvodiť aj kvadrát odhadu vychýlenia odhadu regresnej funkcie. Tento rozklad MSE sa dá použiť na to, aby sme v prípade nesúladu $RMSE$ z metódy Monte Carlo a z bootstrapovej metódy bolo možné posúdiť do akej miery sa na tomto nesúlade podieľa nesúlad v rozptyloch.

Pre každý z 500 dátových súborov vieme odhadnúť $RMSE(\hat{\theta})$ pomocou bootstrapu a to tak, že z daného datasetu opakovaným výberom s vracaním vygenerujeme 100 bootstrapových datasetov, znova o veľkosti $n = 50$(prípadne $n = 30$ pre nasýtený model). Pre jednotlivé bootstrapové vzorky odhadneme metódou

najmenších štvorcov prvotný model a použijeme vybranú data analytickú funkciu k vytvoreniu finálneho modelu.[1] Pre každý zo 100 bootstrapových datasetov získame bootstrapový odhad $RMSE$ odhadu podľa vzťahu

$$\widehat{RMSE}^{boot}(\hat{\theta}) = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\theta_i^{boot} - \hat{\theta})^2}, \tag{3.4}$$

kde θ_i^{boot} je odhad (v našom prípade regresnej funkcie v bode $\boldsymbol{x}_0$ alebo regresného parametru β_1) pre model vytvorený na základe i-tého bootstrapového výberu. Pre porovnanie jednotlivých odhadov $RMSE$ získaných rôznymi spôsobmi je nutné nájsť nejakú spoľahlivú metriku. Priemer hodnôt v rozdelení sa hneď ponúka ako možnosť, ale môže sa stať, že rozdelenie bude obsahovať dostatočný počet vzdialených pozorovaní, ktoré silno ovplyvnia priemer. Preto sme vyhodnocovali rozdelenie v bode mediánu. Teda z rozdelenia 500 hodnôt $\widehat{RMSE}^{boot}(\hat{\theta})$ získaných z jednotlivých bootstrapových výberov vyberieme medián týchto hodnôt.

Na porovnanie klasickej metódy vypočítame naivný odhad $RMSE(\hat{\theta})$ získaný ako odhad smerodatnej odchylky odhadu regresnej funkcie(označíme $s(\hat{\eta}_{\boldsymbol{x}_0})$), pre všetky finálne modely získané z 500 vygenerovaných dátových súborov

$$\widehat{RMSE}(\hat{\eta}_{\boldsymbol{x}_0})^{naive} = s(\hat{\eta}_{\boldsymbol{x}_0}) = s(e)\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}, \tag{3.5}$$

kde $s(e)$ je reziduálna smerodatná odchylka, $\boldsymbol{X}$ je regresná matica, bod $\boldsymbol{x}_0 =$ (0.2,0.2,...,0.2). Všetky tieto hodnoty sú z finálneho modelu vybraného na základe použitej data analytickej funkcie. Hodnotu $\widehat{RMSE}(\hat{\theta})^{naive}$ vypočítame pre každý z 500 datasetov vygenerovaných metódou *Monte Carlo* podľa vzorca 3.5. Získame rozdelenie o 500 hodnotách, ktoré vyhodnotíme v bode mediánu. Nesmieme zabudnúť, že pri aplikovaní transformácie vysvetľovanej premennej musíme, kvôli korektnému porovnaniu naivného odhadu $\widehat{RMSE}(\hat{\theta})^{naive}$ a bootstrapového odhadu $\widehat{RMSE}^{boot}(\hat{\theta})$, transformovať naivný estimátor. V prípade, že na vysvetľovanú premennú aplikujeme funkciu g tak ako je zobrazené v modeli 3.1, bude výsledný naivný estimátor RMSE rovný

$$\widehat{RMSE}(\hat{\eta}_{\boldsymbol{x}_0})^{naive} = (g^{-1})'(\hat{\eta}_{\boldsymbol{x}_0})s(\hat{\eta}_{\boldsymbol{x}_0}), \tag{3.6}$$

kde $(g^{-1})'$ je derivácia zinvertovanej funkcie g v bode $\boldsymbol{x}_0 =$ (0.2,0.2,...,0.2). Získame tým rozdelenie hodnôt naivných odhadov $RMSE$. Ako výslednú hodnotu vyberieme medián z tohto rozdelenia naivných estimátorov. V prípade, že je použitá transformácia odmocnina na vysvetľovanú premennú, výsledná hodnota naivného odhadu $RMSE$ bude

$$\widehat{RMSE}(\hat{\eta}_{\boldsymbol{x}_0})^{naive} = 2\hat{\eta}_{\boldsymbol{x}_0}s(\hat{\eta}_{\boldsymbol{x}_0}). \tag{3.7}$$

V prípade, že je použitá transformácia logaritmus, výsledná hodnota naivného odhadu $RMSE$ bude

$$\widehat{RMSE}(\hat{\eta}_{\boldsymbol{x}_0})^{naive} = \exp(\hat{\eta}_{\boldsymbol{x}_0})s(\hat{\eta}_{\boldsymbol{x}_0}). \tag{3.8}$$

[1] Z jednej bootstrapovej vzorky získam 3 finálne modely, pretože aplikujem 3 data analytické funkcie zakaždým na pôvodnú bootstrapovú vzorku a následne odhadnem finálny model.

Ak nás zaujíma naivný odhad $RMSE$ odhadu regresného parametru, ten zistíme zo vzorca 1.46 tak, že zoberieme príslušný diagonálny prvok z matice 1.46. Teda pre b_1 platí

$$\widehat{RMSE}(b_1)^{naive} = s(b_1) = \left(s(e)\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}\right)_{[2,2]}, \tag{3.9}$$

kde index [2,2] značí, že sa jedná o druhý diagonálny prvok matice 1.46. Ďalej $s(b_1)$ je odhad smerodatnej odchylky b_1. V prípade, že je použitá transformácia odmocnina, výsledná hodnota naivného odhadu $RMSE$ bude

$$\widehat{RMSE}(b_1)^{naive} = b_1 \exp(b_0 + 0.2(b_1 + b_2 + b_3 + b_4 + b_5))s(b_1). \tag{3.10}$$

‘

V prípade, že je použitá transformácia logaritmus, výsledná hodnota naivného odhadu $RMSE$ bude

$$\widehat{RMSE}(b_1)^{naive} = 2b_1((b_0 + 0.2(b_1 + b_2 + b_3 + b_4 + b_5))s(b_1) \tag{3.11}$$

Stojí za to v krátkosti zhrnúť celý proces ešte raz. Zo zadaného teoretického modelu vygenerujeme 500 datasetov a pre každý dataset odhadneme prvotný regresný model. Na prvotný model aplikujeme jednu z DAF a odhadneme finálny model. Na základe finálného modelu odhadneme regresnú funkciu v bode $\boldsymbol{x}_0$, prípadne máme odhad regresného parametru β_1. Poznáme skutočnú hodnotu regresnej funkcie v bode $\boldsymbol{x}_0$, tým pádom podľa vzorca 3.3 vypočítame $\widehat{RMSE}^{MC}(\hat{\theta})$. Naivný estimátor $\widehat{RMSE}(\hat{\theta})^{naive}$ vypočítame ako odhad smerodatnej chyby odhadu pre každý z 500 datových súborov. Výsledné rozdelenie $\widehat{RMSE}(\hat{\theta})^{naive}$ vyhodnotím v bode mediánu. Pre každý z 500 dátových súborov metódou bootstrap vytvorím 100 bootstrapových vzoriek. Z každej bootstrapovej vzorky odhadnem prvotný model. A rovnako ako pri 500 vygenerovaných dátových súboroch aplikujeme jednu z DAF a odhadneme finálny bootstrapový model. Znova na základe finálného bootstrapového modelu odhadneme regresnú funkciu v bode $\boldsymbol{x}_0$, prípadne máme odhad regresného parametru β_1. Namiesto skutočnej hodnoty teraz použijeme na výpočet $\widehat{RMSE}^{boot}(\hat{\theta})$ odhadu(odhadu regresnej funkcie v bode $\boldsymbol{x}_0$ alebo odhadu regresného parametru β_1) z dátového súboru, z ktorého sa vzorkovali bootstrapové výbery. Z rozdelenia hodnôt $\widehat{RMSE}^{boot}(\hat{\theta})$ opäť vyberám medián.

3.1 Odhad regresnej funkcie v bode

Vyskúšame metódu či funguje na dostatočnom počte dát, $n = 50, n = 500, n = 5000$, *bez použitia akýchkoľvek data analytických funkciu.* Vygenerujeme 100, 1 000 a 10 000 datových súborov a takisto 100, 1 000 a 10 000 bootstrapových vzoriek pre každý vygenerovaný datový súbor pochádzajúci z teoretického modelu vanilla. Avšak pri počte simulácii 1 000 a 10 000 sme museli upustiť od 10 000 bootstrapových výberov kvôli výpočtovej náročnosti. Cieľom bude, okrem iného,

aj zistiť nakoľko záleži na počte pozorovaní n, počtu simulácii a počtu bootstapových výberov pri presnosti odhadu RMSE. Budeme sledovať výsledok Monte Carlo estimátora pri 100, 1 000 a 10 000 simuláciach, naivného estimátora a estimátora získaného pomocou bootstrapu. Ak nepoužijeme žiadné data analytické funkcie a pri dostatočnom počte Monte Carlo a bootstrap simulácii, očakávame, že budú výsledky estimátorov veľmi podobné. Dosiahli sme výsledky len s malou odchylkou, viď tabuľka 3.2.

	RMSE Estimátor									
	Monte Carlo			Naivný			Bootstrap			
	Počet dát									
Simul	50	500	5000	50	500	5000	**Boot**	50	500	5000
100	0.150	0.048	0.015	0.163	0.050	0.016	100	0.160	0.049	0.015
	0.158	0.053	0.017	0.163	0.049	0.016	1000	0.161	0.049	0.015
	0.170	0.049	0.015	0.158	0.049	0.015	10000	0.160	0.049	0.015
1000	0.164	0.051	0.015	0.162	0.049	0.015	100	0.163	0.049	0.015
	0.166	0.049	0.016	0.162	0.049	0.015	1000	0.162	0.049	0.015
10000	0.164	0.050	0.016	0.162	0.049	0.015	100	0.162	0.049	0.015
	0.164	0.051	0.169	0.162	0.049	0.158	1000	0.163	0.049	0.159

Tabuľka 3.2: Hodnoty odhadu $RMSE$ odhadu regresnej funkcie v bode $\boldsymbol{x}_0$ bez použitia data analytických funkcí.

