UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Rankové metódy v lineárnej regresii viacerých premenných

DIPLOMOVÁ PRÁCA

2018 Bc. Juraj ŠNEGOŇ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Rankové metódy v lineárnej regresii viacerých premenných

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2018 **Bc. Juraj Šnegoň**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Juraj Šnegoň
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Rankové metódy v lineárnej regresii viacerých premenných.
Rank-based approach to multiple linear regression.

Cieľ: Úlohou bude porovnať pomocou počítačových simulácií v jazyku R bežné a rankové metódy [testy či intervaly spoľahlivosti] v lineárnej regresii s viacerými vysvetľujúcimi premennými.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 26.01.2017

Dátum schválenia: 27.01.2017

prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou sa chcem poďakovať svojmu vedúcemu diplomovej práce Mgr. Jánovi Somorčíkovi, PhD. za príležitosť pracovať na veľmi zaujímavej téme a za cenné rady, pripomienky a usmernenia, ktoré mi veľmi pomohli počas celej tvorby diplomovej práce. Taktiež sa chcem poďakovať svojim rodičom za ich trpezlivosť a podporu počas celého vysokoškolského štúdia.

Abstrakt

ŠNEGOŇ, Juraj: Rankové metódy v lineárnej regresii viacerých premenných [Diplomova práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2018.

V tejto práci sme sa zaoberali porovnaním parametrických a neparametrických prístupov vo viacnásobnej lineárnej regresii. Cieľom bolo prezentovať výhody neparametrických metód v prípadoch, v ktorých náhodné chyby pochádzajú z rozdelenia s ťažšími chvostami, t. j. výskyt outlierov bol častejší. Porovnanie spomínaných dvoch prístupov sme vykonali z troch pohľadov. Praktická časť našej práce je rozdelená do troch častí: odhady regresných parametrov, testovanie signifikantnosti regresných parametrov a intervaly spoľahlivosti pre regresné parametre. Napríklad ak náhodné chyby pochádzali z Cauchyho rozdelenia, tak neparametrické metódy dosiahli oveľa lepšie výsledky ako klasické parametrické metódy vo viacnásobnej lineárnej regresii. Neparametrická regresia je veľmi užitočný nástroj, pretože neparametrické metódy sú robustné, čo sa prejavilo aj vo výsledkoch tejto práce.

Kľúčové slová: Neparametrická lineárna regresia pomocou viacerých premenných, Scale parameter, Score function, Odhad regresných parametrov, Testovanie významonsti regresných parametrov, Intervaly spoľahlivosti

Abstract

ŠNEGOŇ, Juraj: Rank-based approach to multiple linear regression [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2018.

In this paper we were engaged in comparison parametric and nonparametric approach in multiple linear regression. The aim is to present the advantages of nonparametric methods in cases where random errors are from heavy-tailed distribution. The comparison of the mentioned approaches was made on the basis of three criteria. The practical part of our master thesis is divided into three parts: regression coefficients estimation, significance test and confidence intervals for regression parameters. For instance, if the random errors were from Cauchy distribution, the nonparametric methods reached much better results than classical parametric methods in multiple linear regression. Nonparametric regression is a very useful tool because non-parametric methods are robust, which was also reflected in the results of this work.

Keywords: Nonparametric multiple linear regression, Scale parameter, Score function, Regression coefficients estimation, Significance test, Confidence interval

Obsah

Úvod

Dnešný svet je mimoriadne zložitý. Je čoraz komplikovanejšie odhaliť vzťahy a vzájomné závislosti medzi rôznymi objektmi okolo nás. Schopnosť popísať takéto vzťahy je mimoriadne dôležitá, pretože v mnohých prípadoch máme obmedzený zdroj dát, ktorý je častokrát aj malého rozsahu a získanie ďalších dodatočných údajov môže byť finančne a časovo náročné. Vo veľa situáciách sa ukazuje, že vzťahy medzi veličinami sú lineárne, a preto je veľmi dôležité vedieť správne modelovať a popisovať lineárnu závislosť.

Jedným z najpoužívanejších matematických nástrojov na popísanie závislosti medzi rôznymi premennými je lineárna regresia. Jedným z dôvodov prečo je lineárna regresia veľmi populárnym štatistickým nástrojom pri vytváraní modelu, ktorý popisuje predpokladanú závislosť, je jej jednoduchá aplikácia. Problematike lineárnej regresie sa venuje veľmi veľa publikácií ako napríklad [7] alebo [17]. Existuje nespočetne veľa softwarov, ktoré majú prostriedky lineárnej regresie v sebe efektívne zabudované. Stačí, ak vložíme do programu vstupné údaje a veľmi rýchlo získame potrebné výsledky. Často sa stáva, že ľudia vôbec netušia ako získali tieto výsledky. Nezaoberajú sa podstatou použitých metód ani predpokladmi, ktoré musia byť splnené na to, aby dané postupy fungovali správne a s potrebnou spoľahlivosťou.

Jedným z hlavných predpokladov pri používaní klasickej lineárnej regresie je normálne rozdelenie náhodných chýb. Z toho dôvodu by sa pred samotným použitím tejto metódy malo otestovať, či náhodné chyby pochádzajú z normálneho rozdelenia. Pokiaľ sa ukáže, že chyby z normálneho rozdelenia nepochádzajú, niektoré metódy lineárnej regresie nebudú mať požadovanú kvalitu. V praxi sa však lineárna regresia veľmi často používa bez toho, aby sa spomínaný predpoklad overil. Dôsledkom toho je to, že sa výsledky interpretujú veľakrát nesprávne a vôbec nespĺňajú kvalitatívne ukazovatele, ktoré sa od nich vyžadujú: napríklad použitý test môže mať skutočnú pravdepodobnosť chyby 1. druhu inú než je nominálna hladina alebo vypočítaný interval spoľahlivosti môže mať skutočnú spoľahlivosť inú než nominálna hladina.

Existujú metódy, ktoré spomínaný nedostatok klasickej lineárnej regresie odstraňujú. Inými slovami nevyžadujú predpoklad normálneho rozdelenia pre náhodné chyby. Mnohé tieto metódy a postupy sú založené na rankoch, a preto sa často označujú aj

ako rankové metódy a keďže nie sú viazané na konkrétne rozdelenie, zaraďujú sa k neparametrickým metódam. Predovšetkým pri rozdeleniach s ťažšími chvostami ako je normálne rozdelenie by neparametrické metódy mali fungovať oveľa lepšie, pretože by mali byť menej citlivé na výskyt outlierov v dátach. Z toho dôvodu má zmysel sa zaoberať neparametrickou lineárnou regresiou a ukázať jej reálne výhody v praxi v porovnaní s klasickými metódami.

Hlavným cieľom našej diplomovej práce je porovnať parametrickú a neparametrickú lineárnu regresiu pomocou počítačových simulácií v softwari R. Našou snahou bude ukázať, že neparametrické metódy sú kvalitnejšie a presnejšie, ak nie je splnený predpoklad normálneho rozdelenia náhodných chýb. Pri jednotlivých simuláciách budeme meniť podmienky simulácií a nastavenia jednotlivých metód a pokúsime sa zistiť ich vplyv na výsledky. Parametrické a neparametrické metódy porovnáme na základe viacerých kritérií. Najprv sa pozrieme na odhady neznámych regresných parametrov získaných jednotlivými metódami a budeme vyhodnocovať, ktorá z nich dosiahla presnejšie výsledky. Ako druhý krok sa pozrieme, ako si metódy parametrickej a neparametrickej regresie budú počínať pri testovaní signifikantnosti regresných parametrov. Budú nás zaujímať chyby testov ako aj samotné silofunkcie jednotlivých testov, ktoré nám pomôžu pri vzájomnom porovnaní. Na záver sa zameriame na intervaly spoľahlivosti pre regresné parametre a počínanie si parametrických a neparametrických metód pri ich konštrukcii.

Prvé dve kapitoly sa zaoberajú teóriou lineárnej regresie. 1. kapitola je venovaná parametrickej lineárnej regresii, v ktorej v stručnosti rozoberáme základne pojmy a postupy. 2. kapitola je tvorená poznatkami z neparametrickej lineárnej regresie, ktorá je menej rozšírená. Táto kapitola vychádza z knihy [5], v ktorej je veľmi dobre spracovaná nielen neparametrická lineárna regresia, ale aj veľa iných metód z neparametrickej štatistiky. Ďalšie 3 kapitoly sú určené pre výsledky našich simulácií, na základe ktorých porovnáme metódy parametrickej a neparametrickej lineárnej regresie.

1 Klasická parametrická lineárna regresia

1.1 Základné pojmy a charakteristiky v klasickej lineárnej regresii

V tejto kapitole sme uviedli niektoré základné poznatky z klasickej lineárnej regresie s viacerými vysvetľujúcimi premennými. Vychádzali sme z knihy [7] prípadne [17]. Model popisujúci vzťah medzi vysvetľovanou premennou Y a vysvetľujúcimi premennými $x_1, x_2, \ldots, x_p$ je definovaný pomocou vzťahu 1.1

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i,\ i = 1, \ldots, n. \tag{1.1}$$

Index i označuje $i-te$ pozorovanie v rámci všetkých n meraní, ktoré máme k dispozícii. Vektor pozostávajúci zo všetkých závislých premenných sme označili ako $Y = (Y_1, Y_2, \ldots, Y_n)^T$, podobne vektor náhodných chýb ako $e = (e_1, e_2, \ldots, e_n)^T$, vektor regresných parametrov ako $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ a matica X je definovaná ako

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n-1} & x_{2,n-1} & \ldots & x_{p,n-1} \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{p,n} \end{bmatrix}. \tag{1.2}$$

Model 1.1 sme zapísali maticovo a dostali sme analogický model definovaný vektorovým vzťahom

$$Y = X\beta + e. \tag{1.3}$$

Z dôvodu vyhnutia sa rôznym patologickým situáciám sme predpokladali, že počet meraní n je väčší ako počet neznámych parametrov, ktoré chceme odhadnúť t. j. $n > p+1$, a stĺpce matice X sú lineárne nezávislé. Platnosť týchto 2 podmienok budeme vyžadovať aj v nasledujúcich kapitolách našej diplomovej práce. Hlavný predpoklad v celej parametrickej lineárnej regresii je však viazaný na náhodné chyby. Predpokladali sme, že náhodné chyby $e = (e_1, e_2, \ldots, e_n)^T$ sú iid a pochádzajú z normálneho rozdelenia so strednou hodnotou 0 a disperziou σ^2. Uvedomme si, že predpoklad iid je pomerne prísny, pretože to znamená, že v modeli 1.1 nemôže byť medzi chybami $e = (e_1, e_2, \ldots, e_n)^T$ autokorelácia a ani heteroskedasticita.

1.2 Odhad neznámych parametrov pomocou metódy najmenších štvorcov

Myšlienka metódy najmenších štvorcov je veľmi jednoduchá a priamočiara. Hlavná idea je obsiahnutá už v samotnom názve tejto metódy. Odhad neznámych regresných parametrov sa konštruuje tak, aby súčet štvorcov zvislých odchýlok medzi skutočnými realizáciami závislej premennej Y a príslušnými odhadmi týchto realizácií bol čo najmenší. Uvedenú ideu môžeme matematicky zapísať ako

$$\hat{\beta} = arg\min_{\beta} \|Y - X\beta\|_2^2 . \tag{1.4}$$

Našou úlohou je nájsť minimum konvexnej funkcie. Na jeho nájdenie sme využili poznatky z diferenciálneho počtu. Musí platiť, že

$$\frac{\partial \|Y - X\beta\|_2^2}{\partial \beta} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{p+1} . \tag{1.5}$$

Na nájdenie odhadov neznámych parametrov je teda potrebné vyriešiť sústavu $p+1$ rovníc o $p+1$ neznámych $\beta_0, \beta_1, \ldots, \beta_p$. Po krátkom odvodzovaní získame analytické vyjadrenie pre odhad neznámych regresných parametrov dane vzťahom

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y. \tag{1.6}$$

Do pozornosti dávame, že pri celom odvodení odhadov neznámych parametrov sa vôbec nevyužíva predpoklad týkajúci sa normálneho rozdelenia chýb. Taktiež ani fakt, že chyby sú rovnako rozdelené a nezávislé, nie je pri určení odhadu $\hat{\beta}$ potrebný. Na druhej strane pokiaľ chyby $e_1, e_2, \ldots, e_n$ pochádzajú z rozdelenia s ťažšími chvostami ako je normálne rozdelenie, teda výskyt outlierov je častejší, odhady nájdené metódou najmenších štvorcov môžu byť dosť nepresné. Jedným z našich cieľov je práve tento nedostatok metódy najmenších štvorcov demonštrovať pomocou simulácií v ďalších častiach. Čo sa týka predpokladov kladených na náhodné chyby $e_1, e_2, \ldots, e_n$, ich platnosť je nutná, pokiaľ chceme, aby napríklad testy hypotéz o parametroch mali požadovanú kvalitu alebo intervaly spoľahlivosti pre parametre požadovanú presnosť.

1.3 Test signifikantnosti regresných parametrov

V tejto časti sme sa stručne zaoberali testovaním hypotéz o významnosti jednotlivých regresných parametrov $\beta_1, \beta_2, \ldots, \beta_p$ v našom modeli danom vzťahom 1.1. Skúmali sme, či niektoré z vysvetľujúcich premenných $x_1, x_2, \ldots, x_p$ prislúchajúcich k daným regresným parametrom nemôžeme z modelu 1.1 vyhodiť, pričom by však nedošlo k výraznému zníženiu kvality modelu. Každý regresný model je zjednodušením reality resp. nejakého reálneho vzťahu prípadne závislosti medzi vysvetľovanou premennou a vysvetľujúcimi premennými. Je celkom prirodzené, že čím viac vysvetľujúcich premenných v modeli ponecháme, tým bude mať model väčšiu modelovaciu silu. Na druhej strane, čím viac vysvetľujúcich premenných model obsahuje, tým je aj zložitejší. Naším cieľom je hľadať model, ktorý čo najpresnejšie popisuje problém alebo závislosť, s ktorou sa zaoberáme, a súčasne chceme, aby bol tento model čo najjednoduchší. Preto ma zmysel zaoberať sa otázkou signifikancie jednotlivých parametrov $\beta_1, \beta_2, \ldots, \beta_p$, pretože môžeme nájsť model, ktorý je menej zložitý, pričom popisuje závislosť, ktorá je predmetom nášho záujmu veľmi podobne ako nejaký iný komplexnejší model s viacerými vysvetľujúcimi premennými.

Nech I je množina indexov $1, \ldots, p$ t. j. $I = \{1, 2, \ldots, p\}$ a množina Q je jej podmnožinou, $Q \subseteq I$. Hypotéza, ktorej testovaním sme sa zaoberali vyzerá nasledovne:

$$H_0 : \beta_l = 0,\ l \in Q\ vs.\ H_1 : \beta_l \neq 0,\ l \in Q. \tag{1.7}$$

Hypotéza H_0 hovorí, že vysvetľujúce premenné $x_l, l \in Q$ nie sú dôležité v modeli 1.1, môžeme ich zanedbať a lineárny model zjednodušiť. Ak hypotéza H_0 platí, modelujeme závislú premennú Y pomocou tzv. *submodelu*, ktorý je tvorený vysvetľujúcimi premennými $x_l, l \in I \setminus Q$. Pokiaľ je množina Q totožná s množinou I, vtedy horeuvedenou hypotézou testujeme významnosť celej regresie. Inými slovami, či model 1.1 má vôbec zmysel konštruovať na popísanie závislosti medzi vysvetľujúcimi premennými $x_1, x_2, \ldots, x_p$ a vysvetľovanou premennou Y. Počet prvkov v množine Q sme označili q. Myšlienka tohto testu je veľmi jednoduchá. Pozrieme sa na súčet štvorcov rezíduí RSS_M v modeli 1.1 a RSS_{SM} v *submodeli*. Čím je súčet štvorcov rezíduí menší, tým je model presnejší. Pokiaľ platí, že $RSS_{SM} \gg RSS_M$ znamená to, že *submodel* je oveľa nekvalitnejší než model 1.1. Teda nie je postačujúci na popis nášho problému a hypotézu H_0 v takomto prípade zamietame. Testovacia štatistika riadiaca sa za platnosti

H_0 Fisherovým-Snedecorovým rozdelením s q a $n-p-1$ stupňami voľnosti vyzerá nasledovne:

$$F = \frac{(RSS_{SM} - RSS_M)/q}{RSS_M(n-p-1)}. \tag{1.8}$$

Vidíme, že hlavná myšlienka celého testu je obsiahnutá v čitateli testovacej štatistiky F. Hypotézu H_0 budeme zamietať pokiaľ $F \gg 0$ a vzhľadom na to, že testovacia štatistika sa riadi Fisherovým-Snedecorovým rozdelením so spomínanými stupňami voľnosti, tak ako medzu extrémnej kladnosti použijeme 5% kritickú hodnotu tohto rozdelenia t. j. $F_{0.05}(q, n-p-1)$. Teda H_0 zamietneme, ak bude platiť $F > F_{0.05}(q, n-p-1)$.