Z výsledkov z tabuľky 3.2 vyplýva, že bez použitia data analytických funkcii sú hodnoty odhadu $RMSE$ získané z Monte Carlo simulácii, naivného estimátora a bootstrapových simulácii veľmi podobné. Z tabuľky takisto vidíme, že prirodzene s rastúcim počtom dát sa nám aj zmenšuje odhad RMSE. S rastúcim počtom dát sme totiž schopní lepšie odhadnúť skutočný teoretický model a znížiť tým pádom strednú štvorcovú chybu odhadu regresnej funkcie. Z tabuľky 3.2 usudzujeme, že u počtu simulácii 100 je prítomná variabilita odhadu RMSE estimátora u Monte Carlo. Pri počte simulácii 1 000 sa už aproximácia takmer stabilizovala. U bootstrapu je už pri 100 bootstrapových výberov odhad stabilný. Keďže budeme robiť veľký počet simulácii snažíme sa čo najviac ušetriť výpočetný výkon. Preto v ďalších výpočoch budeme používať 500 simulácii a 100 bootstrapových výberov pre každý vygenerovaný dataset. Počet dát zvolíme $n = 50$. Najskôr budeme odhadovať model z vygenerovaných dát bez použitia data analytických funkcí. Potom ten istý model budeme odhadovať za použitia jednotlivých data analytických funkcí osobitne. Označíme si 1. data analytickú funkciu ako DAF1 - táto DAF identifikuje a odstraňuje z dátového súboru vzdialené a vplyvné pozorovania. 2. DAF ako DAF2 - určí, či je vhodné transformovať odhadovaný model a ak áno následne transformuje vysvetľovanú premennú odhadovaného modelu. Nakoniec 3. DAF ako DAF3 - určí vhodné vysvetľujúce premenné, ktoré budú zakomponované vo finálnom modeli.

3.1.1 Vanila model

Začneme s analýzou vanila modelu s použitím data analytických funkcii. Vanila je jednoduchý model bez výrazných vzdialených pozorovaní. Je to celkom

	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	0.156	0.178	0.213	0.156
Naivný	0.161	0.145	0.142	0.161
Bootstrap	0.161	0.195	0.220	0.161

Tabuľka 3.3: Vanila model - Hodnoty odhadu $RMSE$ odhadu regresnej funkcie na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.

lineárny model, to znamená, že je lineárny ako v parametroch, tak z hľadiska vysvetľujúcich premenných. Vyzerá následovne

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i, \tag{3.12}$$

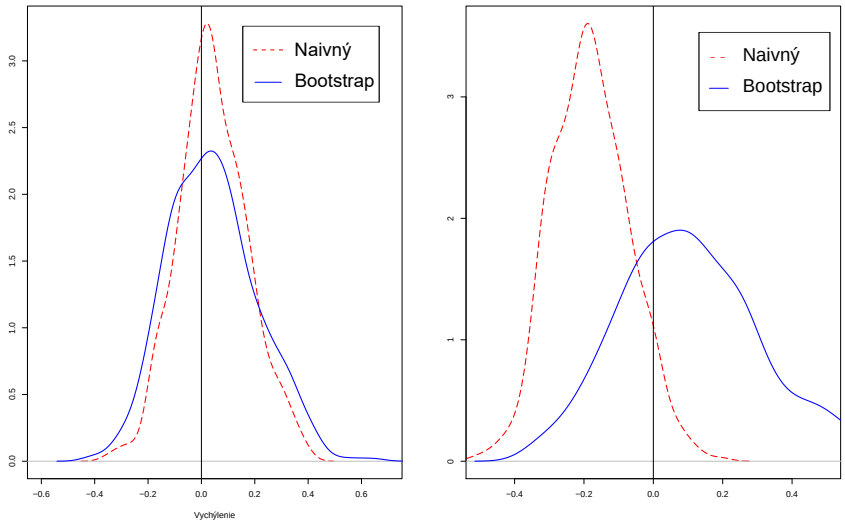
kde $x_{ij}, i = 1,...,50, j = 1,...,5$ sú hodnoty náhodných veličín X_j s normálnym rozdelením, nulovou strednou hodnotou a jednotkovým rozptylom $X_j \sim N(0,1)$, X_j sú nekorelované, preto by nemala byť aplikovaná žiadna z potencionálnych transformácii pri výstavbe regresného modelu. $\epsilon_i \sim N(0,1)$ a je IID(Independent and identically distributed), teda hodnoty ϵ_i sú nezávislé a rovnako rozdelené. Získanú hodnotu $\widehat{RMSE}(\hat{\theta})^{naive}$ a $\widehat{RMSE}(\hat{\theta})^{boot}$ normujeme, to znamená, že od výsledných hodnôt odčítame odhad RMSE získaný metódou Monte Carlo $\widehat{RMSE}^{MC}(\hat{\theta})$ a tento rozdiel takisto vydelíme hodnotou $\widehat{RMSE}^{MC}(\hat{\theta})$, túto znormovanú hodnotu zobrazíme ako rozdelenie graficky. Hodnota $\hat{\theta}$ je ľubovoľný estimátor, v našom prípade to je buď odhad regresnej funkcie v bode X_0(označíme $\hat{\eta}_{X_0}$), alebo odhad regresného koeficientu β_1(označíme b_1). V nasledujúcom grafe celé rozdelenie odhadov $RMSE$ odhadu regresnej funkcie v bode X_0 normujeme podľa hodnoty Monte Carlo odhadu a získavame Z^{Naive} a Z^{Boot} podľa vzorca

$$Z^{Naive} = \frac{\widehat{RMSE}^{Naive}(\hat{\eta}_{X_0}) - \widehat{RMSE}^{MC}(\hat{\eta}_{X_0})}{\widehat{RMSE}^{MC}(\hat{\eta}_{X_0})}, \tag{3.13}$$

$$Z^{Boot} = \frac{\widehat{RMSE}^{Boot}(\hat{\eta}_{X_0}) - \widehat{RMSE}^{MC}(\hat{\eta}_{X_0})}{\widehat{RMSE}^{MC}(\hat{\eta}_{X_0})}, \tag{3.14}$$

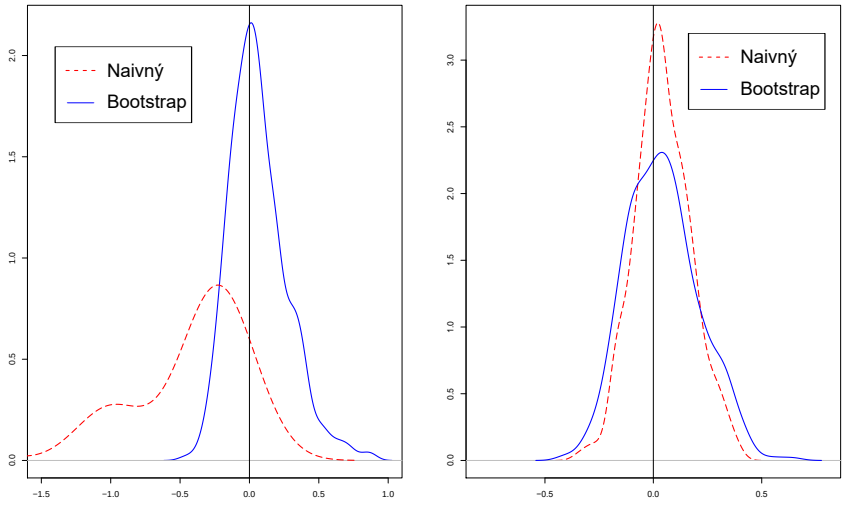
kde $\hat{\eta}_{X_0}$ je odhad regresnej funkcie v bode X_0. Vo výslednom obrázku 3.1 zobrazíme jadrový odhad hustoty Z^{Naive} a Z^{Boot}. Na obrázku 4.1a vidíme, že naivný a bootstrapový odhad skoro dokonale kopíruje hodnotu z MC simulácii. Je to niečo, čo by sme očakávali pri vanila modeli bez použitia akejkoľvek data analytickej funkcie. Ku každému modelu budeme zobrazovať tabuľku s konkrétnymi výsledkami simulácii pred aj po použití jednotlivých DAF samostatne. Hodnoty v nasledujúcej tabuľke 3.3 a ďalších tabuľkách budú vypočítane v bode mediánov rozdelenia odhadov, i keď pri niektorých dvojvrcholových rozdeleniach medián nemusí byť najlepšia štatistika.

Po jednotlivom aplikovaní data analytických funkcí vidíme, že naivný odhad $RMSE$ je podhodnotený pri odstraňovaní vzdialených a vplyvných pozorovaní



(a) Neboli použité DAF. (b) Použitá iba DAF1.

Obr. 3.1: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre vanila model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.2: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre vanila model s aplikovaním DAF2 a DAF3 samostatne.

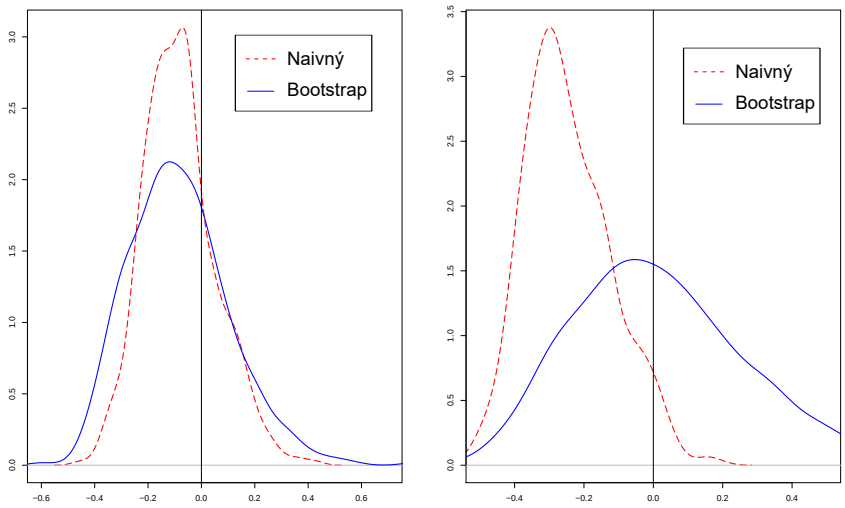
3.1b a pri transformácii vysvetľovanej premennej 3.2a. Naopak bootstrapový odhad $RMSE$ celkom dobre kopíruje hodnotu Monte Carlo odhadu $RMSE$ odhadu regresnej funkcie v bode aj po aplikovaní jednotlivých DAF, obzvlášť je lepší výsledok pri DAF1. Čo je trochu prekvapivé pretože očakávali by sme, že na získanie výsledného modelu nebudú jednotlivé data analytické funkcie pozmeňovať výrazne naše dáta a model. Po krátkej analýze výpočtu zisťujeme, že v 44% prípadov výstavby regresného modelu bola použitá transformácia odozvy - buď odmocnina alebo logaritmus. Pri použití DAF3 vidíme, že obrázky 3.1a a 3.2b sú skoro totožné. Z toho výplýva, že pri výbere premenných metóda spätnej eliminácie ponechala vo väčšine prípadov všetky premenné modelu.

3.1.2 Nelineárny model

Ďalej simulujeme nelineárny model, ktorý je v našej práci definovaný nasledovne

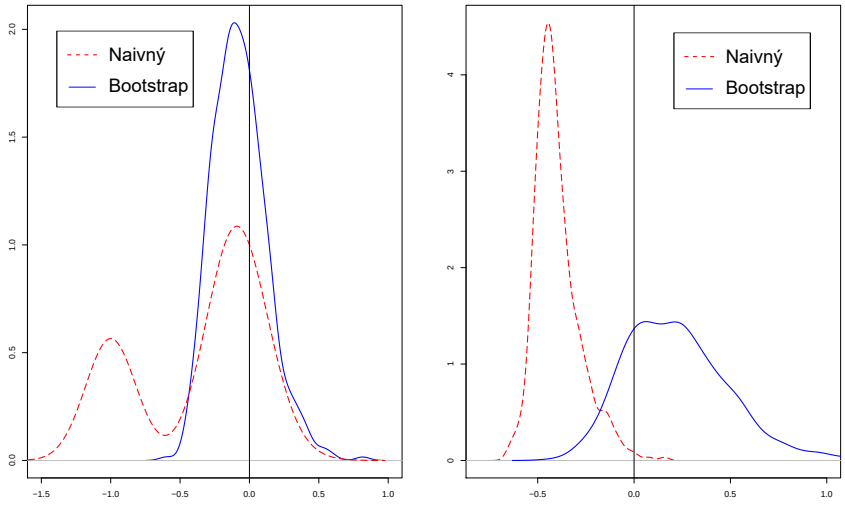
$$y_i = \exp\{(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})/5\} + \epsilon_i. \tag{3.15}$$

Dáta vygnerované z tohto modelu odhadneme lineárnym modelom pomocou metódy najmenších štvorcov. Očakávame, že keďže sa dáta pochádzajúce z nelineárného modelu snažíme odhadnúť lineárnym modelom 3.2, bude odhad $RMSE$ získaný Monte Carlo metódou výrazne odlišný ako naivný odhad a odhad $RMSE$ získaný metódou Bootstrap. Výsledky simulácii máme v tabuľke 3.4. Pri absencii DAF sa potvrdil náš úsudok nesúladu hodnôt z Monte Carlo metódy a naivného a bootstrapového odhadu. Naivný a bootstrapový odhad sú približne rovnaké, Monte Carlo odhad je výrazne vyšší. Na nasledujúcej tabuľke 3.4 si ukážeme aký vplyv na odhad $RMSE$ odhadu regresnej funkcie majú osobitne použité DAF. Teda aplikujeme najprv iba odstránenie vzdialených a vplyvných pozorovaní(DAF1) a odhadneme $RMSE$. Potom proces opakujeme s tým, že aplikujeme iba DAF, ktorá hľadá vhodnú transformáciu(DAF2) vysvetľovanej premennej. Takisto simulujeme iba použitie DAF, ktorá robí výber premenných(DAF3) do modelu na základe AIC štatistiky. Výrazný odklon naivného odhadu RMSE od skutočnej hodnoty nastáva už pri aplikovaní data analytickej funkcie, ktorá odstráni outliery a vplyvné pozorovania(DAF2)(MC:0.716 oproti naivnému:0.527) a ďalej sa prehlbuje krok po kroku výberom premenných(DAF3) na základe AIC(MC:0.509 oproti naivnému:0.293). V tomto prípade sme vyskúšali aplikovať odstránenie vzdialených a vplyvných pozorovaní a zároveň transformácie(DAF1 a DAF2). Výsledok(MC:0.717 oproti naivnému:0.460) ukazuje, že pri aplikovaní viac ako jednej DAF sa rozdiel medzi skutočnou hodnotou reprezentovanou Monte Carlo odhadom a naivným odhadom prehlbuje zatiaľ čo bootstrapový odhad skoro dokonale presne kopíruje hodnotu odhadu pomocou Monte Carlo metódy.
Po aplikovaní jednotlivých DAF samostatne je pri nelineárnom modeli vidieť, že bootstrapový odhad $RMSE$ má vrchol rozdelenia bližšie skutočnému odhadu, respektíve je menej vychýlený pri použití DAF1 viď. 3.3b, DAF2 viď. 3.4a aj DAF3 viď. 3.4b. Hodnota $RMSE$ získaná pomocou Monte Carlo simulácii z hľadiska mediánu je u jednotlivých DAF okrem výberu premenných DAF3 najvyššia. Bootstrapový odhad je vo všetkých prípadoch použitia DAF bližší skutočnej hodnote $RMSE$ odhadu regresnej funkcie než naivný odhad. V prípade, že by



(a) Neboli použité DAF. (b) Použitá iba DAF1.

Obr. 3.3: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre nelineárny model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.4: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre nelineárny model s aplikovaním DAF2 a DAF3 samostatne.

	Medián				
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3	DAF1 a DAF2
Monte Carlo	0.648	0.716	0.647	0.509	0.717
Naivný	0.589	0.527	0.514	0.293	0.460
Bootstrap	0.584	0.712	0.594	0.612	0.718

Tabuľka 3.4: Nelineárny model - Hodnoty odhadu $RMSE$ odhadu regresnej funkcie na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.

sme porovnávali hodnotu vychýlenia, kde sa nachádza vrchol rozdelenia, zistili by sme, že vrchol bootstrapového rozdelenia je bližšie skutočnej hodnote ako vrchol rozdelenia naivného odhadu aj napriek tomu, že u DAF3 medián nadhodnocuje skutočnú hodnotu. Za zmienku stojí aj výsledok po použití transformácie vysvetľovanej premennej 3.4a. Vidíme, že naivný odhad ma dvojvrcholové rozdelenie, čož bolo spôsobené transformáciou. Vyšší vrchol relatívne presne aproximuje skutočnú hodnotu $RMSE$ odhadu regresnej funkcie. V tomto prípade nie je použitie mediánu rozdelenia odhadov $RMSE$ najlepší indikátor. Bootstrapový odhad nie je dokonalý ale u nelineárného modelu je lepší ako naivný odhad pri porovnávaní vychýlenia vrcholov rozdelenia odhadu $RMSE$ odhadu regresnej funkcie.

3.1.3 Hetero model

Ďaľším modelom je heteroskedastický model, skrátene hetero model. Model s konštantným rozptylom sa verbálne označuje ako homoskedastický model. V prípade nekonštantného rozptylu hovoríme o heteroskedastickom modeli [6]. Definujeme teda model s nekonštantným rozptylom nasledovne

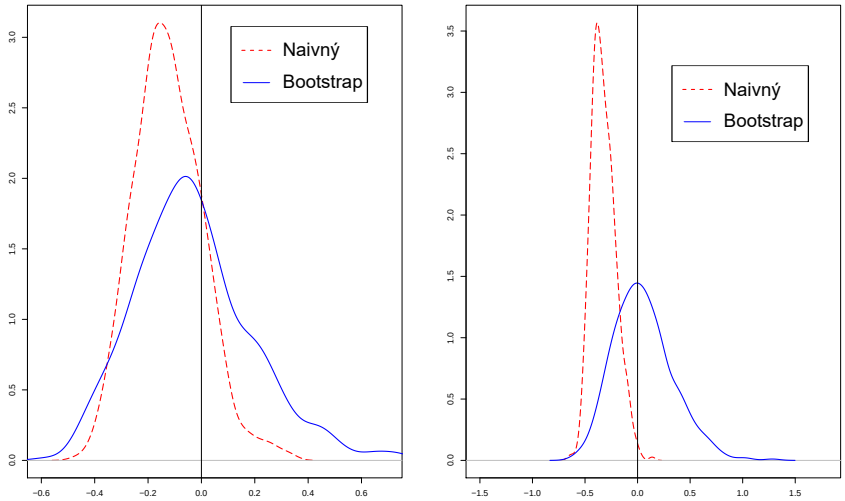
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \sum_{j=1}^{5} \beta_j x_{ij} \epsilon_i \tag{3.16}$$

pre $i = 1,...,n$, kde x_{ij} sú hodnoty generované z náhodných veličín $X_j \sim U(0,0.2)$, teda z rovnomerného rozdelenia z intervalu (0,0.2). Vidíme, že posledná zložka modelu 3.16 udáva váhu chybovej zlozky $w_i = \sum_{j=1}^{5} \beta_j x_{ij}$. S meniacou sa hodnotou w_i sa následne mení aj variabilita celého modelu. Tentokrát vidíme, že naivný a bootstrapový odhad $RMSE$ odhadu regresnej funkcie neaproximujú skutočnú hodnotu tak presne ako u iných modeloch. Výsledok sa nezmení ani po zvýšení počtu simulácii na 2000. Overíme s väčším počtom simulácii a bootstrapových výberov, že skutočne bootstrapový odhad je lepší oproti naivnému aj keď neaplikujeme žiadnu data analytickú funkciu. Na tento prípad použijeme 2000 simulácii a 1000 bootstrapových výberov, počet dát $n = 50$ ponecháme rovnaký. Z tabuľky 3.5 vidieť, že s presnejším výpočtom sú len minimálne odchylky od simulácii s počtom 500. Rozdelenie odhadov bolo takmer identické grafu 3.5a, preto sme sa rozhodli ho nezobraziť. V prípade, že nepoužijeme žiadne DAF, naivný odhad má užšie rozdelenie a bootstrapový odhad má o niečo ťažšie chvosty. Číselne má medián bootstrapového odhadu vo všetkých prípadoch bližšie k hodnote Monte Carlo odhadu teda ku skutočnému odhadu. Najväčší rozdiel naivného odhadu

$RMSE$ a skutočného $RMSE$ je v prípade použitia funkcie výberu premenných, teda DAF3.

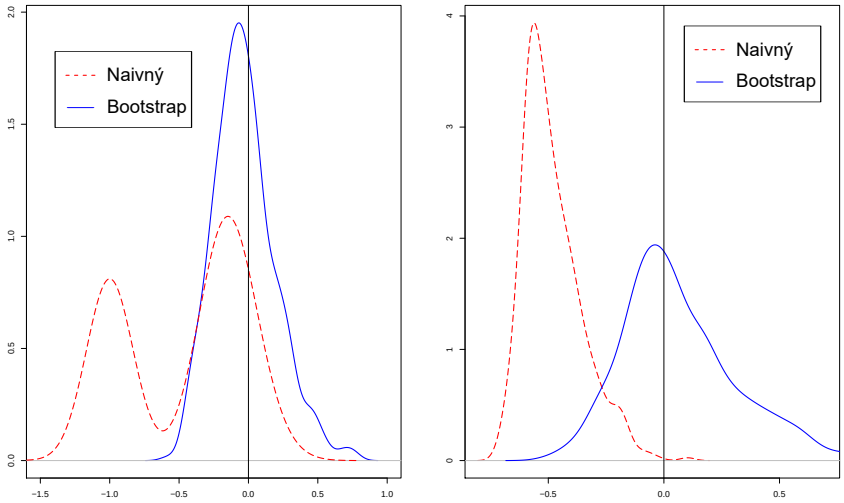
	Medián				
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	Bez DAF(2000 simulácii)	DAF1	DAF2	DAF3
Monte Carlo	0.351	0.352	0.405	0.357	0.344
Naivný	0.304	0.304	0.267	0.256	0.168
Bootstrap	0.333	0.333	0.415	0.340	0.353

Tabuľka 3.5: Hetero model - Hodnoty odhadu $RMSE$ odhadu regresnej funkcie na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.



(a) Neboli použité DAF. (b) Použitá iba DAF1.

Obr. 3.5: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre hetero model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.6: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre hetero model s aplikovaním DAF2 a DAF3 samostatne.