1.4 Interval spoľahlivosti

Okrem klasických bodových odhadov neznámych regresných parametrov $\beta_1, \beta_2, \ldots, \beta_p$ nás často zaujímajú aj intervalové odhady. Odvodenie klasického intervalu spoľahlivosti pre parameter $\beta_l, l = 1, \ldots, p$ vychádza z predpokladu, že náhodné chyby majú normálne rozdelenie.Potom platí aj:

$$\hat{\beta}_l \sim N(\beta_l, \sigma^2 (X^T X)^{-1}_{ll}), \tag{1.9}$$

kde $(X^T X)^{-1}_{ll}$ označuje $l-ty$ diagonálny prvok matice $(X^T X)^{-1}$. Taktiež vieme, že náhodná premenná $\frac{(n-p-1)S^2}{\sigma^2}$ sa riadi Chi-kvadrát rozdelením s $n-p-1$ stupňami voľnosti, teda platí:

$$\frac{(n-p-1)S^2}{\sigma^2} \sim \chi^2_{n-p-1}. \tag{1.10}$$

Po aplikovaní štandardizácie vo vzťahu 1.9 s cieľom, aby sa náhodná premenná riadila štandardizovaným normálnym rozdelením a následne predelením náhodnou premennou zo vzťahu 1.10 získame náhodnú premennú, ktorá sa riadi Studentovým rozdelením

$$\frac{\hat{\beta}_l - \beta_l}{S\sqrt{(X^T X)^{-1}_{ll}}} \sim t_{n-p-1}. \tag{1.11}$$

Na základe uvedených skutočností je veľmi jednoduché odvodiť 95% interval spoľahlivosti pre $\beta_l, l = 1, \ldots, p$:

$$(\hat{\beta}_l - t(2.5\%)_{n-p-1}\sqrt{S(X^T X)^{-1}_{ll}}\ ;\ \hat{\beta}_l + t(2.5\%)_{n-p-1}\sqrt{S(X^T X)^{-1}_{ll}}). \tag{1.12}$$

2 Neparametrická lineárna regresia

2.1 Priamkový model

V tejto časti sme sa v stručnosti zaoberali priamkovou neparametrickou regresiou. Ako hlavný zdroj informácií sme používali knihu [5]. Model, ktorý sme uvažovali je definovaný nižšie vzťahom (2.1)

$$Y_i = \xi + \beta x_i + e_i,\ i = 1, \ldots, n. \tag{2.1}$$

Môžeme si všimnúť iné označenie pre intercept v porovnaní s modelom 1.1, čo sme spravili schválne kvôli lepšej prehľadnosti a orientácii v ďalších častiach našej práce. Hlavný rozdiel oproti klasickej parametrickej regresii spomínanej v 1. kapitole je v predpokladoch kladených na náhodné chyby $e_1, e_2, \ldots, e_n$. Predpokladali sme, že náhodné chyby $e_1, e_2, \ldots, e_n$ pochádzajú z nejakého nám neznámeho spojitého rozdelenia a sú *iid*. Upúšťa sa od predpokladu normálneho rozdelenia a taktiež sa nevyžaduje ani symetrické rozdelenie. Na druhej strane pomerne striktný predpoklad *iid* zabezpečujúci neautokorelovanosť a homoskedastickosť náhodných chýb ostáva zachovaný.

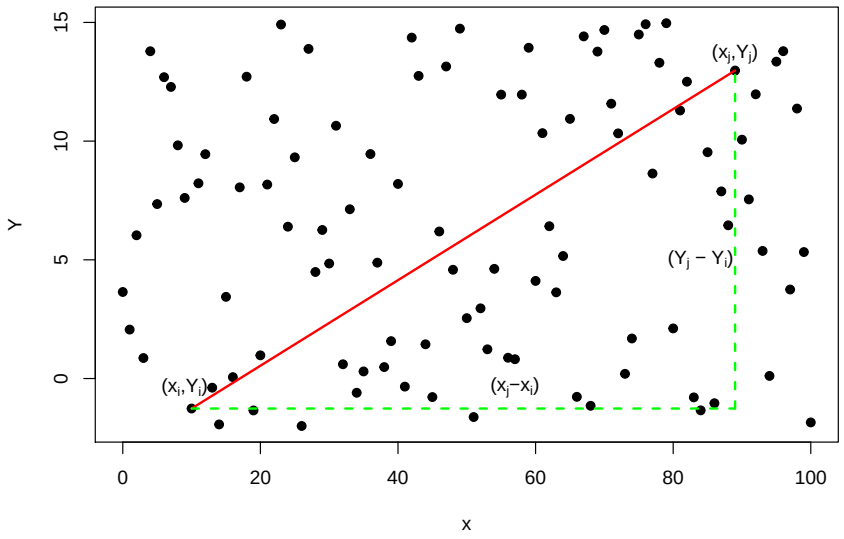
V prvom rade sme sa venovali nájdeniu odhadu parametra β. Myšlienka je pomerne priamočiara a jednoduchšia oproti metóde najmenších štvorcov. Zoberme si $i-te$ a $j-te$ meranie a spojme ich priamkou. Sklon tejto priamky označme S_{ij} a je definovaný ako

$$S_{ij} = \frac{Y_j - Y_i}{x_j - x_i},\ 1 \leq i < j \leq n. \tag{2.2}$$

Analogicky zostrojíme všetky možné priamky cez všetky dvojice bodov. Spolu získame $\binom{n}{2}$ priamok s rovnakým počtom sklonov. Uvedené myšlienky zobrazujeme na Obr. 2.1. Odhad sklonu regresnej priamky je následne určený ako medián zo všetkých možných sklonov S_{ij},

$$\hat{\beta} = median\left\{S_{ij} = \frac{Y_j - Y_i}{x_j - x_i}, 1 \leq i < j \leq n\right\}. \tag{2.3}$$

Uvedený odhad sa nazýva Theil-Senov odhad. Veľmi podobný odhad pre parameter β sa nazýva Siegelov odhad. Dostaneme ho tak, že z každého bodu $(x_i, y_i),\ i = 1, \ldots, n$ zostrojíme $n-1$ priamok do zvyšných bodov $(x_j, y_j),\ j = 1, \ldots, n$ a $j \neq i$. Pre každý bod (x_i, y_i) si spočítame príslušné sklony $n-1$ priamok, ktoré z daného bodu vychádzajú a zistíme z nich medián. Takto získame n mediánov, z ktorých opätovne



Obr. 2.1: Závislosť medzi vysvetľujúcou premennou x a závislou premennou Y

zrátame medián a to bude Siegelov odhad pre parameter β. Myšlienka metódy najmenších štvorcov je na prvý pohľad zložitejšia. Nie je problém nahliadnuť, že odhad sklonu regresnej priamky pomocou metódy najmenších štvorcov je vážený priemer zo sklonov S_{ij} definovaných vzťahom 2.2. Váha jednotlivých sklonov S_{ij} je úmerná vzdialenosti $x_j - x_i$. Dôkaz predchádzajúcej úvahy nie je vôbec zložitý, nebudeme ho explicitne uvádzať v našej práci, pretože to nie je náš prvoradý cieľ, no považovali sme za potrebné aspoň spomenúť takúto zaujímavú analógiu medzi metódou najmenších štvorcov a Theil-Senovým odhadom.

Na rozdiel od klasického parametrického prístupu z 1.kapitoly odhad interceptu ξ určíme osobitne až po zrátaní odhadu sklonu regresnej priamky β. Idea je veľmi jednoduchá. Zrátajú sa rezíduá $e_i^* = Y_i - \hat{\beta} x_i,\ i = 1, \ldots, n.$ a odhad pre intercept bude daný pomocou vzťahu

$$\hat{\xi} = median\left\{e_i^*, i = 1, \ldots, n\right\}. \tag{2.4}$$

Čo sa týka testovania hypotézy o signifikantnosti parametra β myšlienka testu je pomerne zaujímavá. Predpokladajme, že hypotéza, ktorá nás zaujíma vyzerá nasledovne:

$$H_0 : \beta = 0 \; vs. \; H_1 : \beta \neq 0. \tag{2.5}$$

Ako prvý krok si zavedieme pomocné premenné $Z_i = Y_i - \beta x_i, i = 1, \ldots, n$. Za platnosti hypotézy H_0 si môžeme vyjadrenie pre premenné Z_i upraviť spôsobom uvedeným nižšie

$$Z_i = \xi + \beta x_i + e_i - \beta x_i = \xi + e_i,\ i = 1, \ldots, n. \tag{2.6}$$

Vidíme, že ak hypotéza H_0 platí, tak pomocné premenné $Z_1, Z_2, \ldots, Z_n$ nezávisia od vysvetľujúcich premenných $x_1, x_2, \ldots, x_n$. Teda testovanie signifikantnosti sklonu regresnej priamky sa zmení na test nezávislosti medzi pomocnými premennými $Z_1, \ldots, Z_n$ a $x_1, x_2, \ldots, x_n$. Musíme však vybrať taký test nezávislosti, ktorý nevyžaduje normálne rozdelenie dátových sad. Zaujímavosťou v tomto prístupe je fakt, že pri testovaní parametra β sme vôbec nepotrebovali jeho odhad. Z toho dôvodu sa vyššie opísaný test zaraďuje do skupiny *score testov*.

Theil-Senov odhad, ktorý sme v stručnosti opísali vyššie, sa zaraďuje do skupiny tzv. *complete methods*, pretože využíva všetkých $\binom{n}{2}$ sklonov. Existujú aj tzv. *incomplete methods*, ktoré pracujú s menším počtom sklonov. Môžeme spomenúť, že predpoklady pri využívaní *incomplete methods* sú mierne odlišné. Stále vyžadujeme, aby náhodné chyby $e_1, e_2, \ldots, e_n$ pochádzali zo spojitého rozdelenia a boli nezávislé, no upúšťame už od predpokladu rovnakého rozdelenia. Inými slovami pripúšťame v modeli 2.1 heteroskedasticitu. Navyše však predpokladáme, že spojité rozdelenie náhodných chýb je symetrické s mediánom 0. Tieto metódy nebudeme podrobnejšie v našej práci rozoberať, pretože nie sú súčasťou diplomovej práce. Naším hlavným cieľom je venovať sa modelu s viacerými vysvetľujúcimi premennými. Na druhej strane sme považovali za dôležité spomenúť aspoň základné metódy a postupy v neparametrickej priamkovej regresii, ktoré sú aplikovateľné v najjednoduchšom lineárnom modeli. Hoci metódy opísané v nasledujúcej časti týkajúce sa komplexnejšieho lineárneho modelu sa nám na prvý pohľad budú zdať výrazne odlišné od spôsobov popísaných v tejto časti, ukážeme, že istý súvis medzi nimi existuje.

2.2 Lineárny model s viacerými vysvetľujúcimi premennými

V predchádzajúcej časti bola predmetom nášho záujmu priamková lineárna regresia. V nasledujúcej časti sme sa venovali komplexnejšiemu lineárnemu modelu s viac ako jednou vysvetľujúcou premennou. Cieľom našej diplomovej práce je venovať sa práve takémuto zložitejšiemu modelu. V tejto kapitole sme vychádzali prevažne z knihy [5] prípadne [4]. Model, ktorým sme sa zaoberali je definovaný nasledovne:

$$Y_i = \xi + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i,\ i = 1, \ldots, n. \tag{2.7}$$

Môžeme si všimnúť, že model je tvorený p vysvetľujúcimi premennými $x_1, x_2, \ldots, x_p$, ku ktorým prislúcha p neznámych parametrov $\beta_1, \beta_2, \ldots, \beta_p$, ξ označuje neznámy parameter intercept. Pomocou spomínaných parametrov a premenných a na základe n pozorovaní modelujeme analogicky ako v klasickom parametrickom prístupe z 1. kapitoly závislú resp. vysvetľovanú premennú Y. Chyba v $i-tom$ meraní je označená e_i. Pokiaľ chceme model zapísať maticovo, musíme zaviesť dodatočné označenia. Nech $Y = (Y_1, Y_2, \ldots, Y_n)^T$ predstavuje vektor n závislých premenných, $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T$ je vektor neznámych regresných koeficientov, $e = (e_1, e_2, \ldots, e_n)^T$ je vektor neznámych chýb. Matica X je v tomto prípade tvorená len jednotlivými vysvetľujúcimi premennými bez stĺpca so samými jednotkami

$$X = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{p1} \\ x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n-1} & x_{2,n-1} & \ldots & x_{p,n-1} \\ x_{1,n} & x_{2,n} & \ldots & x_{p,n} \end{bmatrix}. \tag{2.8}$$

Lineárny model 2.7 zapísaný maticovo je zobrazený v nasledujúcom riadku

$$Y = \xi\mathbf{1} + X\beta + e, \; kde \; \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{nx1}. \tag{2.9}$$

Náhodný výber chýb $e_1, e_2, \ldots, e_n$ pochádza z nejakého nám neznámeho symetrického rozdelenia s mediánom 0. Opäť platí, že náhodné chyby musia byť iid. Navyše sa vyžaduje splnenie podmienky $\int_{-\infty}^{\infty} f^2(t)dt < \infty$, kde f je hustota rozdelenia náhodných chýb v našom modeli.

2.2.1 Odhad neznámych regresných parametrov

V prvom rade bol predmetom nášho záujmu odhad neznámych parametrov $\beta_1, \beta_2, \ldots, \beta_p$. Z geometrického pohľadu je myšlienka podobná ako pri klasickej regresii. Snažíme sa nájsť odhad neznámych parametrov $\beta_1, \beta_2, \ldots, \beta_p$ tak, aby vzdialenosť medzi Y a $X\beta$ bola čo najmenšia avšak v tomto prípade neuvažujeme klasickú Euklidovskú normu ale tzv. pseudonormu. Táto pseudonorma je definovaná nasledujúcim spôsobom:

$$\|u\|_{\varphi} = \sum_{i=1}^{n} a(R(u_i))u_i. \tag{2.10}$$

Vo vzťahu 2.10 je výraz $R(u_i)$ označením pre rank premennej $u_i, i = 1, \ldots, n$. Taktiež v tejto definícii vystupuje neznáma funkcia $a(i)$. Táto funkcia sa definuje pomocou tzv. *score function* nasledujúcim spôsobom $a(i) = \varphi(i/(n+1))$. Funkcia $\varphi(u)$, *score function*, definovaná na intervale $[0, 1]$ spĺňa určité predpoklady. Požiadavky kladené na *score function* $\varphi(u)$, sú tieto:

$$(S1) : \varphi(u) \text{ je neklesajuca funkcia na } [0, 1]$$

$$(S2) : -\varphi(u) = \varphi(1-u)$$

$$(S3) : \int_0^1 \varphi(u)du = 0$$

$$(S4) : \int_0^1 \varphi^2(u)du = 1.$$

Vďaka splneniu predošlých štyroch predpokladov kladených na *score function* môžeme v ďalšom texte predpokladať, že aj funkcia $a(i)$ spĺňa dva predpoklady:

$$(A1) : a(i) \text{ je neklesajuca funkcia pre } i = 1, 2, \ldots, n$$

$$(A2) : \sum_{i=1}^{n} a(i) = 0.$$

Vidíme, že z platnosti predpokladu $(S1)$ automaticky vyplýva platnosť predpokladu $(A1)$. Taktiež nie je problém nahliadnuť, že splnená podmienka $(S2)$ zabezpečí splnenie predpokladu $(A2)$. Predpoklady $(S3)$ a $(S4)$ sa týkajú štandardizácie funkcie $\varphi(u)$, ktorej výhody sa prejavia pri testovaní hypotéz resp. pri konštrukcii intervalov spoľahlivosti, ktorým sme sa venovali v ďalších častiach.

Pokiaľ sa lepšie pozrieme na definíciu pseudonormy vektora zo vzťahu 2.10, uvedomíme si, že je to vážený priemer jednotlivých zložiek vektora, pričom váha je úmerná ranku príslušnej zložky. Teda naším cieľom je nájsť minimum funkcie $D(\beta)$, ktorá sa nazýva *Jaeckel's dispersion function* podľa jej autora Louis A. Jaeckel z článku [6] a je definovaná ako

$$D(\beta) = \|Y - X\beta\|_\varphi . \tag{2.11}$$

Na základe uvedených skutočnosti odhad neznámych parametrov $\beta_1, \beta_2, \ldots, \beta_p$ vyzerá nasledovne:

$$\hat{\beta} = arg\min_\beta D(\beta) = arg\min_\beta \|Y - X\beta\|_\varphi = arg\min_\beta \sum_{i=1}^{n} a(R_i(\beta))(Y_i - x_i^T\beta), \tag{2.12}$$

kde $R_i(\beta)$ označuje rank medzi $Y_1 - x_1^T\beta, \ldots, Y_n - x_n^T\beta$ a x_i^T predstavuje $i-ty$ riadok matice X zo vzťahu 2.8.