Po aplikovaní jednotlivých DAF je u hetero modelu bootstrapový odhad výrazne lepší ako naivný. Je to vidieť už po aplikovaní prvej data analytickej funkcie DAF1, viď. 3.5b, ktorá odstráni z modelu outliery a vplyvné pozorovania a pri aplikovaní ďalších funkcii je rozdiel viac a viac viditeľný. U DAF2, viď. 3.6a, sa znova opakuje prípad ako u nelineárného modelu a síce to, že na našom dátovom súbore nie je rozdelenie naivého odhadu $RMSE$ unimodálne, rozdelenie bootstrapového odhadu naopak je unimodálne, čo môže zjednodušiť určenie najpravdepodobnejšej hodnoty odhadu $RMSE$. Bootstrapový odhad $RMSE$ jemne nadhodnocuje skutočnú hodnotu pri použití DAF1 a DAF2 samostatne. Výrazny rozdieľ je vidieť pri použití DAF3, viď. obrázok 3.6b. Pri grafickom rozdelení aj pri jednotlivých hodnotách mediánu je jasne vidieť, že bootstrapový odhad je lepší ako naivný odhad.

3.1.4 Kolineárny model

Postupíme ďalej k simuláciam kolineárného modelu. Dve náhodné veličiny X_1 a X_2 sú presne kolineárne, alebo lineárne závislé, ak existuje lineárna rovnica

$$c_1 X_1 + c_2 X_2 = c_0, \tag{3.17}$$

taká, že pre nejaké konštanty c_0, c_1 a c_2 táto rovnica platí pre všetky dáta [2]. Multikolinearita podobne ako kolinearita označuje situáciu, kedy viac ako dve premenné majú medzi sebou lineárny vzťah [3]. Kolineárny model definujeme nasledovne:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \tag{3.18}$$

kde x_{i1} sú hodnoty generované z náhodnej veličiny X_1, viď. 3.19, x_{i2} sú hodnoty generované z náhodnej veličiny X_2, viď. 3.20, a hodnoty x_{i3} sú generované z náhodnej veličiny X_3, viď. 3.21,

$$X_1 \sim N(0,1) \tag{3.19}$$

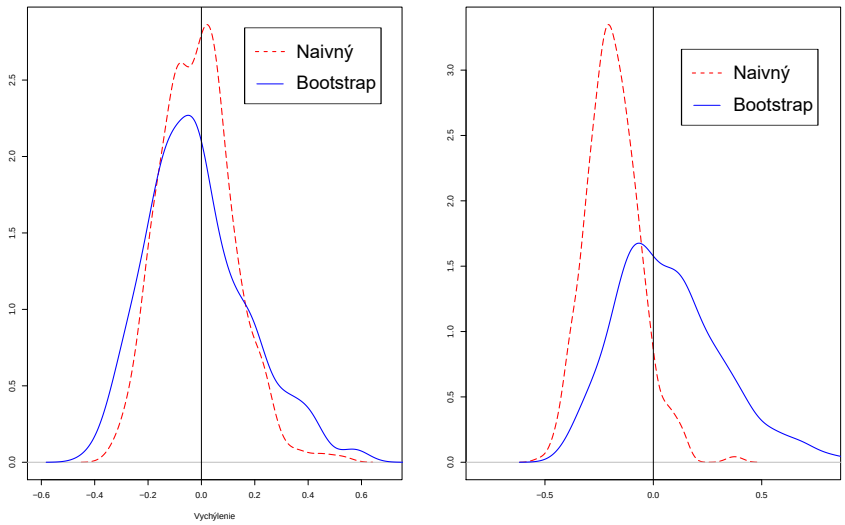
$$X_2, X_3 \sim N(0,0.01) - X_1 \tag{3.20}$$

$$\epsilon_i \sim N(0,1). \tag{3.21}$$

Z definície a názvu modelu sa dá usúdiť, že je prítomná v dátach multikolinearita. Kolineárny model má pri simuláciach bez použitia data analytických funkcí blízke hodnoty odhadu $RMSE$ ku skutočnej hodnote $RMSE$, ale rozhodne je medzi nimi vidieť rozdiel, obzvlášť u bootstrapoveého odhadu $RMSE$.

	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	4.236	4.680	5.223	3.043
Naivný	4.179	3.770	3.586	0.148
Bootstrap	4.075	4.917	4.671	4.128

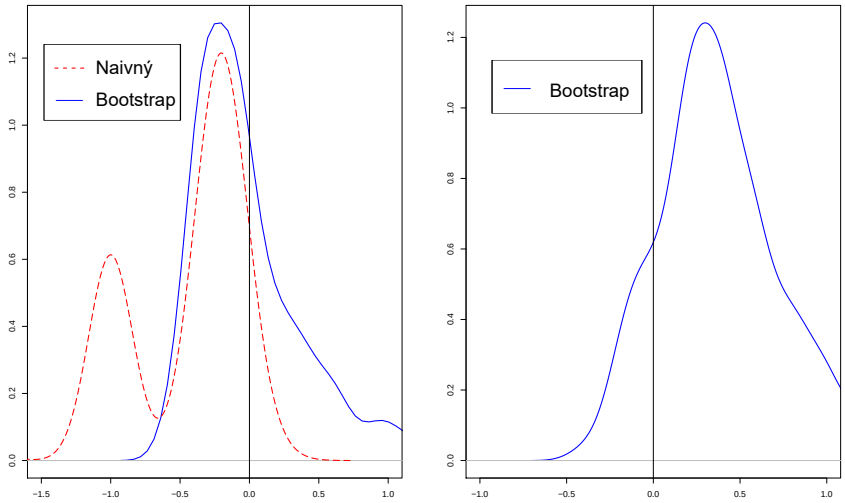
Tabuľka 3.6: Kolineárny model - Hodnoty odhadu $RMSE$ odhadu regresnej funkcie na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.



(a) Bez použitia DAF. (b) Použitá iba DAF1.

Obr. 3.7: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre kolineárny model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.

Po aplikovaní jednotlivých DAF vidíme v tabuľke 3.6, že naivný odhad výrazne podhodnocuje skutočnú hodnotu $RMSE$. Medián bootstrapového odhadu v 2 prípadoch(DAF1 a DAF3) zas nadhodnocuje odhad. Treba tiež dodať, že rozdelenie naivného estimátora pri použití DAF2, viď. obrázok 3.6a, znova nie je unimodálne. Pri použití iba DAF3 náš algoritmus zlyhal pri naivnom odhade $RMSE$ odhadu regresnej funkcie. Preto na obrázku 3.8b je zobrazené iba rozdelenie Z^{Boot}.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.8: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre kolineárny model s aplikovaním DAF2 a DAF3 samostatne.

3.1.5 Nasýtený model

Čo sa týka nasýteného modelu, v tomto prípade ukážeme ako presne aproximuje naivný prístup a metóda bootstrap na počet dát $n = 30$, $n = 50$ a zároveň vysoký počet premenných $p = 15$. Definujeme nasýtený model nasledovne

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} \\ + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} \\ + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} + \beta_{15} x_{i15} + \epsilon_i, \end{aligned} \quad (3.22)$$

pre $i = 1,...,n$, kde máme 16 regresných koeficientov $\beta_0,...,\beta_{15}$ a 15 vysvetľujúcich premenných. Hodnoty x_{ij} sú generované z náhodných veličín X_j pre $j = 1,...,15$, ktoré majú normálne rozdelenie s nulovou strednou hodnotou a jednotkovým rozptylom. Začneme pre $n = 30$. Z hodnôt zo simulácii, kde sme nepoužili žiadne DAF vidíme, že naivný odhad je veľmi blízky skoro skutočnej hodnote ale bootstrapový odhad je úplne inde než by mal byť, je silno nadhodnotený.

	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	0.328	0.399	0.734	0.336
Naivný	0.314	0.248	0.00037	0.312
Bootstrap	0.985	29.745	2.410	0.988

Tabuľka 3.7: Nasýtený model - Hodnoty odhadu $RMSE$ odhadu regresnej funkcie na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne, počet pozorovaní $n = 30$.

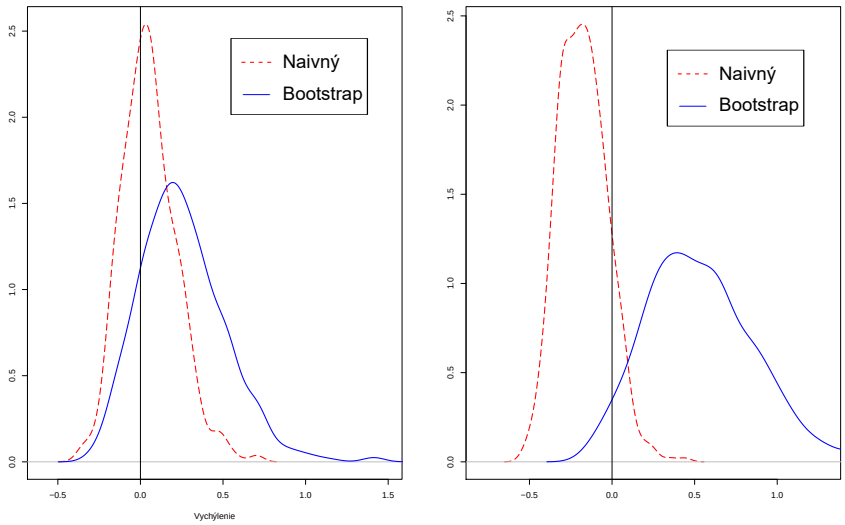
Po aplikovaní jednotlivých DAF vidíme v tabuľke 3.7, že problém s odhadom pomocou bootstrapu je problematický pri $n = 30$ a $p = 15$. Po aplikovaní DAF1 je

rozdiel ešte viac dramatický, zhruba 75-krát nadhodnocuje Monte Carlo odhad, teda skutočnú hodnotu, jednoducho zlyháva. Naopak naivný odhad aproximuje odhad $RMSE$ odhadu regresnej funkcie relatívne dobre, zlyháva iba pri aplikovaní transformácii(DAF2). Konštatujeme, že bootstrap nezvláda veľký pomer premenných na malé množstvo dát a vysoko nadhodnocuje odhad $RMSE$. Teda je pre nasýtený model nevhodný. Z dôvodu veľkých rozdielov v hodnotách neúvádzame grafické rozdelenie jednotlivých odhadov.

Ďalej vyskúšame podobnú simuláciu pre nasýtený model ale s počtom pozorovaní $n = 50$, počet premenných $p = 15$ zostáva rovnaký. Vo všetkých použitých DAF je odhad metódou bootstrap nadhodnotený. Ukazuje sa, že Bootstrap nie je schopný dodať uspokojivé výsledky pri nasýtenom modeli. Z tabuľky 3.8 a obrázkoch 3.9 a 3.10 usudzujeme, že naivný odhad je pri tomto modeli s 50 pozorovaniami lepší ako bootstrapový odhad v prípade, že nepoužijeme žiadnu DAF a v prípade, keď použijeme výber premenných(DAF3).

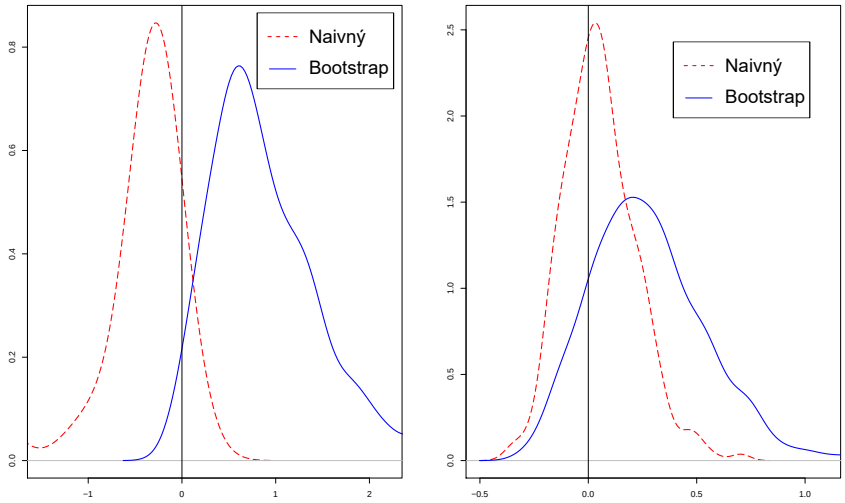
	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	0.201	0.227	0.293	0.201
Naivný	0.209	0.184	0.185	0.209
Bootstrap	0.246	0.343	0.524	0.251

Tabuľka 3.8: Nasýtený model - Hodnoty odhadu $RMSE$ odhadu regresnej funkcie na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne, počet pozorovaní $n = 50$.



(a) Bez použitia DAF. (b) Použitá iba DAF1.

Obr. 3.9: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre nasýtený model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne, $n = 50$.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.10: Hustota rozdelenia Z^{Naive} a Z^{Boot} pre nasýtený model s aplikovaním DAF2 a DAF3 samostatne, $n = 50$.

3.2 Efekt

V predchádzajúcej sekcii sme sa zaoberali s akou veľkou presnosťou odhaduje naivný a bootstrapový odhad $\widehat{RMSE}$ skutočné $RMSE$ odhadu regresnej funkcie v určitom bode. Teraz sa zameriame na podobnú analýzu s tým, že tentokrát sa pokúsime aproximovať $RMSE$ odhadu regresného koeficienta β_1, respektíve efekt β_1. Všeobecne výpočítame efekt β_j podľa [10] ako

$$\text{Efekt}(\beta_j) = \frac{d}{dx_j} g^{-1}\left[\beta_0 + \sum_{i=1}^{p} \beta_i f_i(x_i)\right] |\ X = \boldsymbol{x}_0, \tag{3.23}$$

kde $\frac{d}{dx_j}$ je derivácia podľa j-tej vysvetľujúcej premennej. Ďalej g^{-1} je inverzná funkcia z 3.1 a f je funkcia takisto z 3.1. V našom prípade, keď chceme zistiť β_1 dosadíme za $j = 1$. V prípade, že g, f_j sú identity, potom Efekt(β_j) sú klasické β_j regresné koeficienty. Znova v simuláciach vyskúšame viaceré modely. A takisto ako pri aproximácii odhadu $RMSE$ odhadu regresnej funkcie definujeme normované štatistiky pre odhad RMSE odhadu Efekt(β_1) a to V^{Naive} a V^{Boot}, ktoré nám zjednodušia posúdiť presnosť aproximácie oproti skutočnej hodnote, ktorá je reprezentovaná Monte Carlo odhadom $RMSE$. Definujeme V^{Naive} a V^{Boot} ako

$$V^{Naive} = \frac{\widehat{RMSE}^{Naive}(b_1) - \widehat{RMSE}^{MC}(b_1)}{\widehat{RMSE}^{MC}(b_1}, \tag{3.24}$$

$$V^{Boot} = \frac{\widehat{RMSE}^{Boot}(b_1) - \widehat{RMSE}^{MC}(b_1)}{\widehat{RMSE}^{MC}(b_1)}, \tag{3.25}$$

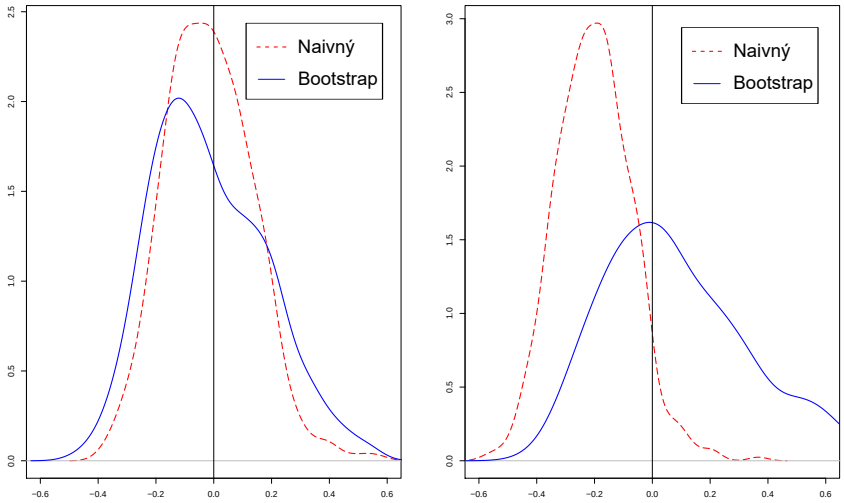
kde b_1 je odhad regresného koeficientu β_1.

3.2.1 Vanila model

Začneme s vanila modelom. Z tabuľky 3.9 a obrázkoch 3.11b a 3.12a pozorujeme významne lepší výsledok bootstrapového odhadu oproti naivnému odhadu. Všeobecne hodnotíme bootstrapový odhad nadradený naivnému odhadu pri odhade $RMSE$ odhadu regresného parametru β_1 vo vanila modeli. Rovnako ako pri odhade $RMSE$ odhadu regresnej funkcie nie je vidieť rozdiel medzi obrázkami 3.11a a 3.12b, z čoho usudzujeme, že neboli eliminované žiadne premenné z prvotného modelu. Situácia, ktorú by sme očakávali pri odhadovaní modelu z dát z vanila modelu.

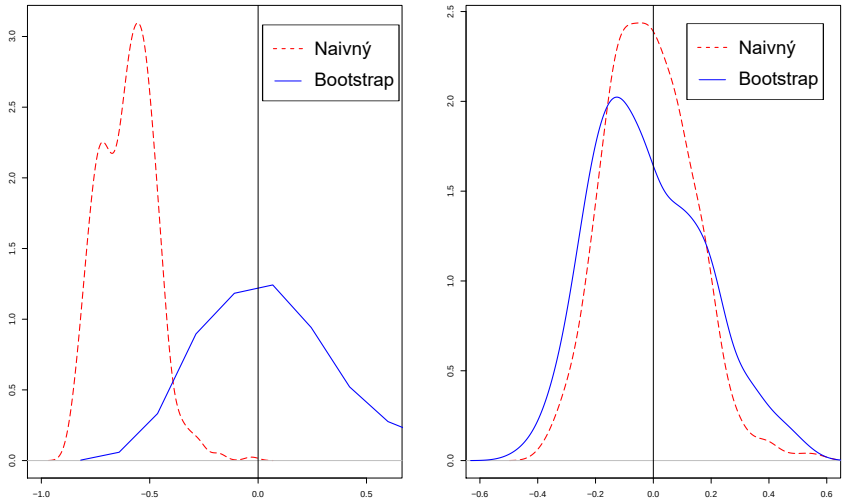
	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	0.154	0.171	0.333	0.154
Naivný	0.151	0.135	0.134	0.151
Bootstrap	0.148	0.179	0.351	0.148

Tabuľka 3.9: Vanila model - Hodnoty odhadu $RMSE$ odhadu parametru β_1 na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.



(a) Bez použitia DAF. (b) Použitá iba DAF1.

Obr. 3.11: Hustota rozdelenia V^{Naive} a V^{Boot} pre vanila model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

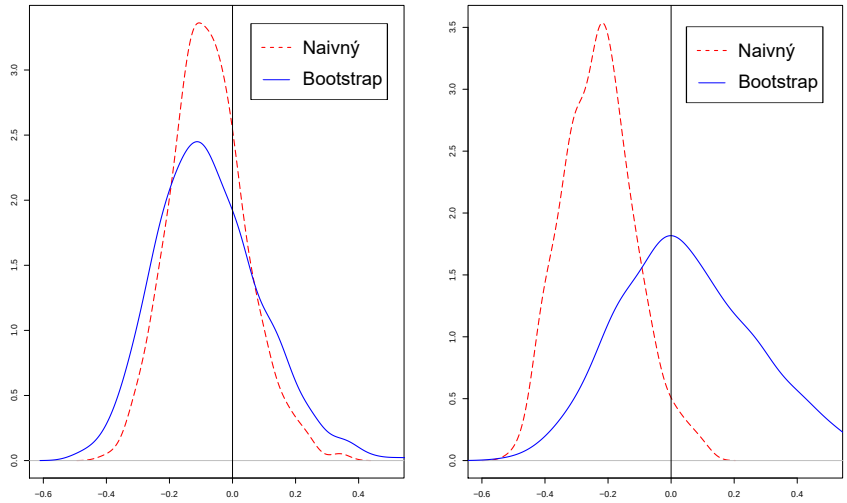
Obr. 3.12: Hustota rozdelenia V^{Naive} a V^{Boot} pre vanila model s aplikovaním DAF2 a DAF3 samostatne.

3.2.2 Nelineárny model

Pri nelineárnom modeli v sekcii 3.1 sme zmienili, že keďže predpokladáme, že dáta boli získané z lineárného modelu ale v skutočnosti boli generované z nelineárného modelu, očakávame nesúlad medzi Monte Carlo odhadom $RMSE$ na jednej strane a naivným a boostrapovým odhadom $RMSE$ na strane druhej. Táto úvaha sa nám potvrdila už pri odhadovaní $RMSE$ odhadu regresnej funkcie a pri odhadovaní $RMSE$ odhadu regresného parametra β_1 sme znova mohli pozorovať diskrepanciu medzi týmito odhadmi.

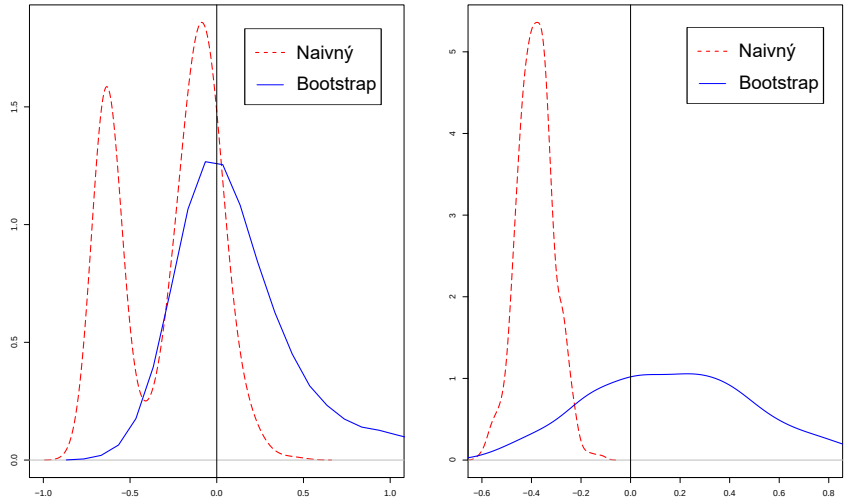
	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	2.816	3.053	2.836	3.957
Naivný	2.589	2.356	2.328	2.431
Bootstrap	2.562	3.153	3.058	4.595

Tabuľka 3.10: Nelineárny model - Hodnoty odhadu $RMSE$ odhadu parametru β_1 na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.