Môžeme si všimnúť, že na rozdiel od metódy najmenších štvorcov spomínanej v 1. kapitole minimalizáciou funkcie $D(\beta)$ nájdeme odhady len neznámych regresných parametrov bez interceptu. Dôvod je veľmi jednoduchý. Hodnota funkcie $D(\beta)$ je invariantná na hodnotu neznámeho parametra ξ, pretože platí $\|Y - X\beta - \xi\|_\varphi = \|Y - X\beta\|_\varphi$, čo je dôsledok predpokladu $(A2)$. Teda odhad interceptu sa hľadá iným osobitým spôsobom. Hľadanie odhadu regresných koeficientov $\beta_1, \beta_2, \ldots, \beta_p$ z modelu 2.7 minimalizáciou *Jaeckel's dispersion function* je optimalizačný problém, a preto je užitočné, poznať nejaké vlastnosti funkcie $D(\beta)$. Pri uvádzaní a zdôvodňovaní jednotlivých vlastností sme vychádzali z článku [6]. Prvá z nich sa vzťahuje na nezápornosť funkcie $D(\beta)$. Ak označíme $z_i(\beta) = Y_i - x_i^T\beta$ a usporiadame ich podľa veľkosti od najmenšieho po najväčšie $z_{(1)} \leq z_{(2)}, \ldots, \leq z_{(n)}$, môžeme vzťah 2.11 pre predpis funkcie $D(\beta)$ upraviť na výraz

$$D(\beta) = \sum_{i=1}^{n} a(i) z_{(i)}(\beta) = \sum_{i=1}^{n} a(i)(Y_{(i)} - x_{(i)}^T\beta). \tag{2.13}$$

Nech s označuje index prvého kladného prvku z postupnosti $(a(1), ..., a(n))$. Na základe uvedených skutočností a predpokladu $(A2)$ sa zápis 2.13 dá upraviť do tvaru 2.14, ktorý potvrdzuje nezápornosť funkcie $D(\beta)$

$$D(\beta) = \sum_{i=1}^{n} a(i) z_{(i)}(\beta) = \sum_{i=1}^{n} a(i)(z_{(i)}(\beta) - z_{(s)}(\beta)). \tag{2.14}$$

Ďalšie dve dôležité vlastnosti funkcie $D(\beta)$ sú spojitosť a konvexnosť. Vďaka predpokladu $(A1)$ a usporiadaniu $z_{(1)} \leq z_{(2)}, \ldots, \leq z_{(n)}$ dosahuje výraz 2.13 maximálnu hodnotu zpomedzi iných možných hodnôt určených na základe iného usporiadania prvkov $z_{(i)}$. Akonáhle by sme ľubovoľným spôsobom zmenili usporiadanie prvkov $z_{(i)}$, hodnota výrazu 2.13 resp. 2.14 by s istotou nevzrástla. Inými slovami ľubovoľná permutácia indexov $1, \ldots, n$, ktorá $z_{(i)}$ zoradí od najmenšieho až po najväčšie maximalizuje výraz 2.14. Ak množinu všetkých permutácií indexov $1, \ldots, n$ označíme $P = \{p(1), p(2), \ldots, p(n)\}$, ktorých je až $n!$, tak vzťah 2.13 s definíciou *Jaeckel's dispersion function* môžeme prepísať nasledovne:

$$D(\beta) = \sum_{i=1}^{n} a(i) z_{(i)}(\beta) = \max_{p \in P} \sum_{i=1}^{n} a(i) z_{p(i)}. \tag{2.15}$$

Vidíme, že funkcia $\sum_{i=1}^{n} a(i) z_{p(i)} = \sum_{i=1}^{n} a(i)(Y_{p(i)} - x_{p(i)}^T \beta)$ je lineárna v premennej β, a teda je spojitá a konvexná. Z vlastnosti konvexných funkcií vieme, že ak sú funkcie $f_1, f_2, \ldots, f_{n!}$ konvexné, potom aj funkcia $f = max(f_1, f_2, \ldots, f_{n!})$ je konvexná. Analogická úvaha platí aj pre spojitosť funkcie. Teda vidíme, že funkcia $D(\beta)$ je konvexná a spojitá v premennej β.

Pre náš problém je veľmi dôležitá konvexnosť funkcie $D(\beta)$, pretože na nájdenie odhadov neznámych regresných parametrov potrebujeme nájsť minimum tejto funkcie. Konvexnosť funkcie nám zabezpečuje oprávnenosť použitia efektívnych algoritmov na jej minimalizáciu ako sú napríklad rôzne kvázinewtonovské metódy. Problém minimalizácie funkcie $D(\beta)$ konkrétnejšie opíšeme v 3. kapitole.

Na rozdiel od metódy najmenších štvorcov, ktorú sme spomínali v 1. kapitole, minimalizáciou *Jaeckel's dispersion function* nie je možné nájsť exaktné analytické vyjadrenie pre odhad neznámych parametrov $\beta_1, \beta_2, \ldots, \beta_p$ s jedinou výnimkou. Vo väčšine prípadov sú odhady numerické riešenia získané iteračnými metódami. Pokiaľ však uvažujeme priamkový model z časti 2.1, da sa ukázať, že prístup minimalizácie funkcie $D(\beta)$ na nájdenie odhadu parametra zodpovedajúceho sklonu regresnej priamky vedie k exaktnému riešeniu. Navyše tento odhad je len určitou modifikáciou Theil-Senovho odhadu, ktorý sme uviedli v predošlej časti. Theil-Senov odhad zo vzťahu 2.3 je medián z $\binom{n}{2}$ sklonov S_{ij} regresných priamok, ktoré vieme vytvoriť z našich dát. Odhad sklonu regresnej priamky, ktorý získame prístupom minimalizácie funkcie $D(\beta)$, je vážený medián z týchto sklonov. Stručné zdôvodnenie predošlého tvrdenia sme uviedli v nasledujúcej časti, pričom sme opäť vychádzali z článku [6], kde je možné nájsť kompletný dôkaz spomínaného tvrdenia. Najprv si uvedieme definíciu váženého mediánu, ktorá je potrebná pri úvahách v ďalšom texte.

Definícia 2.1. *Nech prvky $x_1, x_2, \ldots, x_n$ sú usporiadané od najmenšieho po najväčší, pričom každému prislúcha váha w_i a platí, že $\sum_{i=1}^{n} w_i = 1$. Vážený medián x_k z prvkov $x_1, x_2, \ldots, x_n$ je prvok, ktorý spĺňa nasledujúce podmienky:*

$$\sum_{i=1}^{k-1} w_i \leq 1/2 \quad a \quad \sum_{i=k+1}^{n} w_i \leq 1/2.$$

Z predpokladu $(S2)$ pre *score function* vyplýva platnosť nasledujúceho vzťahu pre

funkciu $a(k)$,

$$a(n+1-k) = \varphi\left(\frac{n+1-k}{n+1}\right) = -\varphi\left(\frac{k}{n+1}\right) = -a(k). \tag{2.16}$$

Vychádzame z definície funkcie $D(\beta)$ zo vzťahu 2.13. Pozrime sa na jej deriváciu $D'(\beta) = \frac{\mathrm{d}D}{\mathrm{d}\beta} = -\sum_{i=1}^{n} a(i)x_{(i)}$. Vieme, že funkcia $D(\beta)$ je konvexná, teda bude existovať bod minima, v ktorom sa derivácia bude rovnať 0. Derivácia funkcie $D(\beta)$ ako funkcia premennej β je neklesajúca, schodkovitá funkcia. Na prvý pohľad β nevystupuje v predpise $D'(\beta)$, no musíme si uvedomiť, že zmena hodnoty premennej β, môže zmeniť hodnotu $z_i(\beta) = Y_i - x_i^T\beta$, a teda v konečnom dôsledku ovplyvniť usporiadanie prvkov z_i od najmenšieho po najväčší. Pokiaľ dôjde k zmene poradia z_i, hodnota $a(i)$ v $D'(\beta)$ bude násobená už inou vysvetľujúcou premennou $x_{(i)}$ prislúchajúcou k $z_{(i)}$ v novom usporiadaní, a teda dôjde k zmene hodnoty vo funkcii $D'(\beta)$. Pre dostatočne veľké β platí:

$$Y_j - \beta x_j < Y_i - \beta x_i, \ pre\ \forall i, j : x_j > x_i. \tag{2.17}$$

Pre veľmi veľké β platí $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(n)}$ a spolu s predpokladom $(A1)$ zabezpečíme, že hodnota $D'(\beta)$ je maximálna. Analogicky pre dostatočne malé β platí:

$$Y_j - \beta x_j < Y_i - \beta x_i, \ pre\ \forall i, j : x_j < x_i. \tag{2.18}$$

Opätovne pre veľmi malé β platí $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ a pri takomto usporiadaní vysvetľujúcich premenných nadobúda funkcia $D'(\beta)$ minimálnu hodnotu.

Vďaka konvexnosti funkcie $D(\beta)$ môžeme povedať, že maximálna hodnota funkcie $D'(\beta)$ je kladná a minimálna hodnota je záporná. Spomínali sme, že funkcia $D'(\beta)$ je schodkovitá, pričom k zmenám funkčnej hodnoty t. j. k skokom dochádza vtedy, keď sa mení poradie prvkov z_i, čo spôsobuje zmena hodnoty β. Nech $x_j > x_i$, ďalej rovnosť $z_i = Y_i - \beta x_i = Y_j - \beta x_j = z_j$ nastáva práve vtedy ak $\beta = (Y_j - Y_i)/(x_j - x_i) = S_{ij}$. Ak β je tesne pod ako S_{ij}, platí $z_i < z_j$, naopak ak β je tesne nad hodnotou S_{ij}, tak platí opačná nerovnosť t. j. $z_i > z_j$. Vidíme, že došlo k zmene v usporiadaní prvkov $z_1, z_2, \ldots, z_n$, teda nastala aj zmena funkčnej hodnoty vo funkcii $D'(\beta)$. Skok nastal vtedy, keď sa β rovnala jednému z možných sklonov regresnej priamky t. j. keď $\beta = S_{ij}$. Na základe horeuvedených skutočností vieme, že skoky vo funkcii $D'(\beta)$ nastavajú v bodoch totožných s jednotlivým možnými sklonmi regresnej priamky, ktorých je $\binom{n}{2}$.

Taktiež by nás ešte zaujímalo, aká je výška jednotlivých skokov. Predpokladajme, že β je tesne pod S_{ij}, teda $z_i < z_j$. z_i zastáva $k - tu$ pozíciu a z_j $k + 1$ pozíciu v usporiadaní prvkov $z_1, z_2, \ldots, z_n$. Pokiaľ β zvolíme len tesne nad S_{ij}, tak ako sme už spomínali, zmena v usporiadaní nastane len medzi prvkami z_i a z_j, ktoré si navzájom vymenia pozície. Teda vieme určiť výšku skoku v danom $\beta = S_{ij}$, pretože sa zmenia len $k - ty$ a $k + 1$ člen v sume definujúcej funkciu $D'(\beta)$. Zmena funkčnej hodnoty $D'(\beta)$ je nasledovná:

$$-a(k)x_j - a(k+1)x_i - (-a(k)x_i - a(k+1)x_j) = (a(k+1) - a(k))(x_j - x_i). \quad (2.19)$$

Vidíme, že výška skoku v bode S_{ij} je úmerná rozdielu medzi vysvetľujúcimi premennými x_j a x_i.

Ďalej uvažujme náhodnú premennú T, ktorej rozdelenie je dané distribučnou funkciou $G(\beta)$, ktorá je definovaná ako $G(\beta) = \frac{D'(\beta) - \min(D')}{\max(D') - \min(D')}$. Funkcia $G(\beta)$ je podobná s funkciou $D'(\beta)$, teda je tak isto schodkovitá, neklesajúca a skoky nastávajú pre $\beta = S_{ij}$. Akurát ako distribučná funkcia má preškálovaný obor hodnôt na $[0, 1]$. Predpokladajme, že $D(\beta)$ je minimalizovaná v bode β_D, ak $D'(\beta) \leq 0$ pre $\beta < \beta_D$ a súčasne ak $D'(\beta) \geq 0$ pre $\beta > \beta_D$. Teraz definujeme β_D ako určitý kvantil rozdelenia daného distribučnou funkciou $G(\beta)$. To znamená, že platí:

$$\begin{aligned} G(\beta_D) = \frac{-\min(D')}{\max(D') - \min(D')} \Leftrightarrow G(\beta_D) \leq \frac{-\min(D')}{\max(D') - \min(D')} \wedge \\ \wedge\, G(\beta_D) \geq \frac{-\min(D')}{\max(D') - \min(D')} \Leftrightarrow P(T < \beta_D) \leq \frac{-\min(D')}{\max(D') - \min(D')} \wedge \\ \wedge\, 1 - P(T > \beta_D) \geq \frac{-\min(D')}{\max(D') - \min(D')} \Leftrightarrow P(T < \beta_D) \leq \frac{-\min(D')}{\max(D') - \min(D')} \wedge \\ \wedge\, P(T > \beta_D) \leq \frac{\max(D')}{\max(D') - \min(D')}. \end{aligned}$$

Môžeme si všimnúť, že bod β_D, ktorý minimalizuje funkciu $D(\beta)$, je ako kvantil distribučnej funkcie $G(\beta)$ definovaný poslednými dvoma nerovnicami. Navyše ak platí vzťah 2.16 po krátkej úvahe zistíme, že platí $\max(D') = -\min(D')$ a dostaneme

$$\begin{aligned} P(T < \beta_D) \leq \frac{1}{2} \wedge P(T > \beta_D) \leq \frac{1}{2} \Leftrightarrow P(T < \beta_D) \leq \frac{1}{2} \wedge P(T < \beta_D) \geq \frac{1}{2} \Rightarrow \\ \Rightarrow P(T < \beta_D) = P(T > \beta_D) = \frac{1}{2} \Rightarrow G(\beta_D) = \frac{1}{2}, \end{aligned}$$

z čoho jednoznačne vidíme, že bod β_D je mediánom rozdelenia daného distribučnou funkciou $G(\beta)$. Teraz si stačí uvedomiť, že ak sčítame všetky výšky skokov v jednot-

livých bodoch rovných S_{ij} funkcie $G(\beta)$ dostaneme hodnotu 1, čo nám zabezpečuje obor hodnôt distribučnej funkcie. Z uvedených skutočností vyplýva, že β_D je naozaj vážený medián z $\binom{n}{2}$ sklonov S_{ij}, pretože ak každému sklonu S_{ij} priradíme váhu 2.19 zodpovedajúcu výške skoku v tomto bode vo funkcii $G(\beta)$ zistíme, že β_D spĺňa Definíciu 2.1 váženého mediánu. Ak sčítame výšky skokov v $S_{ij} > \beta_D$ resp. $S_{ij} < \beta_D$ v oboch prípadoch nepresiahneme 1/2. Naozaj odhad sklonu regresnej priamky nájdený minimalizáciou *Jaeckel's dispersion function* má exaktné analytické vyjadrenie na rozdiel od zložitejších lineárnych modelov a nie je to nič iné ako vážený medián z $\binom{n}{2}$ sklonov S_{ij}, ktoré sme definovali v časti 2.1 pomocou vzťahu 2.3.

Navyše váha je rovnako ako pri metóde najmenších štvorcov úmerná vzdialenosti medzi vysvetľujúcimi premennými. Ako sme v predošlej vete naznačili, istá podobnosť sa dá vidieť aj s odhadom sklonu regresnej priamky získaného metódou najmenších štvorcov, ktorý je možné, ako sme ukázali v 1. kapitole, upraviť do podoby váženého priemeru zo sklonov S_{ij} a odhad rovnakého parametra vypočítaného prístupom minimalizácie funkcie $D(\beta)$ nie je nič iné, ako sme vyššie ukázali, vážený medián zo sklonov S_{ij}.

Odhad regresných parametrov zodpovedajúcich jednotlivým vysvetľujúcim premenným sme už rozoberali vyššie. Povedali sme, že ho získame, ak vyriešime optimalizačný problém založený na minimalizácii konvexnej funkcie. Teraz sa bližšie pozrieme na odhad samotného interceptu, ktorého nájdenie je oveľa jednoduchšie. Vzhľadom na to, že parametre $\beta_1, \beta_2, ..., \beta_p$ už vieme odhadnúť, využijeme tento fakt pri určení odhadu interceptu ξ. Ako prvý krok si zrátame rezíduá $e_i^* = Y_i - x_i^T\hat{\beta},\ i = 1, \ldots, n$. Na základe predpokladov, ktoré sme kládli na model definovaný vzťahom 2.7 vidíme, že rezíduá $e_1^*, e_2^*, \ldots, e_n^*$ pochádzajú zo symetrického rozdelenia s mediánom rovným ξ. Už nie rovným 0 ako pri chybách $e_1, e_2, \ldots, e_n$. Teda intercept ξ je stred symetrie rozdelenia daného rezíduami $e_i^*,\ i = 1, \ldots, n$. Z toho dôvodu je celkom prirodzené hľadať odhad parametra ξ ako medián z Walsh averages počítaných z $e_1^*, e_2^*, \ldots, e_n^*$ t. j.

$$\hat{\xi} = median\left\{\frac{e_i^* + e_j^*}{2},\ 1 \leq i \leq j \leq n\right\}. \tag{2.20}$$

2.2.2 Test signifikantnosti regresných parametrov

V tejto časti sme spomenuli testovanie hypotézy o významnosti jednotlivých regresných parametrov. Formulácia a značenie je úplne rovnaké aké sme použili pri opise podobného testu v podkapitole 1.1.2. Hypotéza H_0 o signifikantnosti regresných parametrov vyzerá nasledovne:

$$H_0 : \beta_l = 0,\ l \in Q\ vs.\ H_1 : \beta_l \neq 0,\ l \in Q. \tag{2.21}$$

Význam sformulovaných hypotéz povedaný ľudskou rečou môžeme nájsť v spomínanej 1. kapitole, kde sme sa ho snažili podrobne opísať. Na otestovanie hore uvedenej hypotézy potrebujeme skonštruovať testovaciu štatistiku. Prvý krok spočíva v nájdení odhadov regresných parametrov $\hat{\beta}$ z pôvodného modelu 2.7 a taktiež z reštrikčného tzv. *submodelu*, ktoré označíme $\hat{\beta}_{SM}$. Reštrikčný model resp. *submodel* je skonštruovaný za platnosti hypotézy H_0.

Ďalej sa pozrieme na rozdiel funkčných hodnôt v *Jaeckel's dispersion function* medzi pôvodným modelom a *submodelom*, ktorý si označíme $D^* = D(\hat{\beta}_{SM}) - D(\hat{\beta})$. Posledný krok na to, aby sme vedeli danú hypotézu otestovať je odhad neznámeho parametra τ, ktorý štandardizuje testovaciu štatistiku a je definovaný ako $\tau = 1/\gamma$, pričom γ vyzerá nasledovne:

$$\gamma = \int_0^1 \varphi(u)\varphi_f(u)du,\ kde\ \varphi_f(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \tag{2.22}$$

kde $F(u)$ a $f(u)$ je distribučná funkcia a hustota rozdelenia chýb $e_1, e_2, \ldots, e_n$. To, že odhad parametra τ je daný len ako prevrátená hodnota parametra γ a nevystupujú v ňom ďalšie dodatočné konštanty je vďaka štandardizácii *score function* v podobe predpokladov $(S3)$ a $(S4)$. Za predpokladu $(S3)$ a $(S4)$ týkajúcich sa štandardizácie *score function* bude testovacia štatistika HM vyzerať ako

$$HM = \frac{2D^*}{q\hat{\tau}}. \tag{2.23}$$

Idea testu sa nachádza v čitateli testovacej štatistiky HM a je analogická ako v 1. kapitole pri testovaní rovnakej hypotézy avšak pomocou klasickej parametrickej metódy. Pokiaľ bude *submodel* výrazne horší pri popisovaní závislosti medzi vysvetľujúcimi premennými a vysvetľovanou premennou, inak povedané nebude postačujúci na opis tejto

závislosti, tak bude platiť $D(\hat{\beta}_{SM}) \gg D(\hat{\beta})$, čiže $D^* \gg 0$. Teda hypotézu H_0 budeme zamietať pokiaľ bude testovacia štatistika $HM \gg 0$. Ešte potrebujeme zistiť medzu extrémnej kladnosti.

Musíme sa pozrieť akým rozdelením sa testovacia štatistika HM riadi. Podľa [5] má testovacia štatistika asymptoticky Fisherovo-Snedecorovo rozdelenie s q a $n-p-1$ stupňami voľnosti. Teda hypotézu H_0 zamietneme, ak $HM > F_{0,05}(q, n-p-1)$, čo je 5% kritická hodnota spomínaného Fisherovho-Snedecorovho rozdelenia, keďže testujeme zvyčajne na hladine významnosti 5%. Ako je poukázané v knihe [5], niekedy sa namiesto Fisherovho-Snedecorovho rozdelenia po odstránení q z menovateľa testovacej štatistiky HM môže použiť aj Chi-kvadrát rozdelenie s q stupňami voľnosti. Avšak pri testovaní pomocou tohto rozdelenia dochádza k príliš častému zamietaniu hypotézy H_0 predovšetkým pri malých a stredných dátových súboroch, čo je spôsobené ľahšími chvostami rozdelenia $\chi^2(q)$ na rozdiel od $F(q, n-p-1)$.

Môžeme si všimnúť, že nie len odhady ale aj hodnota testovacej štatistiky a v konečnom dôsledku aj výsledok celého testu závisí od toho, aký typ *score function* zvolíme. Výsledok testu závisí taktiež od toho, akým spôsobom odhadneme neznámy parameter τ. Práve odhadu parametra τ potrebného na štandardizáciu testovacej štatistiky HM je venovaná ďalšia časť tejto kapitoly.

2.2.3 Odhad parametra τ

V tejto časti sme postupne popísali štyri metódy odhadnutia parametra τ. Neskôr budeme tieto odhady potrebné na štandardizáciu testovacej štatistiky HM medzi sebou porovnávať na základe výsledkov simulácií týkajúcich sa testovania hypotézy o signifikantnosti regresných parametrov. Ako sme už spomínali aj vyššie, tak aj teraz pomocou označení F a f budeme mať na mysli distribučnú funkciu a hustotu rozdelenia, z ktorého pochádzajú chyby $e_1, e_2, \ldots, e_n$. Jednotlivé spôsoby odhadu parametra τ budeme opisovať stručne a niektoré kroky alebo odvodenia uvádzať nebudeme, pretože by sme sa výrazne odchýlili od cieľov našej práce. Podrobnejšie informácie k jednotlivým spôsobom je možné nájsť v uvedených zdrojoch, z ktorých sme vychádzali pri opise jednotlivých spôsobov.

1. spôsob

Pri popise prvého spôsobu odhadu parametra τ sme vychádzali z článku [9] a z knihy [4] konkrétne z časti 3.7.1. Na začiatku predpokladajme, že *score function* spĺňa predpoklady $(S1)$, $(S2)$, $(S3)$ a $(S4)$ a navyše je diferencovateľná. Taktiež, aby sme zabezpečili ohraničenosť *score function* od 0 po 1, sme v ďalšom texte pracovali s funckiou $\varphi^*(u)$ namiesto $\varphi(u)$, ktorá je definované ako $\varphi^*(u) = \frac{\varphi(u)-\varphi(0)}{\varphi(1)-\varphi(0)}$. Na odhadnutie parametra $\tau = 1/\gamma$, potrebujeme v prvom rade odhadnúť parameter γ zo vzťahu 2.22. Vzhľadom na to, že použijeme značenie $\varphi^*(u)$ vzťah 2.22 sa upraví na vzťah

$$\gamma^* = \int_0^1 \varphi^*(u)\varphi_f(u)du = \frac{\gamma}{\varphi(1)-\varphi(0)}. \tag{2.24}$$

Uvažujme substitúciu $u = F(x)$, tým pádom dostávame

$$\gamma^* = -\int_{-\infty}^{\infty} \varphi^*(F(x))f'(x)dx. \tag{2.25}$$

Vďaka predpokladu diferencovateľnosti *score function* a metóde integrovania *per partes*, v ktorej $u = \varphi^*(F(x))$ a $dv = f'(x)dx$ môžeme vzťah 2.25 upraviť na tvary

$$\gamma^* = \int_{-\infty}^{\infty} (\varphi^*)'(F(x))f^2(x)dx \tag{2.26}$$

$$\gamma^* = \int_{-\infty}^{\infty} f(x)d\varphi^*(F(x)). \tag{2.27}$$

Nech Z_1 a Z_2 sú nezávislé náhodné premenné s distribučnými funkciami $F(x)$ a $\varphi^*(F(x))$, $H(y)$ je distribučná funkcia náhodnej premennej $|Z_1 - Z_2|$. Da sa ukázať, že platí:

$$H(y) = \begin{cases} P\left[|Z_1 - Z_2| \le y\right] = \int_{-\infty}^{\infty} \left[F(z_2+y) - F(z_2-y)\right] d\varphi^*(F(z_2)), & \text{ak } y > 0 \\ 0, & \text{ak } y \le 0. \end{cases} \tag{2.28}$$

Predpokladajme, že $h(y)$ je hustota náhodnej premennej $|Z_1 - Z_2|$. Z teórie pravdepodobnosti vieme, že $H'(y) = h(y)$ a dá sa nahliadnuť, že $h(0) = 2\gamma^*$. Na základe toho zistíme, že na to, aby sme odhadli γ^*, potrebujeme odhadnúť $h(0)$. Ak použijeme substitúciu $t = F(z_2)$, môžeme predpis pre distribučnú funkciu prepísať do tvaru

$$H(y) = \int_0^1 \left[F(F^{-1}(t)+y) - F(F^{-1}(t)-y)\right] d\varphi^*(t). \tag{2.29}$$

Ďalším krokom je nájsť odhad pre $H(y)$. Nech $\hat{F}_n$ označuje empirickú distribučnú funkciu rezíduí $e_i^* = Y_i - x_i^T\hat{\beta},\ i = 1, 2, \ldots, n$, ktoré sme už spomínali v predošlej časti a $\hat{F_n^{-1}} = inf\left\{x : F_n(\hat{x}) \geq t\right\}$ označuje inverznú funkciu. Pokiaľ použijeme označenie $e_{(i)}^*$, budeme tým myslieť, že rezíduum $e_{(i)}^*$ zastáva $i-tu$ pozíciu v usporiadaní týchto rezíduí od najmenšieho po najväčšie. Odhad $H(y)$ vyzerá na základe spomínaných úvah nasledovne:

$$\begin{aligned}\hat{H}_n(y) &= \int_0^1 \left[\hat{F}_n(\hat{F_n^{-1}}(t) + y) - \hat{F}_n(\hat{F_n^{-1}}(t) - y)\right] d\varphi^*(t) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\varphi^*\left(\frac{j}{n}\right) - \varphi^*\left(\frac{j-1}{n}\right)\right) I\left(\left|e_{(i)}^* - e_{(j)}^*\right| \leq y\right).\end{aligned} \tag{2.30}$$

Funkcia $I(.)$ sa nazýva tzv. *indicator function*. Pokiaľ je jej vstup vyhodnotený ako pravdivý výraz, tak vráti 1, inak vráti 0. Autori v článku [9] definovali odhad γ^* ako $\hat{H}_n(t_n)/(2t_n)$, kde $t_n = t_{n,\delta}/\sqrt{n}$. Výraz $t_{n,\delta}$ odhadneme ako δ percentný kvantil distribučnej funkcie $\hat{H}_n(y)$, teda $\hat{t}_{n,\delta} = \hat{H}_n{}^{-1}(\delta)$. Odhad pomocného parametra γ je na základe 2.24 daný výrazom

$$\hat{\gamma} = \frac{(\varphi(1) - \varphi(0))\hat{H}_n(\hat{t}_{n,\delta}/\sqrt{n})}{2\hat{t}_{n,\delta}/\sqrt{n}}. \tag{2.31}$$

2. spôsob

Pri popise 2. spôsobu odhadnutia parametra τ sme ako zdroj poznatkov využívali článok [10]. Opäť predpokladáme, že *score function* spĺňa predpoklady $(S1)$, $(S2)$, $(S3)$ a $(S4)$. Tento spôsob odhadu parametra τ je založený na dĺžke intervalu spoľahlivosti pre intercept ξ. V ďalšom texte pod intervalom spoľahlivosti myslíme 95% interval spoľahlivosti. Definujme si pomocnú funkciu $\varphi^+(u) = \varphi(\frac{u+1}{2})$ a $a^+(k) = \varphi^+(\frac{k}{n+1})$. Autori v článku [10] zaviedli funkciu premennej ξ, ktorej predpis je nasledovný:

$$Z(\xi) = n^{-1/2}\sum_{i=1}^{n} sgn(Y_i - x_i^T\hat{\beta} - \xi)a^+\left(R\left(\left|Y_i - x_i^T\hat{\beta} - \xi\right|\right)\right). \tag{2.32}$$

V článku [10] sa uvádza, že funkcia $Z(\xi)$ je klesajúcou schodíkovou funkciou tzv. *step function*, pričom skoky nastávajú v bodoch rovných $\xi_{ij} = \frac{e_i^* + e_j^*}{2},\ 1 \leq i \leq j \leq n$. Následne autori zaviedli dolnú a hornú hranicu intervalu spoľahlivosti ako $\hat{\xi}_L = inf\left\{\xi \mid Z(\xi) < k\right\}$ a $\hat{\xi}_U = sup\left\{\xi \mid Z(\xi) > -k\right\}$, kde k označuje 97.5% kvantil štandardizovaného normálneho rozdelenia. Odhad parametra τ na základe článku [10]

je rovný

$$\hat{\tau} = \frac{n^{1/2}(\hat{\xi_U} - \hat{\xi_L})}{2k}. \tag{2.33}$$

Môžeme si všimnúť, že hranice intervalu spoľahlivosti sú dve konkrétne hodnoty vybraté z Walsh averages počítaných z $e_1^*, e_2^*, \ldots, e_n^*$. Pripomíname, že intercept ξ je mediánom symetrického rozdelenia, z ktorého pochádzajú $e_1^*, e_2^*, \ldots, e_n^*$. To, ktoré konkrétne vyberieme zpomedzi usporiadaných Walsh averages, je zakotvené v samotných definícia dolnej a hornej hranice. Tento spôsob odhadu parametra τ zohľadňuje v sebe mieru kolísavosti rezíduí $e_1^*, e_2^*, \ldots, e_n^*$, pretože čím viac sú tieto rezíduá rozptýlené, tým väčšia bude aj dĺžka intervalu spoľahlivosti a v konečnom dôsledku sa to prejaví aj v samotnom odhade parametra τ. Podobne ako 1. spôsob odhadu parametra τ, tak aj tento druhý odhad má vlastnosť konzistentnosti, ktorej dôkaz je možné nájsť v článku [10].

3. spôsob

Myšlienka 3. spôsobu odhadu parametra τ je založená na odhade funkcie hustoty f rozdelenia náhodných chýb $e_1, e_2, \ldots, e_n$ pomocou kernelovej metódy. Pri vybudovaní tohto spôsobu odhadu parametra τ sme využívali zdroje [1] a [15]. Opäť predpokladáme, že *score function* spĺňa podmienky $(S1)$, $(S2)$, $(S3)$ a $(S4)$. Rovnako ako v 1. spôsobe potrebujeme, aby *score function* bola diferencovateľná a teraz navyše aj lineárna. Ako sme ukázali v 1. spôsobe, výraz pre parameter γ môžeme upraviť na tvar

$$\gamma = \int_{-\infty}^{\infty} \varphi'(F(x)) f^2(x) dx. \tag{2.34}$$

Avšak vďaka lineárnosti *score function* je zabezpečené, že výraz $\varphi'(F(x)) = c$ je konštanta, ktorá môže ísť pred samotný integrál, v ktorom ostane len funkcionál zodpovedajúci štvorcu funkcie hustoty f. Teda sa snažíme nájsť odhad výrazu

$$\gamma = c \int_{-\infty}^{\infty} f^2(x) dx. \tag{2.35}$$

Odhad neznámeho parametra τ je teda ako sme už viackrát spomínali daný prevrátenou hodnotou parametra γ. Vidíme, že hlavná časť tohto 3. spôsobu spočíva v odhade funkcionálu $\int_{-\infty}^{\infty} f^2(x) dx$. Myšlienka je pritom pomerne priamočiara a spočíva v nájdení odhadu funkcie hustoty pomocou jadrového odhadu hustoty z dát $X_1, X_2, \ldots, X_n$,

ktorý je daný ako $f_n(x) = (nh)^{-1} \sum_{i=1}^{n} K\left[(x - X_i)/h\right]$. V našom prípade ako dátový súbor na odhad hustoty f slúžia opätovne rezíduá $e_i^* = Y_i - x_i^T \hat{\beta},\ i = 1, 2, \ldots, n$. Funkcia $K(u)$ sa nazýva *kernel* alebo *window function*. Existuje veľa rôznych možností ako zvoliť túto funkciu. My sme sa rozhodli vybrať jednu z najpoužívanejších *kernel function*, ktorá zodpovedá hustote štandardizovaného normálneho rozdelenia. Označili sme ju ako g. Na základe informácií z článkov [1] a [15] a po sérii rôznych úprav by sa dalo odvodiť, že odhad parametra γ je daný ako

$$\begin{aligned} \hat{\gamma} &= c \int_{-\infty}^{\infty} f_n^2(x) dx \\ &= c(nh)^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{-\infty}^{\infty} g\left[(x - X_i)h^{-1}\right] g\left[(x - X_j)h^{-1}\right] dx \\ &= c2^{-1/2} h^{-1} n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} g\left[(X_i - X_j)2^{-1/2} h^{-1}\right]. \end{aligned} \tag{2.36}$$

Môžeme si všimnúť, že v odhade parametra γ vystupuje nový parameter h. Tento parameter sa nazýva *bandwidth* a je dôsledkom použitia kernelovej metódy na odhad funkcie hustoty f. Existuje veľa možností a pravidiel, ako zvoliť tento parameter. Na druhej strane často je ťažké povedať, ktorá voľba je optimálna. My sa budeme taktiež zaoberať viacerými možnosťami voľby parametra h. Nakoľko je však táto oblasť veľmi obšírna a mohla by jej byť venovaná samostatná diplomová práca, nerozoberáme vybrané metódy príliš do hĺbky. Navyše to ani nie je cieľom našej diplomovej práce.

Teraz opíšeme stručne štyri spôsoby, ktoré sme sa rozhodli aplikovať pri voľbe parametra h. Prvé dve možnosti patria do skupiny nazývajúcej sa *rule of thumb*, pričom pri ich opise sme vychádzali z knihy [13]. Táto voľba sa ukazuje ako vhodná, ak ako *kernel function* volíme hustotu štandardizovaného normálneho rozdelenia a taktiež ak sa hustota, ktorú chceme odhadnúť, podobá na normálne rozdelenie. Pokiaľ zistíme, že dáta pochádzajúce z hustoty, ktorú odhadujeme, výrazne líši od normálneho rozdelenia, musíme byť obozretný, pretože vtedy voľba parametra h pomocou *rule of thumb* môže spôsobiť pomerne veľké nepresnosti. Najčastejšia voľba odhadu parametra h zo skupiny *rule of thumb* je rovná $h_1 = 0.9An^{-1/5}$ resp. $h_2 = 1.6An^{-1/5}$, kde $A = \min(\sigma, IQR/1.34)$. IQR je medzikvartilové rozpätie t. j. rozdiel medzi 3. a 1. kvartilom v dátach $e_1^*, e_2^*, \ldots, e_n^*$ a σ je štandardná odchýlka z týchto dát. Výhoda spomínaných prvých dvoch možností spočíva v jednoduchosti voľby parametra h.

Najčastejším kritériom optimality pri výbere parametra h je zvoliť h tak, aby sa

minimalizoval výraz $MISE$ (*mean integrated squared error*), ktorý je daný ako

$$MISE(h) = E\left[\int (\hat{f}_h(x) - f(x))^2 dx\right]. \tag{2.37}$$

V konečnom dôsledku toto kritérium platilo aj v prvých dvoch možnostiach určenia parametra h a bude základom aj pri ďalších dvoch spôsoboch. Pri popise odhadov h_3 a h_4 sme vychádzali z článkov [3] a [14]. Často sa namiesto ukazovateľa $MISE(h)$ používa asymptotická verzia $AMISE(h)$, ktorá je daná ako

$$AMISE(h) = (nh)^{-1}R(K) + \frac{1}{4}h^4\sigma_K^4 R(f''), \tag{2.38}$$

kde $R(g) = \int g^2(x)dx$ a $\sigma_g^2 = \int x^2 g(x)dx$. Minimálna hodnota výrazu $AMISE(h)$ sa nadobúda pre h, ktoré je riešením rovnice

$$\frac{\partial}{\partial h}AMISE(h) = -\frac{R(K)}{nh^2} + \sigma_K^4 h^3 R(f'') = 0 \; resp. \; h = \frac{R(K)^{1/5}}{\sigma_K^{4/5} R(f'')^{1/5} n^{1/5}}. \tag{2.39}$$

Voľba h_3 je založená práve na riešení hore uvedenej rovnice, pričom kľúčový krok je odhad neznámeho výrazu $R(f'')$. V článku [14] je presne popísaný postup ako implementovať tento spôsob. Navyše aj v softwari R je zabudovaná funkcia, ktorá priamo počíta h_3. Čo sa týka voľby h_4, ako všetky doteraz spomínané voľby parametra h je založený na minimalizácii $MISE(h)$ resp. $AMISE(h)$. Na rozdiel od predošlého spôsobu voľby parametra h sa pri voľbe h_4 použil na dosiahnutie presnejšej aproximácie výrazu $MISE(h)$ výraz vyššieho rádu, ktorý je analogický ako 2.38, akurát obsahuje ešte ďalší člen. Z toho dôvodu sa treba vysporiadať s ďalším dodatočným odhadom výrazu $R(f''')$, čo je cena za vyššiu presnosť spomínanej aproximácie. Implementácia do softwaru R bola bezproblémová, pretože celý postup je kompletne popísaný v článku [3], z ktorého sme vychádzali. V uvedených zdrojoch sú jednotlivé kroky a zdôvodnenia uvádzané oveľa podrobnejšie, no bližšie ich rozoberať nebudeme, pretože je to mimo cieľov našej práce.