(a) Bez použitia DAF. (b) Použitá iba DAF1.

Obr. 3.13: Hustota rozdelenia V^{Naive} a V^{Boot} pre nelineárny model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.14: Hustota rozdelenia V^{Naive} a V^{Boot} pre nelineárny model s aplikovaním DAF2 a DAF3 samostatne.

Podobne ako u vanila modelu konštatujeme, že bootstrapový odhad je jednoznačne lepší v prípade použitia DAF1, viď 3.13b, a DAF2, viď 3.14a. U DAF2, viď 3.14a, znova ako pri odhade RMSE odhadu regresnej funkcie nemá naivný odhad unimodálne rozdelenie. Aj pri nelineárnom modeli na základe hodnôt mediánu a jednotlivých obrázkoch 3.13 a 3.14 považujeme bootstrap za vhodnú metódu aproximovania skutočnej hodnoty $RMSE$ odhadu regresného parametru β_1. Pri použití DAF3, teda výberu premenných, má rozdelenie bootstrapových odhadov $RMSE$ ťažké chvosty, avšak medián tohto rozdelenia aproximuje skutočnú hod-

notu $RMSE$ lepšie ako medián rozdelenia naivných hodnôt.

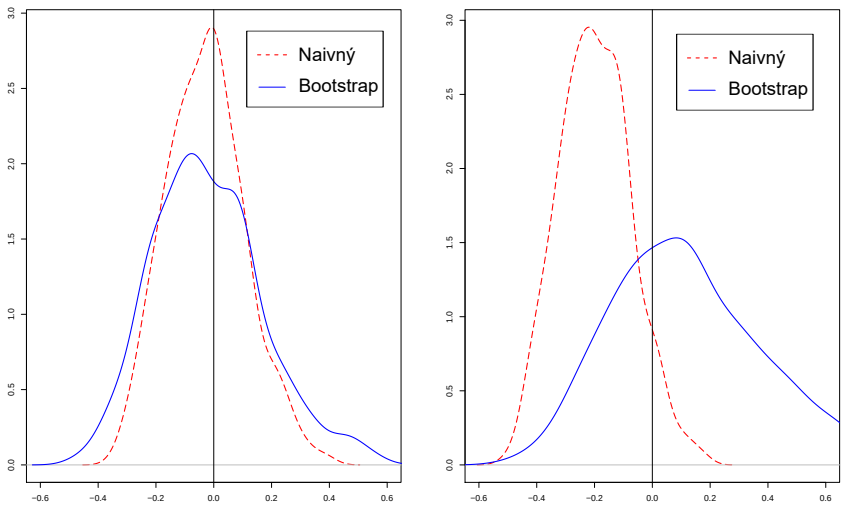
3.2.3 Hetero model

Prípomíname, že hetero model 3.16 má oproti vanila modelu rôznu váhu chybovej zložky $w_i = \sum_{j=1}^{p} \beta_j f_j(x_{ij})$, ktorá spôsobuje nekonštantný rozptyl, respektíve heteroskedasticitu.

	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	1.381	1.463	2.011	1.716
Naivný	1.346	1.167	1.112	1.257
Bootstrap	1.326	1.612	2.652	1.857

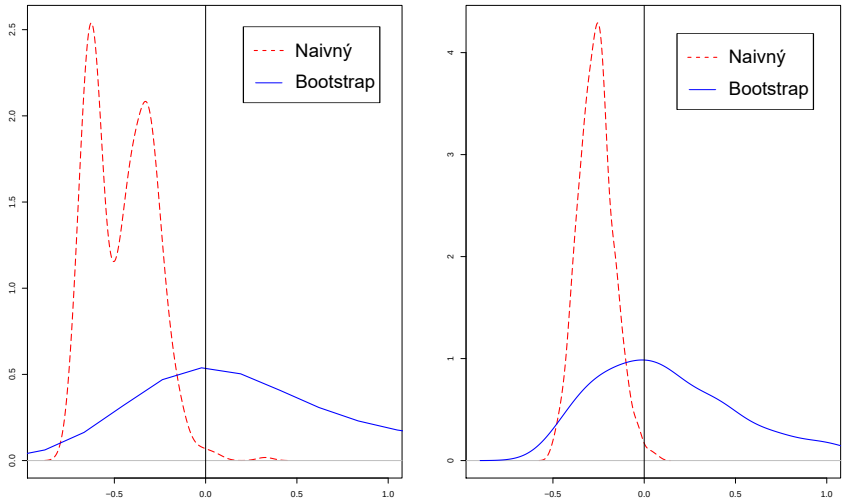
Tabuľka 3.11: Hetero model - Hodnoty odhadu $RMSE$ odhadu parametru β_1 na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.

Všeobecne z obrázkov 3.15 a 3.16 nás zaujal fakt, že rozdelenie bootstrapových odhadov je širšie(respektíve má ťažšie chvosty) ako rozdelenie naivných odhadov. Pri použití DAF2 3.16a sme dostali rovnako ako pri odhade $RMSE$ odhadu regresnej funkcie dovjvrcholové rozdelenie naivného odhadu.



(a) Bez použitia DAF. (b) Použitá iba DAF1.

Obr. 3.15: Hustota rozdelenia V^{Naive} a V^{Boot} pre hetero model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 samostatne.



(a) Použitá iba DAF2. (b) Použitá iba DAF3.

Obr. 3.16: Hustota rozdelenia V^{Naive} a V^{Boot} pre hetero model s aplikovaním DAF2 a DAF3 samostatne.

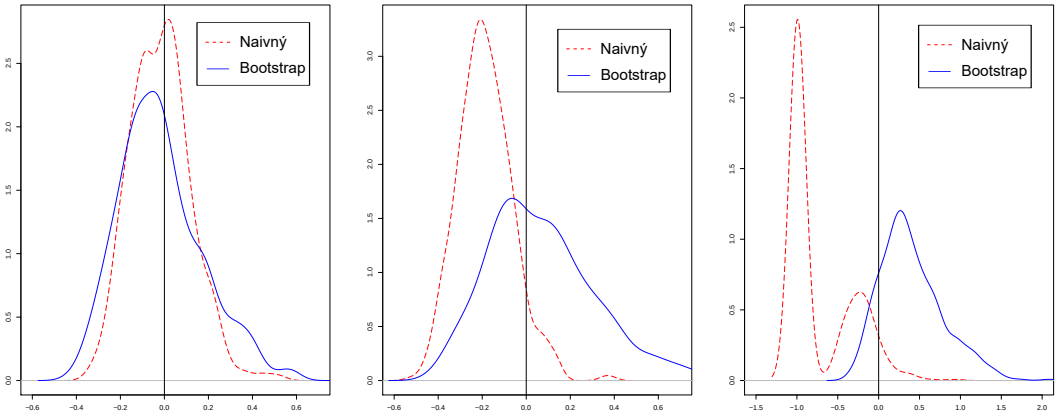
V prípade, že zoberieme do úvahy kombináciu výsledkov z tabuľky 3.11 a obrázkov 3.15 a 3.16, považujeme bootstrapový odhad za lepší, pretože vrcholy bootstrapového rozdelenia sú bližšie skutočnej hodnote a mediány rozdelenia, i keď často nadhodnotené, sú takisto bližšie skutočnej hodnote, teda odhadu RMSE metódou Monte Carlo odhadu regresného koeficientu β_1.

3.2.4 Kolineárny model

Postupujeme v analýze kolineárného modelu. Používame rovnaký model ako v sekcii 3.1. Z tabuľky 3.12 je jasne vidieť zlyhanie Monte Carlo odhadu aj bootstrapového odhadu pri použití DAF2. Hodnoty Monte Carla odhadu a bootstrapového odhadu sú nezmyselné, preto sme obrázok s rozdelením jednotlivých odhadov pri použití DAF2 nezobrazili. Naopak pri použití DAF3 zlyháva naivný odhad, ktorý hlboko podhodnocuje skutočnú hodnotu a navyše jeho rozdelenie nie je unimodálne.

	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo	21.213	23.414	1.161×10^{16}	16.442
Naivný	20.915	18.820	18.458	0.157
Bootstrap	20.388	24.542	11 804 307 232	21.849

Tabuľka 3.12: Kolineárny model - Hodnoty odhadu $RMSE$ odhadu parametru β_1 na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.



(a) Bez použitia DAF. (b) Použitá iba DAF1. (c) Použitá iba DAF3.

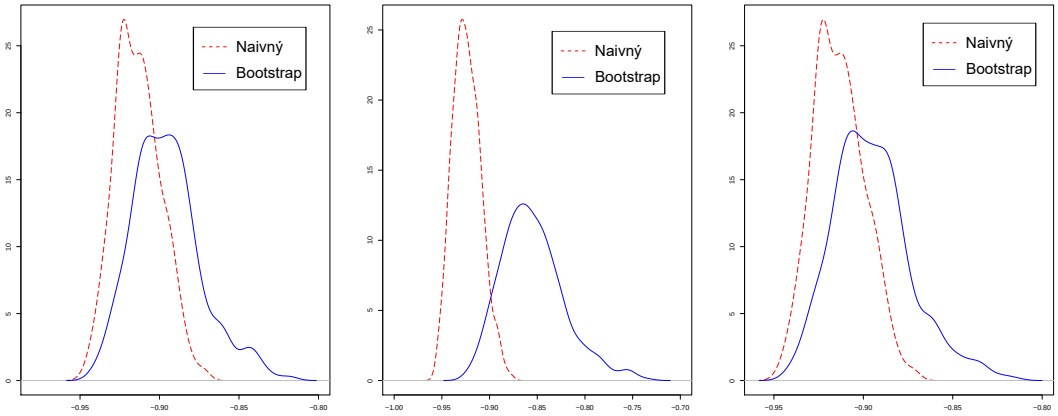
Obr. 3.17: Hustota rozdelenia V^{Naive} a V^{Boot} pre kolineárny model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 a DAF3 samostatne.

3.2.5 Nasýtený model

U nasýteného modelu máme $p = 15$ premenných. Najprv budeme simulovať $n = 30$ a následne $n = 50$ pozorovaní. Pre $n = 30$ sú výsledky pre bootstrapový odhad chaotické. Nedávajú zmysel hlavne pri použití DAF1 a DAF2. Usudzujeme, že po použití DAF1 a DAF2 bootstrap zlyháva pre $n = 30$. Naivný odhad $RMSE$ je zhruba 10-násobne podhodnotený. Pri $n = 50$ situácia nie je o moc lepšia, tentokrát bootstrap zlyháva iba pri DAF2(preto nezobrazujeme graf pri použití DAF2), inak je spolu s naivným odhadom $RMSE$ asi 10-násobne podhodnotený. Usudzujeme, že naivný a bootstrapový odhad $RMSE$ odhadu regresného parametru β_1 nie je presná aproximácia skutočnej hodnoty $RMSE$ odhadu regresného parametru β_1.