4. spôsob

4. spôsob odhadu neznámeho parametra τ je podobný ako pri 3. spôsobe a je založený na odhade hustoty f. Opätovne aj v tomto spôsobe sme využívali na spočítanie odhadu rezíduá $e_1^*, e_2^*, \ldots, e_n^*$ tak ako pri 3. spôsobe a koniec koncov aj ako pri prvých

dvoch spôsoboch. Pri opise 4. spôsobu sme vychádzali z článku [16]. Predpokladáme, že *score function* spĺňa predpoklady $(S1)$, $(S2)$, $(S3)$ a $(S4)$. Navyše vyžadujeme diferencovateľnosť *score function*, no lineárnosť už nie je potrebná tak ako v 3. spôsobe. Deriváciu *score function* označme φ'. Uvažujme parameter γ daný ako

$$\gamma = \int_{-\infty}^{\infty} f(x) d\varphi(F(x)). \tag{2.40}$$

Na základe spomínaného článku [13] a uvedených skutočností je odhad parametra γ daný ako

$$\hat{\gamma} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi' \left(\frac{i}{n} \right) w_n(e^*_{(i)}, e^*_{(j)}), \tag{2.41}$$

kde $e^*_{(i)}$ sú usporiadané rezíduá $e^*_{(1)} \leq e^*_{(2)} \leq \cdots \leq e^*_{(n)}$ a $w_n(x, y) = h^{-1}K((x - y)/h)$ kde funkcia $K(u)$ je *kernel* alebo *window function*. Rovnako ako v 3. spôsobe uvažujeme ako *kernel function* funkciu, ktorá zodpovedá hustote štandardizovaného normálneho rozdelenia. Taktiež aj tu si môžeme všimnúť prítomnosť parametra h, ktorý predstavuje rovnako ako v predošlom spôsobe parameter *bandwidth*. Voliť ho budeme analogicky pomocou štyroch metód, ktoré sme opísali v 3. spôsobe odhadu parametra τ.

2.2.4 Neparametrické intervaly spoľahlivosti

Pri písaní tejto časti sme vychádzali z knihy [4]. Dá sa ukázať, že odhad regresných parametrov $\hat{\beta}$ daný vzťahom 2.12 sa rovnako ako v klasickej regresii riadi normálnym rozdelením, no v tomto prípade tvrdenie platí asymptoticky

$$\hat{\beta}_l \approx N(\beta_l, \tau^2 (X^T X)^{-1}_{ll}), \; l = 1, 2, \ldots, p. \tag{2.42}$$

Môžeme si všimnúť, že oproti vzťahu 1.9 je v uvedenom vzťahu len minimálny rozdiel, a síce σ^2 sa nahradilo τ^2. Prvok $(X^T X)^{-1}_{ll})$ označuje rovnako ako v prvej kapitole $l - ty$ diagonálny prvok matice $(X^T X)^{-1}$. Interval spoľahlivosti pre parameter β_l, $l = 1, 2, \ldots, p$ je definovaný nasledovne:

$$(\hat{\beta}_l - t(2.5\%)_{n-p-1} \hat{\tau} \sqrt{(X^T X)^{-1}_{ll}}; \hat{\beta}_l + t(2.5\%)_{n-p-1} \hat{\tau} \sqrt{(X^T X)^{-1}_{ll}}) \tag{2.43}$$

Oproti klasickému intervalu spoľahlivosti zo vzťahu 1.12 vidíme, že jediná zmena v predošlom vzťahu 2.43 spočíva v nahradení výrazu S odhadom parametra τ. Nahradenie

kvantilov štandardizovaného normálneho rozdelenia kvantilmi Studentovho rozdelenia nie je však tak priamočiare ako pri parametrickom intervale spoľahlivosti. Vo viacerých prácach ako napríklad [11] prípadne [2], ale aj v samotnej knihe [4] môžeme však nájsť odporúčanie použiť kritické hodnoty Studentovho rozdelenia namiesto pôvodného štandardizovaného normálneho rozdelenia. Vzhľadom na fakt, že vzťah 2.42 platí len asymptoticky, tak aj príslušná spoľahlivosť skonštruovaného intervalu spoľahlivosti je asymptotická.

Existujú však aj iné neparametrické verzie intervalu spoľahlivosti pre neznáme regresné parametre $\beta_1, \beta_2, \ldots, \beta_p$. Jedným z nich je aj interval spoľahlivosti navrhnutý Marekom Omelkom z článku [12]. Povedzme, že nás zaujíma 95% interval spoľahlivosti pre $\beta_l,\ l = 1, 2, \ldots, p$. Ako prvý krok sa definuje funkcia $S_l(t) = \frac{1}{n^{3/2}} \sum_{i=1}^{n} x_{il} R_i(t)$, kde prvok x_{il} označuje $i-ty$ riadok a $l-ty$ stĺpec v matici definovanej vzťahom 2.8 a $R_i(t)$ označuje rank náhodnej premennej $e_i^* - tx_{il}$ medzi $e_1^* - tx_{1l}, e_2^* - tx_{2l}, \ldots, e_n^* - tx_{nl}$. Ďalej sa definuje pomocný výraz c nasledovne:

$$c = \frac{T_l^2 \sqrt{n(X^T X)_{ll}^{-1}}}{\sqrt{12}} \sqrt{\frac{n}{n-p-1}},\ kde T_l^2 = \frac{1}{n} \sum_{i=1}^{n} x_{il}^2. \tag{2.44}$$

Na základe vyššie spomenutých skutočností autor článku [12] definoval 95% interval spoľahlivosti pre $\beta_l,\ l = 1, 2, \ldots, p$ nasledujúcim spôsobom:

$$(\hat{\beta}_l + \delta^-; \hat{\beta}_l + \delta^+), \tag{2.45}$$

kde $\delta^- = sup\,\{t < 0 : S_l(t) \geq ct(2.5\%)_{n-p}\}$ a $\delta^+ = inf\,\{t > 0 : S_l(t) \leq -ct(2.5\%)_{n-p}\}$. Tak ako sa píše v samotnom článku [12] vyššie uvedený interval spoľahlivosti je odvodený pre *Wilcoxon score function*, ktorej definíciu uvedieme v ďalšej kapitole. Veľkou výhodou tohto intervalu spoľahlivosti je fakt, že na rozdiel od intervalu spoľahlivosti zo vzťahu 2.43, netreba odhadovať neznámy parameter τ. Táto skutočnosť je mimoriadne dôležitá, keďže vieme, že nesprávny odhad parametra τ môže výrazne ovplyvniť kvalitu testov alebo intervalov spoľahlivosti. Podobne ako predošlý neparametrický interval spoľahlivosti aj tento funguje len asymptoticky.

Ďalšou zaujímavou alternatívou k už spomínaným intervalom spoľahlivosti daných vzťahmi 1.12, 2.43 a 2.45 sú intervaly spoľahlivosti založené na metóde bootstrap. Myšlienka tejto metódy je veľmi jednoduchá a v dnešnej dobe vyspelej výpočtovej techniky aj často aplikovaná v porovnaní s minulosťou. Bootstrap metóda v lineárnej regresii

spočíva vo vytvorení nových dátových sad, z ktorých odhadneme parametre rovnakého lineárneho modelu ako z pôvodnej dátovej vzorky. Nové dátové sady vyberáme s opakovaním z pôvodných n meraní, pričom väzba medzi $i-tou$ závislou premennou a $i-tymi$ vysvetľujúcimi premennými musí ostať zachovaná. Na to, aby bootsptrap metóda mala zmysel, je nutné vytvoriť pomerne veľa nových dátových súborov. V každej dátovej sade vieme následne zrátať odhad neznámych regresných parametrov či už metódou najmenších štvorcov, alebo neparametrickou metódou minimalizácie *Jaeckel dispersion function.*

Prvé dva intervaly spoľahlivosti skonštruované pomocou bootstrap metódy vychádzajú zo skutočnosti, že odhad neznámeho regresného parametra $\beta_l,\ l = 1, 2, \ldots, p$ má normálne rozdelenie dané vzťahom 1.9 alebo 2.42. Prvý z nich definovaný vzťahom 2.46 využíva ako disperziu odhadu $\hat{\beta}_l$ výberovú disperziu D^* zrátanú z odhadov $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$ prislúchajúcich k R dátovým sadám vytvorených bootstrapom

$$(\hat{\beta}_l - u(2.5\%)\sqrt{D^*}; \hat{\beta}_l + u(2.5\%)\sqrt{D^*}), \tag{2.46}$$

kde u(2.5%) označuje 2.5% kritickú hodnotu štandardizovaného normálneho rozdelenia. Druhý z intervalov spoľahlivosti uvedený vo vzťahu 2.47 zohľadňuje navyše aj možné vychýlenie $\hat{\beta}_l$

$$(2\hat{\beta}_l - \bar{\beta}^* - u(2.5\%)\sqrt{D^*}; 2\hat{\beta}_l - \bar{\beta}^* + u(2.5\%)\sqrt{D^*}),\ kde\ \bar{\beta}^* = \frac{1}{R}\sum_{k=1}^{R}\beta_k^*. \tag{2.47}$$

Ďalšie dva intervaly spoľahlivosti sa nespoliehajú vôbec na to, že odhad $\hat{\beta}_l,\ l = 1, 2, \ldots, p$ má normálne rozdelenie. Percentilový interval spoľahlivosti sa zostrojí veľmi jednoducho, pretože stačí odhady $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$ usporiadať od najmenšieho po najväčšie a ak chceme získať 95% interval spoľahlivosti, tak 2.5% najmenších a najväčších odhadov vyhodíme a najmenšia resp. najväčšia hodnota zo zvyšných odhadov bude tvoriť ľavý resp. pravý okraj percentilového intervalu spoľahlivosti. Tento interval spoľahlivosti vychádza len z bootstrap rozdelenia daného odhadmi $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$, ale má dve nevýhody. Nevyužíva odhad $\hat{\beta}_l,\ l = 1, 2, \ldots, p$ regresného parametra zrátaného z pôvodnej dátovej sady a nezohľadňuje prípadnú šikmosť bootstrap rozdelenia daného odhadmi $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$.

Veľmi podobný predošlému intervalu spoľahlivosti je tzv. BCA (*bias-corrected-accelerated*) percentilový interval spoľahlivosti, ktorý spomínané dva nedostatky klasického percentilového intervalu spoľahlivosti odstraňuje. Z toho dôvodu na zostrojenie BCA intervalu spoľahlivosti treba odhadnúť dva dodatočné parametre. BCA interval spoľahlivosti nie nutne využíva 2.5% a 97.5% kritické hodnoty bootstrap rozdelenia daného odhadmi $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$, ale závisí to práve od odhadov dvoch dodatočných parametrov. Jeden z nich sa nazýva *bias-correction parameter*, ktorý súvisí s počtom odhadov $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$, ktoré sú menšie ako odhad $\hat{\beta}_l,\ l = 1, 2, \ldots, p$ zrátaný z pôvodnej dátovej vzorky. Druhý z parametrov sa nazýva *acceleration parameter* a jeho cieľom je zohľadniť zošikmenie bootstrap rozdelenia daného odhadmi $\beta_l^{1*}, \beta_l^{2*}, \ldots, \beta_l^{R*},\ l = 1, 2, \ldots, p$. Ďalšie podrobnosti o BCA intervale neuvádzame.

3 Simulácie odhadov regresných parametrov

3.1 Základné charakteristiky vykonaných simulácií

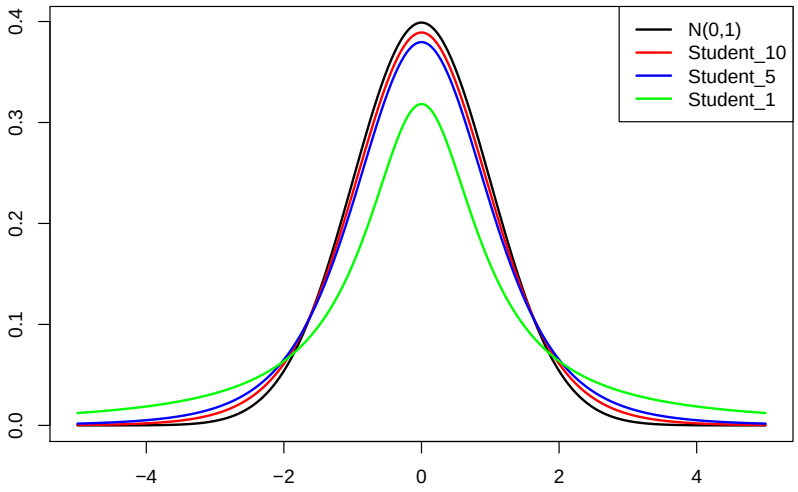
Hlavným cieľom našej diplomovej práce je porovnať parametrické a neparametrické metódy v lineárnej regresii pomocou počítačových simulácií v jazyku R. Ako prvý spôsob porovnania sme sa rozhodli vykonať simulácie pri hľadaní odhadov neznámych regresných parametrov. Metóda najmenších štvorcov zastupuje klasickú parametrickú metódu a neparametrická metóda je reprezentovaná minimalizáciou konvexnej funkcie $D(\beta)$, ktorú sme definovali v 2. kapitole pomocou vzťahu 2.11. Ako sme už spomínali, vďaka konvexnosti funkcie $D(\beta)$ môžeme na nájdenie jej minima použiť efektívne algoritmy. My sme sa rozhodli použiť kvázinewtonovskú BFGS metódu.

Taktiež sa nám podarilo nájsť balík *Rfit* pre R software, ktorý sa zaoberá neparametrickými metódami v lineárnej regresii. Popis balíka a jeho základných funkcií je k dispozícii v článku [8]. Funkcia na rátanie odhadov neznámych parametrov z balíka *Rfit* taktiež využíva kvázinewtonovské metódy, pričom kombinuje BFGS a CG metódy. Do pozornosti dávame, že balík *Rfit* nám slúžil prioritne len na kontrolu našich vlastných metód. Všetky použité metódy sme si naprogramovali samostatne, pretože sme nechceli vykradnúť balík *Rfit*. Navyše sme sa rozhodli porovnať odhady nájdené pomocou našej implemantácie základnej BFGS metódy a komplikovanejšej metódy pozostávajúcej z kombinácie dvoch kvázinewtonovských metód z balíka *Rfit*.

Jedným z prínosov našej práce sú samotné simulácie nielen v tejto časti, ale i v nasledujúcich častiach, pretože sú v porovnaní s inými prácami komplexnejšie. Na rozdiel od iných štúdií sme pri našich simuláciách menili rozdelenie náhodných chýb, počet pozorovaní n a taktiež aj rôzne *score function*. Tak sme mohli pozorovať vplyv spomínaných zmien na výsledky parametrických a neparametrických metód a zahrnúť získané poznatky aj v samotnom porovnaní oboch prístupov.

Čo sa týka rozdelenia náhodných chýb $e_1, e_2, \ldots, e_n$ začínali sme so štandardizovaným normálnym rozdelením. Postupne sme rozdelenie nahradzovali rozdeleniami s ťažšími chvostami, aby sme zabezpečili častejší výskyt extrémnych hodnôt. Cieľom bolo zistiť, či neparametrické metódy budú fungovať lepšie ako parametrická metóda najmenších štvorcov, hoci vieme, že metóda najmenších štvorcov nevyžaduje predpo-

klad normálneho rozdelenia chýb. Teda namiesto normálneho rozdelenia sme zvolili symetrické Studentovo rozdelenie s 10, 5 a 1 stupňom voľnosti. Postupne sme znižovali počet stupňov voľnosti, čím sme zabezpečili, že chvosty týchto rozdelení boli ťažšie a ťažšie a výskyt outlierov častejší. Porovnanie jednotlivých rozdelení uvádzame na Obr. 3.1.



Obr. 3.1: Porovnanie rozdelení náhodných chýb

Pri simuláciách odhadov neznámych parametrov sme taktiež menili počet pozorovaní n v našom modeli. Najprv sme uvažovali $n = 20$ reprezentujúce malý dátový súbor, potom $n = 80$ zastupujúce stredne veľkú dátovú sadu a nakoniec $n = 200$ predstavujúce najväčší počet meraní. Cieľom bolo odsledovať vplyv veľkosti dátového súboru na výsledky jednotlivých metód pri rôznych pravdepodobnostných rozdeleniach. Vieme, že funkcia $D(\beta)$ je ovplyvnená voľbou *score function*. Z toho dôvodu sme usúdili, že bude zaujímavé skúmať vplyv rôznych *score function* na kvalitu odhadov získaných spomínanými metódami. My sme si zvolili tri *score function*, o ktorých sme sa dozvedeli z článku [8]. Pripomíname, že všetky nižšie uvedené *score function* $\varphi(u)$ sú definované pre $u \in [0, 1]$.

1. Wilcoxon score function

$$\varphi(u) = \sqrt{12}(u - \frac{1}{2}) \tag{3.1}$$

Táto *score function* patrí medzi najpoužívanejšie, pretože je diferencovateľná a lineárna, čo sa ukazuje ako výhoda predovšetkým pri odhade neznámeho parametra τ.

2. Kvantil score function

$$\varphi(u) = \phi(u)^{-1} \tag{3.2}$$

Funkcia $\phi(u)$ reprezentuje distribučnú funkciu štandardizovaného normálneho rozdelenia t. j. kvantilova funkcia je jej inverznou funkciou. Niekedy sa zvykne táto funkcia nazývať ako *probit function*.

3. Bent score function

$$\varphi(u) = \begin{cases} -\sqrt{3/2}, & \text{ak } u < 1/4 \\ \sqrt{3/2}, & \text{ak } u > 3/4 \\ \sqrt{3/2}(4u-2), & \text{ak } 1/4 \geq u \leq 3/4. \end{cases} \tag{3.3}$$

Môžeme si všimnúť, že všetky 3 uvedené *score function* spĺňajú predpoklady $(S1)$, $(S2)$, $(S3)$ a $(S4)$ uvedené v 2.kapitole. Podmienky $(S3)$ a $(S4)$ sa týkajú štandardizácie *score function* a ich význam oceníme v ďalších častiach našej práce. Ak by sme túto štandardizáciu nevykonali, museli by sme do rátania testovacej štatistiky resp. odhadovania neznámeho parametra τ zakomponovať dodatočnú konštantu závislú na konkrétnej *score function*, čo by mohlo byť pri rôznych *score function* zbytočne mätúce.

Menením všetkých spomínaných nastavení v simuláciách sme sa snažili o komplexne porovnanie klasického t. j. parametrického prístupu s neparametrickým, čo považujeme za prínos našej práce. Simulácie sme vykonávali na umelých dátach. Bez ujmy na všeobecnosti sme uvažovali lineárny regresný model s interceptom a s piatimi vysvetľujúcimi premennými. Najprv sme náhodne vygenerovali šesť hodnôt neznámych regresných parametrov z postupnosti čísel od -5 po 5 s krokom 0.01. Ako druhý krok sme si vygenerovali z rovnomerného rozdelenia na intervale $[0, 5]$ vektor dĺžky np, ktorý sme následne postupne poukladali do matice X_{nxp}, ktorej stĺpce reprezentovali jednotlivé vysvetľujúce premenné. Náhodné chyby sme generovali z už spomínaných štyroch rozdelení. Vďaka takto umelo vygenerovaným dátam sme si zrátali realizácie závislej premennej Y.

Parametrické rovnako ako aj neparametrické metódy nám na základe zadaných vstupných dát v podobe Y a X zrátali odhady neznámych regresných parametrov, ktoré sme následne porovnávali s pevne zvolenými resp. vygenerovanými parametrami

zo začiatku. Zaujímali sme sa o štatisticky ukazovateľ MSE (*mean squared error*). Teda v každej simulácií pri každom regresnom parametri sme sa pozreli na štvorec odchýlky od pevne zvolenej hodnoty daného parametra a potom sme zrátali priemernú odchýlku. Čím menšie MSE vykazuje metóda pri daných podmienkach, tým je presnejšia a kvalitnejšia. Pri náhodnom generovaní umelých dát sme používali príkaz *set.seed()* v softwari R na zabezpečenie reprodukovateľnosti náhodných postupnosti, ktoré tvoria naše umelé dáta. Z toho dôvodu sme mohli vyhodnocovať výsledky simulácií nielen z globálneho pohľadu, ale aj v rámci každej simulácie. Na zaistenie výpovednej hodnoty výsledkov simulácií sme zvolili počet simulácií rovný 10 000.

3.2 Počítanie funkčných hodnôt *score function*

Pri hľadaní minima funkcie $D(\beta)$ sme potrebovali vedieť počítať funkčnú hodnotu tejto funkcie pre zadané β. Naprogramovali sme si teda na to veľmi jednoducho funkciu, ktorá nám vrátila funkčnú hodnotu $D(\beta)$ a porovnali sme to s funkciou *disp()* z balíka *Rfit*. Bohužiaľ zistili sme, že výsledky sa nám nezhodujú. Začali sme teda pátrať po príčine. Ako prvé sme si otvorili popis funkcie *disp()* v manuáli k balíku *Rfit*. Našli sme tam popis samotnej implementácie funkcie *disp()*, ktorej výstup sa zhodoval s našou implementáciou. Podľa manuálu by funkcia *disp()* mala dávať rovnaké výsledky ako naša funkcia, no realita bola iná. Z toho dôvodu sme sa pozreli na zdrojový kód funkcie *disp()*. Všimli sme si malý rozdiel v kóde pri rátaní *score function* oproti popisu popísanej v manuáli. Výsledok získaný funkciou *disp()* sa líši od našej funkcie rovnako ako aj od funkcie, ktorá je opísaná v manuáli k balíku *Rfit*, pretože ráta *score function* iným spôsobom pomocou funkcie *getScores()*, o ktorej sme doposiaľ netušili na čo slúži.

Ďalším krokom bolo zistiť akým spôsobom funguje spomínaná funkcia *getScores()*. Bohužiaľ to už bolo trochu zložitejšie, pretože získať kód samotnej funkcie nebolo možné. Z webovej stránky k balíku *Rfit* sme si stiahli dodatočné súbory a podarilo sa nám nájsť zdrojový kód funkcie *getScores()*. Zistili sme, že funkcia *getScores()* počíta hodnoty *score function* klasicky ako naša funkcia, ale následne robí ešte dodatočnú úpravu. Úprava spočíva v tom, že od týchto hodnôt odráta ich priemer a následne sú predelené smerodajnou odchýlkou. Teda ide len o nejaké preškálovanie resp. štandardizovanie samotnej *score function*.

Podstatnejšie však bolo, že sme nevedeli prísť na dôvod takejto úpravy. Z toho dôvodu sme sa rozhodli napísať jednému z autorov balíka *Rfit*. Napísali sme profesorovi Josephovi W. McKeanovi. Boli sme veľmi prekvapení, keď sme na druhý deň našli jeho odpoveď s vysvetlením. Najprv si však pripomeňme definíciu funkcie $a(i) = \varphi(i/(n+1))$ z 2. kapitoly. Dôvod dodatočnej úpravy pri rátaní *score function* bol taký, aby bolo zabezpečené splnenie podmienky

$$\frac{1}{n+1}\sum_{i=1}^{n} a(i)^2 = \frac{1}{n+1}\sum_{i=1}^{n} \varphi(\frac{i}{n+1})^2 = 1, \tag{3.4}$$

čo je horný integrálny súčet integrálu $\int_0^1 \varphi^2(u)du$ vystupujúceho v predpoklade $(S4)$. Autori spomínaným preškálovaním resp. dodatočnou štandardizáciou pôvodnej *score function* zabezpečili, že pre ľubovoľné n reprezentujúce počet meraní v dátovej sade platí, že aproximácia integrálu v podobe horného integrálneho súčtu je rovná 1. Je jasné, že pre väčšie n je zmena v zmysle preškálovania pôvodnej *score function* menšia t. j. hodnoty upravenej *score function* sa menej líšia od pôvodných neupravených hodnôt, pretože horný integrálny súčet je presnejšou aproximáciou presného integrálu. Analogická úvaha platí pre menšie n. Samozrejme po tejto úprave už nie je splnená podmienka $(S4)$, pretože presný integrál $\int_0^1 \varphi^2(u)du$ rátaný z upravenej *score function* nie je rovný 1. Na druhej strane máme referenciu od samotného autora balíka *Rfit* profesora J.W. Mckeana, ktorý má v tejto oblasti neparametrickej štatistiky pomerne veľa publikácií a skúsenosti, preto sme si nechali poradiť. Rozhodli sme sa, že zakomponujeme túto úpravu v rátaní *score function* aj do našich metód a vyžadovali sme, aby bol splnený vzťah 3.4. V tomto duchu sme upravili aj predpoklad $(S4)$, v ktorom sme presný integrál nahradili jeho aproximáciou pomocou horného integrálneho súčtu. Spomínali sme, že štandardizácia *score function* v podobe splnenia predpokladov $(S3)$ a $(S4)$ sa ukáže prínosná hlavne pri výpočtoch, v ktorých sa bude vyskytovať odhad parametra τ. Pokiaľ by sme túto štandardizáciu nevykonali, museli by sme do všetkých týchto výpočtov zakomponovať dodatočnú konštantu rovnú odmocnine z hodnoty integrálu resp. jeho aproximácie v podobe horného integrálneho súčtu z predpokladu $(S4)$, čo by bolo komplikovanejšie, neprehľadnejšie a náchylnejšie na vznik zbytočných chýb.

V konečnom dôsledku balík *Rfit* a jeho spôsob rátania *score function* funguje správne, ale v manuáli pri popise funkcií využívajúcich hodnoty *score function* sa

o tejto dodatočnej úprave nič nespomína. Popis funkcií z manuálu *Rfit*, v ktorých je potrebné rátať hodnoty *score function*, nekorešponduje s reálnou implementáciou, čo je dosť mätúce. Aj tento fakt v podobe zistenia určitej nezhody medzi manuálom k balíku *Rfit* a reálnou implementáciou môžeme považovať za určitý prínos našej práce, pretože profesor McKean uznal, že manuál a popis niektorých funkcií je zastaralý a potrebuje úpravu. Zistili sme, že v balíku *Rfit* dochádzalo doteraz aspoň raz za rok k nejakej aktualizácii. Je možné, že aj naše pripomienky sa v nejakej budúcej úprave prejavia a popis niektorých funkcií sa aktualizuje, čo prispeje k vyvarovaniu sa podobným nezrovnalostiam, na aké sme narazili my.

Z toho dôvodu sa nám vyplatilo implementovať si vlastné funkcie, pretože inak by sme na spomínanú nezhodu pravdepodobne neprišli. Funkcie z balíka *Rfit* by rátali *score function* pomocou vyššie opísaného preškálovania a je možné, že by sme podrobnejšie neskúmali ich postup rátania, pretože by sme na to nemali žiaden dôvod. Celý čas by sme boli v tom, že *score function* sú počítané klasicky bez akejkoľvek úpravy a len na základe ich predpisov, ktoré sme uviedli vyššie, čo by nebolo správne.

3.3 Výsledky simulácií odhadov regresných parametrov

V tejto časti sme uviedli výsledky simulácií opísaných v predošlej časti, pomocou ktorých sme porovnávali parametrický a neparametrický prístup pri odhadovaní neznámych parametrov v lineárnej regresii. Kvalitu jednotlivých metód sme posudzovali na základe priemernej odchýlky MSE v každej simulácii, ktorú sme určili ako priemer z rozdielu umocneného na druhú medzi odhadnutými a pevne zvolenými regresnými parametrami. Takto sme pri každej metóde získali 10 000 rôznych MSE, z ktorých sme zrátali priemernú hodnotu, taktiež maximálne a minimálne MSE. Vo výsledkových tabuľkách sme však namiesto priemernej hodnoty z 10 000 MSE uviedli relatívnu efektívnosť danej metódy voči metóde najmenších štvorcov. To znamená, že na získanie relatívnej efektívnosti pre $m-t$ú metódu sme dali do pomeru priemerné MSE metódy najmenších štvorcov a priemerné MSE $m-tej$ metódy. Pri opise výsledkov sme sa však častejšie odvolávali na priemerné MSE. Je úplne jedno, na ktorý ukazovateľ sa pri hodnotení výsledkov odvolávame, pretože oba sú navzájom prepojené. Metódu, ktorá dosiahla relatívnu efektívnosť vyššiu ako 1, sme považovali za presnejšiu a lepšiu v po-

rovnaní s klasickou metódou najmenších štvorcov. Metódy s relatívnou efektívnosťou pod 1 patrili prirodzene k menej presným metódam voči metóde najmenších štvorcov. V jednoduchosti povedané, čím menší priemer z 10 000 MSE metóda dosiahla, tým sme ju považovali za efektívnejšiu, t. j. presnejšiu a kvalitnejšiu.

Spomínali sme, že pri generovaní umelých dát sme využívali funkciu *set.seed()*, aby sme zabezpečili, že pre dané náhodné rozdelenie a dané n budeme merať kvalitu jednotlivých metód na rovnakých dátach. Vďaka tomu sme mohli vyhodnocovať všetky metódy v rámci každej simulácie a nielen z globálneho hľadiska. Pre dané n, dané rozdelenie náhodných chýb a danú *score function* sme zisťovali v každej simulácii, ktorá metóda dosiahla najnižšiu odchýlku MSE. Tieto výsledky sme pri jednotlivých metódach zaznamenali do premennej s názvom *víťazstvo 1*. V rámci neparametrických metód pri danom n a danom rozdelení náhodných chýb sme v premennej *víťazstvo 2* zisťovali, pri ktorej *score function* dosahuje daná neparametrická metóda najmenšie MSE v každej z 10 000 simulácií. Teda v premenných *víťazstvo 1* a *víťazstvo 2* sú uložené percentuálne počty najmenších MSE, ktoré dosiahli jednotlivé metódy po 10 000 simuláciách.

Výsledky sme postupne prezentovali v štyroch tabuľkách, ktoré zodpovedajú štyrom rôznym rozdeleniam náhodných chýb. V tabuľke Tab.3.1 sme zobrazili výsledky simulácií pre štandardizované normálne rozdelenie. Parametrickú metódu najmenších štvorcov sme vo výsledkoch označovali LSE, neparametrická metóda odhadnutia neznámych parametrov na základe prístupu minimalizácie funkcie $D(\beta)$ z balíka *Rfit* je označená NON_1 a naša implementácia tejto neparametrickej metódy založená na analogickom prístupe NON_2. Rozdiel medzi našou implemntáciou neparametrickej metódy a metódou z balíka *Rfit* sme už opísali v predošlej časti. Keďže sme uvažovali náhodné chyby zo štandardizovaného normálneho rozdelenia, tak na základe výsledkov našich simulácií klasický parametrický prístup je lepší a presnejší ako neparametrický. Pre všetky druhy rozsahov dátovej vzorky platí, že priemerné MSE získané metódou najmenších štvorcov je menšie ako priemerné MSE získané neparametrickou metódou z balíka *Rfit* alebo našou implementáciou neparametrickej metódy, čo vidíme aj v príslušných relatívnych efektivitách. Tento výsledok sa potvrdzuje pri všetkých 3 *score function*, ktoré sme v neparametrických metódach menili. Taktiež platí, že priemerné

MSE sa zmenšuje pri rastúcom n vo všetkých metódach, čo je celkom prirodzené. V regresnom modeli uvažujeme viacero dát, metódy majú k dispozícii viacej informácií, a preto sú schopné skonštruovať presnejšie odhady regresných parametrov.

N(0,1)				SCORE FUNCTION		
n	Metóda	Štatistický ukazovateľ		Wilcoxon	Qnorm	Bent
20	LSE	MSE	rel. efe.	1		
			min	0,00055157		
			max	3,863064057		
		víťazstvo 1		**45,94%**	**42,88%**	**51,91%**
	NON_1	MSE	rel. efe.	0,92225338	0,948806253	0,807524845
			min	0,001428673	0,001088423	0,000904763
			max	4,898872026	4,518032717	6,094773561
		víťazstvo 1		**31,89%**	**35,48%**	**26,49%**
		víťazstvo 2		*20,04%*	*44,49%*	*35,47%*
	NON_2	MSE	rel. efe.	0,945130549	0,969810131	0,841606634
			min	0,000708673	0,001777693	0,001612808
			max	4,65931438	4,556810319	5,975299849
		víťazstvo 1		**22,17%**	**21,64%**	**21,60%**
		víťazstvo 2		*19,08%*	*44,85%*	*36,07%*
80	LSE	MSE	rel. efe.	1		
			min	0,000277107		
			max	0,762152442		
		víťazstvo 1		**53,80%**	**49,71%**	**58,54%**
	NON_1	MSE	rel. efe.	0,940304443	0,983105644	0,81197957
			min	0,00019866	0,000192324	0,000166353
			max	0,928672569	0,74478375	1,134066895
		víťazstvo 1		**25,05%**	**28,87%**	**21,26%**
		víťazstvo 2		*16,21%*	*49,86%*	*33,93%*
	NON_2	MSE	rel. efe.	0,940378158	0,984854641	0,80926008
			min	0,000173145	0,000213124	0,000186713
			max	0,928412024	0,74925473	1,123528231
		víťazstvo 1		**21,15%**	**21,42%**	**20,20%**
		víťazstvo 2		*15,80%*	*50,16%*	*34,04%*
200	LSE	MSE	rel. efe.	1		
			min	6,08648E-05		
			max	0,1726541		
		víťazstvo 1		**54,31%**	**50,80%**	**59,88%**
	NON_1	MSE	rel. efe.	0,953260719	0,989429156	0,839531051
			min	4,20122E-05	3,9596E-05	0,00014875
			max	0,1904944	0,176825003	0,272957305
		víťazstvo 1		**21,86%**	**23,52%**	**19,24%**
		víťazstvo 2		*16,34%*	*50,38%*	*33,28%*
	NON_2	MSE	rel. efe.	0,953111115	0,98954399	0,838820343
			min	4,03351E-05	4,04948E-05	0,000148448
			max	0,1906082	0,176426274	0,267718675
		víťazstvo 1		**23,83%**	**25,68%**	**20,88%**
		víťazstvo 2		*16,56%*	*50,43%*	*33,01%*

Tab 3.1: Simulácie odhadov regresneých parametrov pri N(0,1)

Naše výsledky potvrdili to, čo sme mohli očakávať. Ak máme chyby z normálneho rozdelenia, tak metóda najmenších štvorcov sa ukazuje ako najlepšia. Dokazuje to aj fakt, ktorý sa odzrkadľuje v premennej *víťazstvo 1*. Metóda LSE zaznamenala vo väčšine prípadov nadpolovičný počet zo všetkých 10 000 simulácií, v ktorých dosiahla najmenšie MSE v danej simulácií v porovnaní s neparametrickými metódami. Na druhej strane ani neparametrické metódy neboli výrazne horšie. Rozdiel v priemernom MSE nie je vôbec veľký. Môžeme sa pozrieť na výsledky neparametrických metód vzhľadom na rôzne *score function*. Z Tab.3.1 vidíme, že pri všetkých troch rôznych n dosahujú

neparametrické metódy najmenšie priemerné MSE t. j. najlepší výsledok, ak použijeme *Kvantil score function*. Za ňou nasleduje *Wilcoxon score function* a najväčšie MSE zaznamenali neparametrické metódy pri aplikovaní *Bent score funtion*. Uvedené tvrdenie opätovne platí bez ohľadu na to, či sme odhady rátali neparametrickou metódou z balíka *Rfit* alebo pomocou metódy NON_2. Hoci v jednotlivých tabuľkách neuvádzame explicitne priemernú hodnotu zo všetkých 10 000 získaných MSE v jednotlivých metódach, vyššie opísané závery sa prirodzene preniesli aj do relatívnych efektivít.