	Medián			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Monte Carlo, n = 30	2.016	2.028	2.185	2.009
Naivný, n = 30	0.252	0.201	0.249	0.251
Bootstrap, n = 30	0.806	26.698	48 458.98	0.815
Monte Carlo, n = 50	1.999	2.001	2.296	1.999
Naivný, n = 50	0.171	0.151	0.156	0.171
Bootstrap, n = 50	0.206	0.284	328.328	0.206

Tabuľka 3.13: Nasýtený model - Hodnoty odhadu $RMSE$ odhadu parametru β_1 na základe Monte Carlo metódy, naivného prístupu a bootstrapového prístupu bez a s aplikovaním jednotlivých DAF samostatne.



(a) Bez použitia DAF. (b) Použitá iba DAF1. (c) Použitá iba DAF3.

Obr. 3.18: Hustota rozdelenia V^{Naive} a V^{Boot} pre nasýtený model bez aplikovania akejkoľvek DAF a s aplikovaním DAF1 a DAF3 samostatne, $n = 50$.

3.3 Algoritmus odhadu RMSE odhadu regresnej funkcie

Algoritmus implementovaný v RStudiu a použitý v predchádzajúcej kapitole stručne vysvetlíme s použitím vanilla modelu. Kód bude zameraný na odhad $RMSE$ odhadu regresnej funkcie v bode $\boldsymbol{x}_0$. Výpočet odhadu $RMSE$ odhadu regresného parametru β_1 je v zásade veľmi podobný, dokonca jednoduchší, pretože po získaní finálneho modelu už poznáme hodnotu b_1.

Dáta generujeme pomocou funkcie *rnorm* so strednou hodnotou 0 a smerodatnou odchylkou rovnou 1. *sim* je premenná funkcie *vanilla* a označuje počet simulovaných dát v jednotlivých dátových súboroch.

```
vanilla <- function(sim){
yy <- matrix(,nrow = sim, ncol = 6);
a <- rnorm(sim); b <- rnorm(sim); c <- rnorm(sim);
d <-rnorm(sim); e <- rnorm(sim)
res <- rnorm(sim)
y <- a + b + c + d + e + res
yy[,1] <- y; yy[,2] <- a; yy[,3] <- b;
yy[,4] <- c;yy[,5] <- d; yy[,6] <- e
return(yy)}
```

Funkcia *vanilla* z kódu vyššie vygeneruje maticu yy, do ktorej postupne vkladáme vygenerované vektory, ktoré reprezentujú hodnoty vysvetľujúcich premenných. Následne po prvotnej iniciácii základných premenných a funkcí spustíme hlavný for cyklus, v ktorom vygenerujeme z modelu zakaždým dátový súbor o rovnakom počte pozorovaní. Takýto datový súbor označíme *data_van*.

```
data_van <- vanilla(pocdat)
```

Dostaneme dátovú maticu, ktorá má 50 riadkov a v každom riadku je jedna

instancia vygenerovaných dát, teda hodnota odozvy a hodnoty jednotlivých vysvetľujúcich premenných. Ďalej zvýšime hodnotu odozvy o nejakú konštantu, tak aby vygenerované hodnoty odozvy boli nezáporné. V našom kóde túto konštantu označíme *const*. Robíme to kvôli tomu, aby sme mohli dáta transformovať pomocou logaritmu a odmocniny v data analytických funkciách. Pri generovaní dát z normálneho rozdelenia totiž dostávame aj záporné hodnoty a v prípade, že naša vysvetľovaná premenná má zápornú hodnotu nie je možné ju zlogaritmovať ani odmocniť, pretože logaritmus a odmocnica zápornej hodnoty nie je definovaná. Ďalej vytvoríme prvotný model za účelom identifikácie vzdialených a vplyvných pozorovaní. Na pripomenutie, prvotný model je taký, ktorý vznikne po odhadnutí dát, ktoré neboli nijako pozmenené, metódou najmenších štvorcov. Funkcia *outlier_fun* identifikuje vzdialené a vplyvné pozorovania

```
outlier_fun = function(model,pocetdat,mydata){

### X je regresna matica
X = mydata[,-1]
X1 = rep(1,pocetdat)
X = as.matrix(cbind(X1,X))

### H je projekcna matica, di znaci Cookovu vzdialenost
H = X %*% solve(t(X) %*% X) %*% t(X)
s2 <- summary(model)$sigma^2
swi <- c(seq(1,pocetdat))
di <- c(seq(1,pocetdat))
rezidua = as.matrix(resid(model))
rownames(rezidua) = seq(1,pocetdat)
for (k in seq(1,pocetdat)) {
swi[k] <-
sqrt(((pocetdat-p)*s2 - rezidua[k]^2/(1-H[k,k]))/(pocetdat-p-1))
di[k] = rezidua[k]^2*H[k,k]/s2/p/(1-H[k,k])^2
}

### reswi su externe studentizovane rezidua
reswi <- c(seq(1,pocetdat))
for (l in seq(1,pocetdat)) {
reswi[l] <- rezidua[l]/sqrt((1-H[l,l]))/swi[l]
}

bonf_kv = qt(0.975/pocetdat,pocetdat-p-1)
c_points = which(di > 1)
outlier_points = which(abs(reswi) > abs(bonf_kv))
outlier_points = c(outlier_points, c_points)
outlier_points = unique(outlier_points)
d=c(d,di)
return(outlier_points)}
```

Výsledkom je zoznam vzdialených a vplyvných pozorovaní. Vzdialené pozorovania(outliery) sú počítané pomocou externe študentizovaných reziduí a následne tieto hodnoty pomocou *bonferoniho korekce* 1.3.1 identifikujú vzdialené pozorovania. Podobne pri zisťovaní vplyvných pozorovaní použijeme Cookovú vzdialenosť,

ktorá ak je väčšia ako 1 označuje vplyvné pozorovanie. V našom kóde sú externe študentizované reziduá označené ako *reswi* a Cooková vzdialenosť ako *di*.

Výber premenných modelu je určený na základe hodnoty AIC. V R programe máme balíček *olsrr*, ktorý obsahuje už naprogramovanú funkciu *ols stepaic backward*, viď. 1.3.4, a *ols stepaic forward*, viď. 1.3.4, funkciu. Rozdiel v týchto funkciách je ten, že prvá funkcia začína s plným modelom a postupne podľa hodnoty AIC vyraďuje premenné, ktoré po vyradení z modelu najviac znížia hodnoty AIC štatistiky. V každom kroku skúša akú hodnotu AIC štatistiky by mal model bez určitej vysvetľujúcej premennej. Takto vyskúša všetky vysvetľujúce premenné a vyberá taký model, ktorý má bez určitej vysvetľujúcej premennej zo všetkých vysvetľujúcich premenných najnižšiu hodnotu AIC štatistiky. *Ols stepaic forward* začína s modelom bez vysvetľujúcich premenných, iba s konštantou. Postupne pridáva premenné, podľa toho, ktorá najviac znížia hodnotu AIC štatistiky modelu. Do funkcie vstupuje súčasný model a výsledkom tejto funkcie je model, ktorý obsahuje iba premenné opísané pomocou metódy v 1.3.4 teda iba vybrané premenné podľa opísaného postupu. Funkcia *aicback* ako výstupnú hodnotu vydá odhadnutý model, na ktorý bola aplikovaná spätná eliminačná metóda, teda výber premenných.

```
aicback <- function(model,daata){
pom11 <- ols_stepaic_backward(model, details = FALSE)
k <- pom11$predictors
if(is.null(k)) return(model) else
{pom11 <- update(as.formula(model),
as.formula(paste('.~.','-',paste(k,collapse = '-'))));
pom11 = lm(pom11,daata)
return(pom11)}}
```

Funkcia *aicforw* aplikuje metódu postupného výberu premenných na základe AIC.

```
aicforw <- function(model){
    pomfor1 <- ols_stepaic_forward(model, details = FALSE)
    l <- pomfor1$predictors
    pomfor2 <- update(model,as.formula(paste('y ~',paste(l,collapse =
        '+'))))
    return(pomfor2)
  }
```

Ďalej sú tieto funkcie použité v hlavnom aj vo vnútornom for cykle. Funkcia *tryCatch* je kvôli tomu, aby keď funkcia *ols stepaic backward* nevyradí žiadnu premennú zvládla bez chyby vrátiť výsledný neredukovaný model. Premenná *final_mod* udáva finálny model.

```
final_mod <- tryCatch(aicback(van_model_new,bc_data_van2),
error = function(e) final_mod = aicforw(van_model_new))
```

Nasleduje hlavné telo programu a ten sa skladá z dvoch for cyklov, jeden vnorený vo vnútri toho druhého. Postupne sú aplikované data analytické funkcie v poradí ako sme opísali v kapitole 1.3. Transformácia spočíva v porovnaní modelu, ktorý má najnižšiu šikmosť reziduí. Porovnáme 3 modely, klasický netransformovaný,

taký, v ktorom je zlogaritmovaná odozva a taký, v ktorom je odmocnená odozva.

```
van_model2 <- lm(y ~ vanx1 + vanx2 + + vanx3 + vanx4 + vanx5, data =
= bc_data_van2)
van_model3 = lm(log(y) ~ vanx1 + vanx2 + vanx3 + vanx4 + vanx5, data =
= bc_data_van2)
van_model4 = lm(sqrt(y) ~ vanx1 + vanx2 + vanx3 + vanx4 + vanx5, data =
 = bc_data_van2)
s2 = skewness(residuals(van_model2))
s3 = skewness(residuals(van_model3))
s4 = skewness(residuals(van_model4))
s = c("s2", "s3", "s4")[which.min(c(abs(s2),abs(s3),abs(s4)))]
if(s == "s3") final_mod = van_model3 else
if(s == "s4") final_mod = van_model4 else
final_mod = van_model2
```

Z finálneho modelu určíme odhad regresnej funkcie v bode $\boldsymbol{x}_0$ ako

```
### theta_i je odhad regresnej funkcie v bode x0 = (0.2,0.2,...,0.2)
    theta_i <- predict(final_mod, x0, se.fit = TRUE).
```

Hodnotu $\widehat{RMSE}^{MC}(\hat{\theta})$ vypočítame podľa vzorca 3.3. Hodnotu $\widehat{RMSE}^{Naive}(\hat{\theta})$ získame ako medián rozdelenia odhadov smerodatných odchyliek odhadu regresnej funkcie v bode $\boldsymbol{x}_0$. Rozdelenie smerodatných odchyliek odhadu regresnej funkcie v bode $\boldsymbol{x}_0$ získam združením hodnôt

```
naive_theta_i = theta_i$se.fit.
```

Hodnotu $\widehat{RMSE}^{Boot}(\hat{\theta})$ získame podobne ako $\widehat{RMSE}^{MC}(\hat{\theta})$ vo vnorenom for cykle. Vnorený for cyklus má skoro rovnaký kód ako hlavný for cyklus len s takým rozdielom, že dáta sú bootstrapovým výberom z prvotného dátového súboru a pri výpočte $\widehat{RMSE}^{Boot}(\hat{\theta})$ podľa vzorca 3.4 za hodnotu $\hat{\theta}$ použijeme odhad regresnej funkcie v bode $\boldsymbol{x}_0$ získaný z odhadnutého modelu z hlavného for cyklu.

3.4 Výpočtová náročnosť

Pripomeňme, že naše analýzy bežali pri počte dát $n = 50$(respektíve aj $n = 30$ pre nasýtený model), počtu simulácii 500 a počte bootstrapových výberov pre každú simuláciu 100. Simulácie bežali vo viacerých variantách, prvá bola bez aplikovania data analytických funkcí a ďalšie boli s jednotlivým aplikovaním DAF samostatne. To znamená, že simulácia pre jeden model s jednou variantou použitej DAF vypočítala $100 * 500 = 50000$ odhadnutých modelov, každý na základe iného dátového súboru. Celkovo sme teda pre každý model(s výnimkou nasýteného modelu) robili 8 simulácii, z toho 4 pre odhad $RMSE$ odhadu regresnej funkcie a 4 pre odhad $RMSE$ odhadu regresného parametru β_1. Simulácie bežali na osobnom počítači so štvorjadrovým CPU o frekvencii 3,4 GHz. Avšak program R funguje tak, že výpočet beží iba na 1 jadre CPU. Pri ostatných výpočtoch sme si vytvorili nový projekt a týmto spôsobom sme dokázali spustiť paralelne výpočet

pre ďalšie modely, signifikantne sme tým ušetrili čas. U obidvoch typoch simulácii bola približne rovnaká výpočtová náročnosť. Pochopiteľne pretože sme počítali výsledný model u obidvoch, rozdiel bol iba v poslednej časti výpočtu, ktorý je výpočtovo jednoduchý. Celkový čas jednej simulácie záležal od toho, ktorá data analytická funkcia bola použitá a takisto, či boli dáta z nasýteného modelu alebo nie. Z tabuľky 3.14 vidíme, že ak neboli použité DAF, celkový čas sa pohyboval od 220 sekúnd(3,6 minúty) u vanila modelu až po 400 sekúnd(6,6 minúty) u nasýteného modelu. Výpočet s DAF1(identifikácia a odstránenie vzdialených a vplyvných pozorovaní) zaberal o niečo viac času ako výpočet bez DAF a výpočet s DAF2(transformácia vysvetľovanej premennej) zabral o trochu viac času ako výpočet s DAF1. Jednoznačne najviac časovo náročný bol výpočet s DAF3(výber premenných), ktorý zabral u vanila modelu 5 387 sekúnd(1,5 hodiny) a u nasýteného modelu 17 313 sekúnd(4,8 hodiny). Dôvodom tak vysokého nárastu časovej náročnosti je použitie funkcii *ols_stepaic_backward* a *ols_stepaic_forward*, ktoré sme opísali v 1.3.4. U ostatných modeloch sa čas potrebný na výpočet pohyboval medzi hodnotami vanila a nasýteného modelu.

	Celkový čas výpočtu v sekundách			
$\widehat{RMSE}(\hat{\theta})$	Bez DAF	DAF1	DAF2	DAF3
Vanila	220	326	456	5 387
Nasýtený, n = 50	400	533	804	17 313

Tabuľka 3.14: Celkový čas výpočtu pre vanila a nasýtený model.

Záver

Cieľom našej práce bolo v duchu analýzy [10] priblížiť aký vplyv má výstavba lineárného regresného modelu(použitie data analytických funkcii(DAF) = identifikácia a odstránenie vzdialených a vplyvných pozorovaní, transformácia vysvetľovanej premennej a výber premenných) na presnosť klasického odhadu smerodatnej chyby odhadu regresnej funkcie v bode a odhadu regresného parametru β_1. Kvalitu sme usudzovali pomocou odmocniny zo strednej štvorcovej chyby($RMSE$), viď 2.3. Hlavný prínos tejto práce je posúdenie do akej miery dokáže metóda bootstrap priblížiť skutočnú hodnotu kvality modelu meranú pomocou $RMSE$. Na rozdiel od práce [10] sme aplikovali jednotlivé data analytické funkcie osobitne a skúmali sme aký čiastkový vplyv majú na odhad kvality finálného modelu. Rozhodli sme sa tak preto, lebo pri aplikovaní všetkých DAF sme zistili, že použitie transformácie vysvetľovanej premennej a zároveň výber premenných prináša výsledky, ktoré sa silno nezhodujú s hodnotou, ktorú by sme očakávali. V tomto prípade nepomohol ani bootstrap. Ako transformáciu vysvetľovanej premennej sme použily buď odmocninu alebo logaritmus vysvetľovanej premennej. Na takto transformovaný model sme aplikovali výber premenných a dostali sme finálny model. Ďalej sme pomocou finálného modelu získali odhad regresnej funkcie v určitom bode. V prípade, že bola použitá transformácia modelu, bolo nutné kvôli porovnateľnosti hodnôt spätne transformovať odhad regresnej funkcie, aby sme dostali hodnotu, ktorá bola škáľovo rovnaká ako mal finálny model bez transformácie. V prípade odmocniny, spätná transformácia zahŕňa v sebe druhú mocninu, v prípade logaritmusu spätná transformácia zahŕňa exponenciálu. Exponenciála aj druhá mocnina určitej hodnoty dokážu aj pri malých zmenách tejto hodnoty dostať diametrálne odlišný výsledok. Zaradenie alebo vyradenie určitej vysvetľujúcej premennej preto mohlo mať veľký vlpyv na výslednú hodnotu odhadu regresnej funkcie respektíve odhadu regresného parametru β_1. Kombinácia týchto dvoch DAF viedla k nerozumným výsledkom, i keď osobitné použitie transformácie alebo výberu premenných prinášalo očakávané výsledky. Výstavba regresného modelu môže byť zložitý proces a urobiť procedúru, ktorá vždy určí vhodnú transformáciu a správny výber premenných je náročné. Vráťme sa ale k výsledkom našej analýzy.

Výsledky analýzy bez použitia DAF môžme zosumarizovať nasledovne. Najprv zhrnieme výsledky odhadu $RMSE$ odhadu regresnej funkcie v určitom bode. A priori predpokladáme, že naše dáta pochádzajú z lineárného regresného modelu, preto výsledky Monte Carlo odhadu $RMSE$ sú zhruba rovnaké pre vanila, hetero a kolineárny model. Nasýtený model dosiahol silno odlišné výsledky, ktoré zhrnieme na konci. U nelineárného modelu je diskrepancia medzi odhadmi, ktorú sme aj očakávali. Konkrétne medzi hodnotou Monte Carlo odhadu RMSE, naivným odhadom smerodatnej odchylky a bootstrapovým odhadom RMSE. Použitie

prvej DAF, teda odstránenie odľahlých a vplyvných pozorovaní spôsobilo podhodnotenie naivného odhadu. Bootstrapový odhad vo všetkých modeloch okrem nasýteného dosiahol lepšie výsledky ako naivný odhad. Problém s aproximáciou skutočnej $RMSE$ odhadu regresnej funkcie mal bootstrap v prípade kolineárného modelu, výsledky boli avšak stále lepšie ako u naivného odhadu. Čo sa týka nasýteného modelu pre $n = 30$ bootstrap zlyhal, viac sa darilo naivnému odhadu, pre $n = 50$ bootstrap nezlyhal ale hodnoty boli značne nadhodnotené, znova bol naivný odhad bližšie skutočnému odhadu reprezentovaným Monte Carlo odhadom. Na otázku, ktorá z DAF spôsobila najväčšie podhodnotenie naivného odhadu $RMSE$ odhadu regresnej funkcie v bode sa nedá jednoznačne odpovedať. Závisí to od zvoleného modelu.

Pri analýze odhadu $RMSE$ odhadu regresného parametru β_1 boli výsledky podobné ako pri odhade $RMSE$ odhadu regresnej funkcie. U nelineárného a hetero modelu mal bootstrap tendenciu jemne nadhodnocovať odhad $RMSE$. Metóda Monte Carlo a bootstrapová metóda tentokrát zlyhala pri kolineárnom modeli. U nasýteného modelu(pre $n = 30$ aj $n = 50$) zlyhal bootstrap pri použití transformácie vysvetľovanej premennej. Bootstrap v ostatných prípadoch aproximoval veľmi dobre skutočnú hodnotu $RMSE$ oproti naivnému odhadu. Pri obidvoch typoch simuláciach bolo zaujímavé pozorovať ako po použití transformácii vysvetľovanej premennej rozdelenie naivného odhadu nebolo unimodálne u nelineárneho, hetero a kolineárného modelu. Vo väčšine prípadov bolo rozdelenie bootstrapových odhadov menej špicaté než rozdelenie naivných odhadov.

V našej práci sme ukázali ako pri výstavbe lineárného regresného modelu, použité procedúry(odstránenie odľahlých a vplyvných pozorovaní, transformácia a výber premenných) spôsobia, že naivný odhad $RMSE$ odhadu regresnej funkcie alebo regresného parametru je výrazne podhodnotený oproti skutočnej hodnote. Demonštrovali sme ako je možné, aj v prípade použitých data analytických funkcii, pomocou metódy bootstrap úspešne aproximovať skutočnú hodnotu $RMSE$ odhadu regresnej funkcie alebo odhadu regresného parametru. Navyše sme ukázali do akej miery jednotlivé data analytické funkcie spôsobia podhodnotenie naivného odhadu RMSE.

Literatúra

[1] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Boca Raton, Florida, USA. Chapman & Hall/CRC, 1993. p. 426. ISBN 978-0412042317.

[2] Sanford Weisberg. *Applied Linear Regression*. Third Edition. Hoboken, New Jersey: John Wiley & Sons, Inc., 2005. 330 p. ISBN 0-471-66379-4.

[3] Sanford Weisberg. *Applied Linear Regression*. Second Edition. New York: J. Wiley & Sons, Inc., 1985. 324 p. ISBN 0-471-87957-6.

[4] Bradley Efron. *Bootstrap Methods: Another Look at the Jackknife*. Ann. Statist., 1979, vol. 7, no. 1, p. 1-26.

[5] J. Scott Long and Laurie H. Ervin.*Correcting for Heteroscedasticity with Heteroscedasticity Consistent Standard Errors in the Linear Regression Model: Small Sample Considerations*. January 1998.

[6] Tomáš Cipra. *Finanční Ekonometrie*. 1. vydanie. Praha: Ekopress, s.r.o., 2008. 519 s. ISBN 978-80-86929-43-9.

[7] C. F. J. Wu. *Jackknife, bootstrap and other resampling methods in regression analysis*. Ann. Statist., 1986, vol. 14, no. 4, p. 1261-1295.

[8] Zuzana Prášková. *Metoda Bootstrap*. Univerzita Karlova, Fakulta matematicko-fyzikální.

[9] Mohammad Mohammadi. *On the bounds for diagonal and off-diagonal elements of the hat matrix in the linear regression model*. REVSTAT – Statistical Journal, 2016, vol. 14, no. 1, p. 75-87.

[10] Julian J. Faraway. *On the Cost of Data Analysis*. Journal of Computational and Graphical Statistics, 1992, vol. 1, no. 3, p. 1-15.

[11] Peter J. Bickel and David A. Freedman. *Some asymptotic theory for the bootstrap*. Ann. Statist., 1981, vol. 9, no. 6, p. 1196-1217.