V texte sme rozlišovali medzi neparametrickou metódou z balíka *Rfit* a našou implementáciou tejto metódy, hoci metódy sú analogické a drobný rozdiel je v optimalizačnom algoritme, ktorý sme už opísali. Z Tab.3.1 pozorujeme, že výsledky metód NON_1 a NON_2 sú veľmi podobné a rozdiely sú minimálne. Hoci pre $n = 20$ je naša metóda pri všetkých 3 *score function* o niečo lepšia ako metóda NON_1 z balíka *Rfit* z hľadiska priemerného MSE. Pre $n = 20$ a $n = 80$ a pre všetky *score function* zaznamenala metóda NON_1 z balíka *Rfit* väčší počet najmenších MSE zo všetkých 10 000 simulácií ako naša metóda, čo vidíme v premennej *víťazstvo 1*. Pre $n = 200$ sa situácia otočila a naša metóda dosiahla väčší počet minimálnych MSE. Na základe výsledkov z premennej *víťazstvo 2* vidíme pri všetkých n, že pri neparametrických metódach sa ukazuje ako najlepšie použitie *Kvantil score function*. Rovnaký záver sme získali aj keď sme neparametrické metódy porovnávali na základe priemerného MSE. Na druhej strane pri tomto porovnaní sa ukázalo ako druhé najlepšie použitie *Wilcoxon score function* a až nakoniec *Bent score funcion*. Pri hodnotí neparametrických metód podľa premennej *víťazstvo 2* sa poradie spomínaných *score function* na 2. a 3. mieste vymenilo.

V Tab.3.2 sme zobrazili výsledky simulácií pre náhodné chyby pochádzajúce zo Studentovho rozdelenia s 10 stupňami voľnosti, ktoré má o niečo ťažšie chvosty ako štandardizované normálne rozdelenie. Môžeme badať, že pri $n = 20$ je priemerné MSE nižšie pri metóde LSE ako pri neparametrických metódach NON_1 a NON_2. Avšak pri väčších dátových súboroch t. j. pre $n = 80$ a $n = 200$ vidíme, že neparametrické metódy sú presnejšie ako parametrická metóda najmenších štvorcov, ak volíme *Wilcoxon score function* alebo *Kvantil score function*. Pri voľbe *Bent score function* má metóda LSE stále menšie priemerné MSE ako neparametrické metódy. Čo sa týka výsledkov

v premennej *víťazstvo 1* vo všetkých prípadoch dosahuje najlepšie hodnotenie metóda najmenších štvorcov. Pri neparametrických metódach s $n = 20$ sa ako najvhodnejšia voľba *score function* ukazuje *Kvantil score function*, pretože dosahuje najmenšie priemerné MSE a taktiež najlepšie výsledky v premennej *víťazstvo 2*. Pre $n = 80$ a $n = 200$ sa na základe priemerného MSE ako najlepšia voľba ukazuje *Wilcoxon score function*, hoci podľa premennej *víťazstvo 2* je stále na prvom mieste *Kvantil score function*, za ňou *Bent score function* a až na poslednom mieste je *Wilcoxon score function*.

Student_10				SCORE FUNCTION		
n	Metóda	Štatistický ukazovateľ		Wilcoxon	Qnorm	Bent
20	LSE	MSE	rel. efe.	1		
			min	0,000390505		
			max	5,2085596		
		víťazstvo 1		**43,51%**	**42,00%**	**48,36%**
	NON_1	MSE	rel. efe.	0,965162845	0,975154919	0,864097889
			min	0,001854388	0,000307413	0,002069391
			max	6,305815124	5,418575054	7,833721907
		víťazstvo 1		**34,00%**	**35,35%**	**28,51%**
		víťazstvo 2		*21,40%*	*41,85%*	*36,75%*
	NON_2	MSE	rel. efe.	0,980706181	0,993171861	0,897869254
			min	0,00100189	0,000686646	0,001748658
			max	6,268108458	5,281153218	8,199522832
		víťazstvo 1		**22,49%**	**22,65%**	**23,13%**
		víťazstvo 2		*19,25%*	*42,01%*	*38,74%*
80	LSE	MSE	rel. efe.	1		
			min	0,000145415		
			max	1,261681709		
		víťazstvo 1		**46,77%**	**45,53%**	**51,82%**
	NON_1	MSE	rel. efe.	1,030748633	1,024579271	0,940327211
			min	0,000225038	8,03374E-05	0,000204595
			max	1,088146401	1,304600946	1,187682095
		víťazstvo 1		**26,95%**	**30,03%**	**23,54%**
		víťazstvo 2		*17,94%*	*44,00%*	*38,06%*
	NON_2	MSE	rel. efe.	1,031200944	1,025037133	0,938361711
			min	0,000225976	9,29246E-05	0,000188741
			max	1,087961329	1,262682591	1,203908694
		víťazstvo 1		**26,28%**	**24,44%**	**24,64%**
		víťazstvo 2		*18,07%*	*43,96%*	*37,97%*
200	LSE	MSE	rel. efe.	1		
			min	5,40123E-05		
			max	0,250076467		
		víťazstvo 1		**46,27%**	**44,56%**	**51,21%**
	NON_1	MSE	rel. efe.	1,046325073	1,031776141	0,968421702
			min	9,37048E-05	0,00012419	0,000109309
			max	0,20878627	0,231749696	0,245590777
		víťazstvo 1		**26,20%**	**26,62%**	**23,67%**
		víťazstvo 2		*17,49%*	*43,51%*	*39,00%*
	NON_2	MSE	rel. efe.	1,046203183	1,031699986	0,967854323
			min	9,3387E-05	0,000118844	0,000117429
			max	0,208582715	0,23167581	0,245938631
		víťazstvo 1		**27,53%**	**28,82%**	**25,12%**
		víťazstvo 2		*17,58%*	*43,40%*	*39,02%*

Tab 3.2: Simulácie odhadov regresneých parametrov pri rozdelení t_{10}

V Tab.3.3 dokumentujeme výsledky našich simulácií pre Studentovo rozdelenie s menším počtom stupňov voľnosti v porovnaní s predošlým prípadom. Konkrétne sa jedná o 5 stupňov voľnosti, čo nám zabezpečuje ešte častejší výskyt extrémnych hodnôt v porovnaní s predchádzajúcimi dvomi rozdeleniami náhodných chýb. V tomto prípade

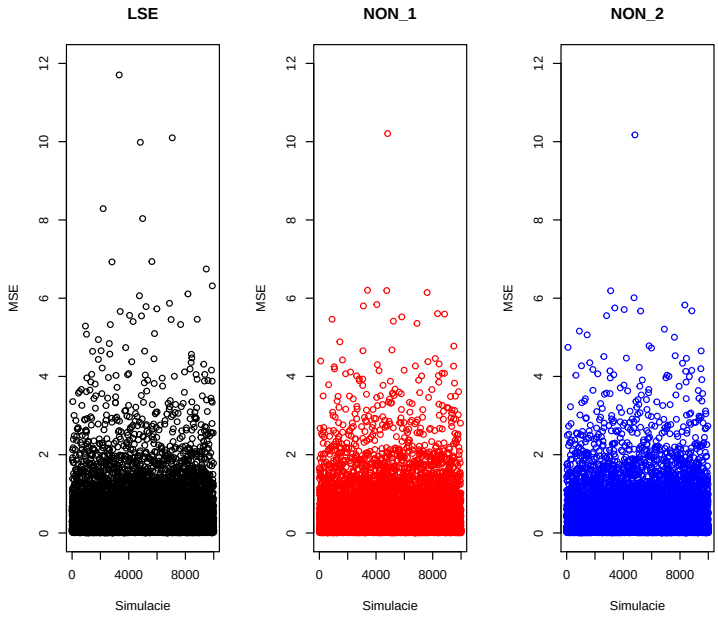
začíname pozorovať dominanciu neparametrických metód, pretože sa začínajú ukazovať lepšie vlastnosti a menšia citlivosť na prítomnosť outlierov v prípadoch rozdelení s ťažšími chvostami. Metódy NON_1 a NON_2 dosahujú takmer vo všetkých prípadoch menšie priemerné MSE v porovnaní s metódou LSE okrem jediného prípadu pri metóde NON_1, pri $n = 20$ a pri *Bent score function*.

Student_5				SCORE FUNCTION		
n	Metóda	Štatistický ukazovateľ		Wilcoxon	Qnorm	Bent
20	LSE	MSE	rel. efe.	1		
			min	0,000884937		
			max	11,70594189		
		víťazstvo 1		**40,49%**	**39,89%**	**45,30%**
	NON_1	MSE	rel. efe.	1,055460918	1,04317254	0,970456451
			min	0,001841706	0,000999709	0,000870601
			max	10,20777525	10,14442623	9,774809284
		víťazstvo 1		**34,67%**	**36,33%**	**28,80%**
		víťazstvo 2		*21,28%*	*39,88%*	*38,84%*
	NON_2	MSE	rel. efe.	1,065907963	1,052494421	1,000860451
			min	0,001030321	0,001256891	0,001248459
			max	10,1732343	10,20062518	10,26983175
		víťazstvo 1		**24,84%**	**23,78%**	**25,90%**
		víťazstvo 2		*20,16%*	*39,78%*	*40,06%*
80	LSE	MSE	rel. efe.	1		
			min	0,000193993		
			max	1,908187101		
		víťazstvo 1		**40,91%**	**39,76%**	**45,69%**
	NON_1	MSE	rel. efe.	1,180957501	1,134266829	1,12184958
			min	0,00035201	8,15581E-05	0,000157735
			max	1,574753846	1,646368479	1,542355933
		víťazstvo 1		**29,62%**	**31,71%**	**25,98%**
		víťazstvo 2		*17,20%*	*40,81%*	*41,99%*
	NON_2	MSE	rel. efe.	1,181845377	1,134195937	1,120203407
			min	0,000308478	0,000138075	0,000177657
			max	1,579623981	1,636426413	1,543490471
		víťazstvo 1		**29,47%**	**28,53%**	**28,33%**
		víťazstvo 2		*17,32%*	*40,85%*	*41,83%*
200	LSE	MSE	rel. efe.	1		
			min	0,000187076		
			max	0,395029667		
		víťazstvo 1		**37,66%**	**36,09%**	**43,12%**
	NON_1	MSE	rel. efe.	1,218175699	1,152893883	1,166362815
			min	9,12357E-05	9,48604E-05	9,48555E-05
			max	0,301889206	0,339893323	0,263443135
		víťazstvo 1		**30,28%**	**30,82%**	**27,36%**
		víťazstvo 2		*18,33%*	*39,11%*	*42,56%*
	NON_2	MSE	rel. efe.	1,218203577	1,153194577	1,165945122
			min	9,11254E-05	0,000102291	0,000108064
			max	0,302275292	0,339106204	0,260397422
		víťazstvo 1		**32,06%**	**33,09%**	**29,52%**
		víťazstvo 2		*18,52%*	*39,17%*	*42,31%*

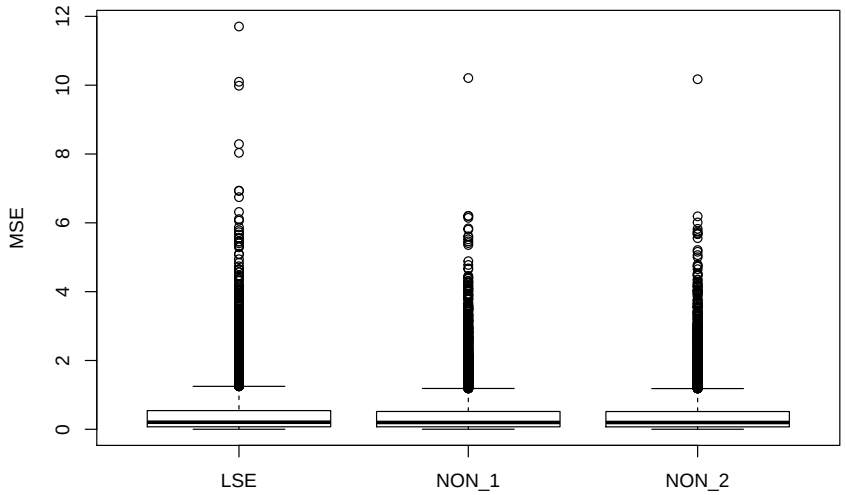
Tab 3.3: Simulácie odhadov regresneých parametrov pri rozdelení t_5

Hodnoty MSE v jednotlivých 10 000 simuláciách v prípade *Wilcoxon score function* a $n = 20$ prehľadne zobrazujeme na nasledujúcich obrázkoch, konkrétne na Obr.3.2 porovnávame metódu najmenších štvorcov, neparametrickú metódu NON_1 a metódu NON_2. Na záver na Obr.3.3 je grafické porovnanie jednotlivých spomínaných metód pomocou *boxplotu*. Čo sa týka výsledkov v premennej *víťazstvo 1* stále je na

prvom mieste parametrická metóda LSE. V rámci neparametrických metód vychádza ako najlepšia na základe priemerného MSE neparametrická metóda s *Wilcoxon score function*. Na druhej strane ak sa pozrieme na výsledky v premennej *víťazstvo 2*, tak vo všetkých prípadoch okrem jedného vyhráva voľba *Bent score function* bez ohľadu na to, či sa pozeráme na metódu NON_1 alebo NON_2.



Obr. 3.2: Porovnanie jednotlivých metód



Obr. 3.3: Porovnanie jednotlivých metód pomocou *boxplotu*

Z hľadiska cieľov našej práce sú pre nás najdôležitejšie výsledky, ktoré sme uviedli v Tab.3.4. V tomto prípade sme generovali náhodné chyby so Studentovho rozdelenia s

1 stupňom voľnosti, ktoré sa nazýva aj Cauchyho rozdelenie. Podľa Obr.3.1 môžeme vidieť, že toto rozdelenie má výrazne ťažšie chvosty ako štandardizované normálne rozdelenie, čo zabezpečuje oveľa častejší výskyt extrémnych hodnôt. Práve v takomto prípade by sa mali ukázať výhody použitia neparametrických metód, ktoré sú robustnejšie. Povedané ľudskou rečou nenechajú sa pomýliť prítomnosťou outlierov v dátovej vzorke, pretože sú oveľa menej citlivé na ich výskyt ako metóda najmenších štvorcov. Tieto očakávania sa jednoznačne napĺňajú z výsledkov uvedených v Tab.3.4.

Student_1				SCORE FUNCTION		
n	Metóda	Štatistický ukazovateľ		Wilcoxon	Qnorm	Bent
20	LSE	MSE	rel. efe.	1		
			min	0,007280791		
			max	896284809		
		víťazstvo 1		**12,29%**	**13,54%**	**0,1326**
	NON_1	MSE	rel. efe.	41746,87744	31251,4247	55409,60174
			min	0,003775398	0,001172286	0,002851098
			max	346,0568725	457,5572305	129,0779519
		víťazstvo 1		**44,61%**	**44,60%**	**42,64%**
		víťazstvo 2		*19,21%*	*24,45%*	*56,34%*
	NON_2	MSE	rel. efe.	41410,22151	31106,42675	55248,52457
			min	0,001974779	0,00317315	0,001405374
			max	461,7730631	432,7358808	131,37316
		víťazstvo 1		**43,10%**	**41,86%**	**44,10%**
		víťazstvo 2		*19,87%*	*23,64%*	*56,49%*
80	LSE	MSE	rel. efe.	1		
			min	0,003256501		
			max	367676547,8		
		víťazstvo 1		**4,17%**	**4,50%**	**3,92%**
	NON_1	MSE	rel. efe.	196287,9389	142984,8987	260774,2921
			min	0,001008843	0,000881025	0,0004487
			max	5,118714908	7,336593748	4,114668321
		víťazstvo 1		**46,08%**	**44,57%**	**47,56%**
		víťazstvo 2		*13,97%*	*19,29%*	*66,74%*
	NON_2	MSE	rel. efe.	197043,8837	143705,4973	261404,7723
			min	0,000910352	0,00093485	0,000421366
			max	5,109785848	7,202967606	4,172686739
		víťazstvo 1		**49,75%**	**50,93%**	**48,52%**
		víťazstvo 2		*14,10%*	*19,44%*	*66,46%*
200	LSE	MSE	rel. efe.	1		
			min	0,002153831		
			max	71838969,7		
		víťazstvo 1		**1,02%**	**1,22%**	**0,89%**
	NON_1	MSE	rel. efe.	194351,7221	141114,2247	257864,6411
			min	0,000197825	0,000179867	0,000170353
			max	0,85323956	1,231636774	0,583640914
		víťazstvo 1		**45,78%**	**43,98%**	**47,60%**
		víťazstvo 2		*12,40%*	*16,06%*	*71,54%*
	NON_2	MSE	rel. efe.	195825,2351	142675,7243	259339,3056
			min	0,000169304	0,000159321	0,000194682
			max	0,857297545	1,198330134	0,581659166
		víťazstvo 1		**53,20%**	**54,80%**	**51,51%**
		víťazstvo 2		*13,29%*	*16,26%*	*70,45%*

Tab 3.4: Simulácie odhadov regresneých parametrov pri rozdelení t_1

Vidíme, že vo všetkých prípadoch rôznej veľkosti dátového súboru alebo použitia rôznych *score function* je metóda LSE nepresnejšia ako metódy NON_1 a NON_2, či už na základe výsledkov priemerného MSE, resp. relatívnej efektivity alebo výsledkov z premennej *víťazstvo 1*. Netvrdíme, že metóda najmenších štvorcov zlyhala v

úplne všetkých 10 000 simuláciách. Na základe údajov z premennej *víťazstvo 1* sme schopný nájsť nejaké prípady, kde dokonca metóda LSE dosiahla najmenšie MSE v danej simulácií v porovnaní s neparametrickými metódami, no je ich oveľa menej ako pri neparametrických metódach. Môžeme si všimnúť, že pre $n = 200$ sú tieto počty v premennej *víťazstvo 1* pre metódu LSE celkom zanedbateľné. Rovnaké závery týkajúce sa porovnania parametrickej metódy LSE a neparametrických metód platia aj vtedy, ak sa pozrieme na výsledky minimálneho resp. maximálneho MSE zo všetkých 10 000 simulácií. Tak ako vo všetkých troch predošlých rozdeleniach platí, že čím máme väčší počet meraní, tým sú metódy vo všeobecnosti presnejšie.

Naozaj sa ukazuje, že v prípade rozdelenia s ťažšími chvostami môže viesť aplikovanie metódy najmenších štvorcov pri odhade neznámych regresných parametrov k nepresnostiam, ktoré môžu ovplyvniť analýzu skúmanej závislosti medzi závislou premennou a vysvetľujúcimi premennými v negatívnom zmysle. Teda má zmysel bližšie sa zaoberať použitím neparametrických metód v lineárnej regresii. Čo sa týka porovnania metódy NON_1 s metódou NON_2, tak nájdeme prípady, v ktorých je lepšia metóda NON_1, ako i prípady, kde je presnejšia metóda NON_2. Napríklad pre $n = 20$ je metóda NON_1 v rámci všetkých troch *score function* lepšia ako NON_2 na základe priemerného MSE a až na jeden prípad aj na základe výsledkov v premennej *víťazstvo 1*. Ak sa však pozrieme na analogické porovnanie v prípadoch $n = 80$ a $n = 200$ uvidíme, že naša metóda NON_2 dosahuje o niečo presnejšie výsledky ako metóda NON_1 z balíka *Rfit* pre všetky druhy *score function*.

Taktiež zaujímavé závery získame, ak sa pozrieme na porovnanie jednotlivých *score function* v neparametrických metódach. Na základe priemerného MSE sa v neparametrických metódach NON_1 a NON_2 pri všetkých druhoch voľby n ukazuje ako najlepšia voľba *Bent score function*, za ňou nasleduje *Wilcoxon score function* a posledná voľba je reprezentovaná *Kvantil score function*. Podľa výsledkov premennej *víťazstvo 2* opäť dominuje *Bent score function* pri metódach NON_1 a NON_2 pri všetkých n, poradie *Wilcoxon* a *Kvantil score function* sa však zmenilo.

Z vyššie opísaných výsledkov vzťahujúcich sa na použitie Cauchyho rozdelenia vyplýva, že pri rozdeleniach s častým výskytom outlierov je vhodnejšie použitie neparametrických metód, ktoré sú odolnejšie na prítomnosť extrémnych hodnôt. Čo sa týka

voľby *score function*, tak na základe výsledkov našich simulácií sa ukazuje ako najlepšie použitie *Bent score function*, pretože odhady neznámych regresných parametrov sú najpresnejšie.

Taktiež si môžeme všimnúť, že bez ohľadu na voľbu n, *score function* alebo použitú metódu platí, že MSE narastalo pri rozdelení s ťažšími chvostami. Častejší výskyt outlierov spôsobil, že parametrické ale aj neparametrické metódy boli menej presné. V niektorých prípadoch bol nárast MSE pomerne zanedbateľný, inokedy bol rozdiel až príliš veľký. Pri neparametrických metódach boli dosiahnuté výsledky v poriadku pri všetkých štyroch rozdeleniach, pretože nárast MSE súvisiaci s prechodom od rozdelenia s ľahšími chvostami k rozdeleniu s ťažšími chvostami bol akceptovateľný. Na druhej strane pri parametrickej metóde LSE bol nárast MSE pri prechode na Cauchyho rozdelenie príliš veľký, čo jednoznačne dokazuje nevhodnosť použitia tejto metódy pri rozdeleniach s častým výskytom extrémnych hodnôt.

4 Simulácie testov o významnosti regresných parametrov

4.1 Opis vykonaných simulácií

V tejto kapitole sme sa zaoberali výsledkami simulácií testov o významnosti regresných parametrov $\beta_1, \beta_2, \ldots, \beta_p$. Vzhľad hypotéz, ktoré boli predmetom nášho záujmu, sme spomenuli v predchádzajúcich kapitolách pomocou vzťahov 1.7 resp. 2.21. V prvom kroku pri počítaní simulácií testov sme parametre $\beta_l,\ l \in Q$, ktoré vystupovali v hypotézach daných vzťahmi 1.7 a 2.21, položili rovné 0. Cieľom bolo určiť v koľkých prípadoch dôjde k tomu, že test zamietne hypotézu H_0 napriek tomu, že hypotéza H_0 platí. Teda sme chceli odhadnúť skutočnú pravdepodobnosť chyby 1. druhu, ktorá by mala byť ideálne rovná hladine významnosti testu. V našom prípade sme testovali na úrovni 5%. Na druhej strane vieme, že kvalitu testu nie je možné posúdiť len na základe pravdepodobnosti chyby 1. druhu. Z toho dôvodu sme postupne testované regresné parametre $\beta_l,\ l \in Q$ v hypotéze H_0 nastavili na hodnotu c, ktorú sme vždy mierne zvyšovali, pokým miera zamietania hypotézy H_0 pri väčšine testov nebola na úrovni 100%. Chceli sme zistiť pravdepodobnosť chyby 2. druhu, teda percentuálny podiel simulácií, v ktorých dôjde k nezamietnutiu hypotézy H_0 napriek tomu, že hypotéza H_0 neplatí. Z toho vieme veľmi jednoducho odvodiť silu testu, ktorá sa vypočíta ako 1 – pravdepodobnosť chyby 2. druhu.

Vykonali sme viacero verzii neparametrických testov opísaných v časti 2.2.2, pretože sme menili *score function*, rôzne odhady parametra τ, taktiež sme menili počet meraní n a rozdelenie náhodných chýb $e_1, e_2, \ldots, e_n$ podobne ako v predošlej kapitole pri simuláciách odhadov. Celkovo sme získali pomerne veľa neparametrických testovacích metód, ktoré sme následne medzi sebou mohli porovnávať. Hlavným cieľom bolo taktiež porovnať neparametrický prístup s klasickým parametrickým prístupom opísaným v časti 1.1.2 a zistiť reakciu jednotlivých metód na zmenu vyššie spomínaných parametrov. Vzájomné porovnanie jednotlivých metód sme vykonali pomocou simulovaných silofunkcií a podrobnejšiemu popisu sa budeme venovať v ďalšej časti. Za prínos tejto časti považujeme to, že porovnanie, ktoré sme vykonali, je opäť dosť komplexné. V iných simulačných prácach autori často využívali len jednu *score function* a taktiež

len jeden odhad parametra τ.

Náhodné chyby sme v simuláciách vyberali podobne ako v 3. kapitole z normálneho a Studentovho rozdelenia s počtom stupňov voľnosti 10, 5 a 1. Počet meraní v modeli n sme uvažovali obdobne ako pri simuláciách odhadov t. j. 20, 80 a nakoniec 200. Čo sa týka voľby *score function*, tak sme pracovali s doterajšími funkciami, ktoré boli *Wilcoxon score function*, *Kvantil score function* a *Bent score function*. Na spočítanie testovacej štatistiky danej vzťahom 2.23 sme museli odhadnúť parameter τ. V 2. kapitole sme spomínali štyri spôsoby na odhad tohto parametra. Vieme, že 1., 2. a 3. spôsob vyžadovali, aby použitá *score function* bola diferencovateľná a 3. spôsob navyše vyžadoval aj lineárnosť funkcie. Tieto vlastnosti spĺňala jedine *Wilcoxon score function*. To je dôvod najčastejšieho používania práve tejto funkcie v neparametrických metódach. To znamená, že pri použití tejto *score function* sme mali na výber pomerne veľkú škálu rôznych odhadov neznámeho parametra τ. Napriek tomu, že pri zvyšných dvoch *score function* sme boli značne limitovaný možnosťami akým spôsobom odhadnúť parametra τ, neodradilo nás to, aby sme ich zakomponovali do našich simulácií a mohli výsledky medzi sebou porovnať.

Pri použití *Kvantil score function* a *Bent score function* sme teda použili na odhad parametra τ len 2. spôsob opísaný v 2. kapitole, ktorý je založený na dĺžke intervalu spoľahlivosti pre intercept ξ, ktorý je počítaný z $e_1^*, e_2^*, \ldots, e_n^*$. Nie je problém nahliadnuť, že čím viac sú chyby $e_1^*, e_2^*, \ldots, e_n^*$ rozptýlené, tým bude dĺžka intervalu spoľahlivosti väčšia. Táto skutočnosť nás primäla k tomu, že sme si povedali, že skúsime odhadnúť parameter τ len na základe kolísavosti chýb $e_1^*, e_2^*, \ldots, e_n^*$, ktorú sme vypočítali pomocou smerodajnej odchýlky. Výhodou tohto spôsobu odhadu parametra τ je jeho jednoduché spočítanie oproti iným spôsobom. Na druhej strane sme nemali príliš veľké očakávania ohľadne získaných výsledkov, no zaujímalo nás preveriť aj takúto jednoduchú možnosť odhadovania parametra τ. Pri použití *Wilcoxon score function* sme okrem predchádzajúcich dvoch možností odhadov aplikovali aj 3. a 4. spôsob založené na odhade funkcie hustoty f rozdelenia náhodných chýb $e_1, e_2, \ldots, e_n$ pomocou kernelovej metódy. Pri oboch spôsoboch však bolo potrebné zvoliť *kernel function* a taktiež určiť *bandwidth* h. Ako sme už spomenuli v 2. kapitole ako *kernel function* sme zvolili hustotu štandardizovaného normálneho rozdelenia a parameter *bandwidth*

h sme spočítali štyrmi spôsobmi h_1, h_2, h_3 *a* h_4 opísanými v rovnakej kapitole.

Taktiež pri použití *Wilcoxon score function* sme na odhad parametra τ využili aj 1. spôsob popísaný v 2. kapitole. Vzhľadom na to, že v práci sme sa usilovali všetky potrebné výpočty implementovať pomocou vlastných funkcií, výnimku sme nechceli spraviť ani v tomto prípade, hoci sme vedeli, že v balíku *Rfit* je zabudovaná funkcia rátajúca tento spôsob odhadu parametra τ. Našu funkciu sme naprogramovali teda podľa článku [9] resp. knihy [4] a chceli sme si výsledok porovnať s výsledkom funkcie *gettau()* z balíka *Rfit*, pretože v jej popise boli uvedené práve vyššie spomenuté zdroje. Na naše prekvapenie sme dostali rozdielne výsledky. V prvom rade sme sa snažili nájsť chybu v našej funkcii, no nedarilo sa nám to. Keď sme však otvorili zdrojový kód funkcie *gettau()*, zistili sme, že je tam istá časť, ktorú sme nevedeli nájsť v zdrojoch [9] a [4], z ktorých sme vychádzali my pri implementácii našej vlastnej funkcie. Z toho dôvodu sme sa rozhodli pri kontaktovaní profesora McKeana opýtať aj na túto skutočnosť. Profesor McKean nám vysvetlil, že pri počítaní tohto odhadu parametra τ spravili dodatočnú korekciu opísanú v článku [11]. Nanešťastie tento článok sme v popise funkcie *gettau()* nenašli, a preto sme nevedeli určiť o akú korekciu išlo resp. zdôvodniť nezhodu nášho výsledku s výsledkom z balíka *Rfit*. Nakoniec sme sa rozhodli v našich simuláciách použiť aj našu funkciu a taktiež funckiu *gettau()* z balíka *Rfit* s cieľom porovnať získané výsledky. Táto skúsenosť nám opäť potvrdila správnosť nášho rozhodnutia pri implementovaní si svojich vlastných funkcií v softwari R, pretože pri využití len samotného balíka *Rfit* by sme na spomínaný rozdiel neprišli.

Ako sme uviedli v 2. kapitole, tak testovacia štatistika daná vzťahom 2.23 sa môže porovnávať s Fisherovým-Snedecorovým alebo Chi-kvadrát rozdelením. Z teórie vieme, že je vhodnejšie použiť práve kritické hodnoty Fisherovho-Snedecorovho rozdelenia. Zdôvodnenie sme uviedli v teoretickej časti. V našich simuláciách sme zrátané testovacie štatistiky porovnávali s oboma rozdeleniami, no v záverečných výsledkoch uvádzame len tie s Fisherovým-Snedecorovým rozdelením. Taktiež aj samotné odhady neznámych regresných parametrov $\beta_1, \beta_2, \ldots, \beta_p$ sme rátali pomocou funkcie z balíka *Rfit*, ale aj pomocou našej vlastnej funkcie. Ako sme mohli vidieť v predošlej kapitole rozdiely v získaných odhadoch pomocou oboch metód neboli príliš výrazne, a preto by bolo zbytočné uvádzať v záverečných výsledkoch oba spôsoby. Z toho dôvodu získané

výsledky publikované v ďalšej časti sú z odhadov parametrov $\beta_1, \beta_2, \ldots, \beta_p$ spočítaných len pomocou funkcie z balíka *Rfit*. V našich simuláciách sme podobne ako v 3. kapitole uvažovali lineárny regresný model s 5 vysvetľujúcimi premennými. V hypotézach sme konkrétne testovali významnosť 1., 2. a 4. vysvetľujúcej premennej. Počet simulácií sme nastavili na 10 000.

4.2 Výsledky simulácií testov o významnosti regresných parametrov

V tejto časti porovnáme medzi sebou pomocou simulovaných silofunkcií viacero metód slúžiacich na testovanie signifikantnosti regresných parametrov. Vzhľadom na to, že sa jedná o grafické porovnanie, uvedieme označenia použitých metód, aby bolo jasné, ktorá silofunkcia reprezentuje akú metódu. V legende každého obrázku sme zobrazili čísla od 1 do 17, pričom každé číslo reprezentuje inú metódu. Číslo 1 je označenie pre klasickú parametrickú metódu opísanú v časti 1.1.2. Čísla 2 až 13 sú metódy, pri ktorých sme použili *Wilcoxon score function*. Číslo 2 je označenie pre metódu, kedy bol parameter τ odhadnutý na základe smerodajnej odchýlky rezíduí $e_1^*, e_2^*, \ldots, e_n^*$. Číslo 3 sme rezervovali metóde, v ktorej sa parameter τ odhadoval pomocou 1. spôsobu a číslo 4 metóde, v ktorej bol odhad parametra τ vykonaný pomocou funkcie z balíka *Rfit* s dodatočnou korekciou. Číslo 5 reprezentuje metódu, v ktorej sa parameter τ odhaduje pomocou 2. spôsobu, teda na základe dĺžky intervalu spoľahlivosti pre intercept počítaný z rezíduí $e_1^*, e_2^*, \ldots, e_n^*$. Čísla 6, 7, 8 a 9 sme rezervovali pre metódy, v ktorých sa parameter τ odhadoval pomocou 3. spôsobu, pričom parameter h (*bandwidth*) bol určený pomocou štyroch metód, ktoré sme v časti 2.2.3 označili h_1, h_2, h_3 *a* h_4. Analogicky je to aj pri číslach 10, 11, 12 a 13 s tým rozdielom, že parameter τ bol odhadnutý pomocou 4. spôsobu. Čísla 14 a 15 reprezentujú metódy, kde sme použili *Kvantil socore function*. Číslo 14 resp. 15 je označenie pre metódu, v ktorej sa parameter τ odhadol na základe smerodajne odchýlky z $e_1^*, e_2^*, \ldots, e_n^*$ resp. na základe dĺžky intervalu spoľahlivosti určeného z rezíduí $e_1^*, e_2^*, \ldots, e_n^*$. Úplne rovnako je to aj pri čísla 16 a 17, ktoré zastupujú metódy, v ktorých sme použili *Bent score function*. V Tab.4.1 zobrazujeme jednotlivé metódy v prehľadnejšej forme.

Číslo metódy	Metóda	Score function	Odhad parametra τ
1	parametrikcá	-	-
2	neparametrická	Wilcoxon	smerodajná odchýlka z rezíduí
3	neparametrická	Wilcoxon	1.spôsob
4	neparametrická	Wilcoxon	1.spôsob pomocou Rfit
5	neparametrická	Wilcoxon	2. spôsob
6	neparametrická	Wilcoxon	3.spôsob a bandwidth ako h1
7	neparametrická	Wilcoxon	3.spôsob a bandwidth ako h2
8	neparametrická	Wilcoxon	3.spôsob a bandwidth ako h3
9	neparametrická	Wilcoxon	3.spôsob a bandwidth ako h4
10	neparametrická	Wilcoxon	4.spôsob a bandwidth ako h1
11	neparametrická	Wilcoxon	4.spôsob a bandwidth ako h2
12	neparametrická	Wilcoxon	4.spôsob a bandwidth ako h3
13	neparametrická	Wilcoxon	4.spôsob a bandwidth ako h4
14	neparametrická	Kvantil	smerodajná odchýlka z rezíduí
15	neparametrická	Kvantil	2. spôsob
16	neparametrická	Bent	smerodajná odchýlka z rezíduí
17	neparametrická	Bent	2. spôsob

Tab 4.1: Prehľad jednotlivých metód

Pre upresnenie na vodorovnej osi v jednotlivých obrázkoch sme zobrazili spomínaný parameter c, ktorý sme postupne zvyšovali a sledovali správanie jednotlivých testovacích metód. Na zvislej osi je miera zamietania hypotézy H_0. Čo sa týka generovania umelých dát, na ktorých sme vykonávali vyššie opísané simulácie, tak sme postupovali analogicky ako v 3. kapitole.

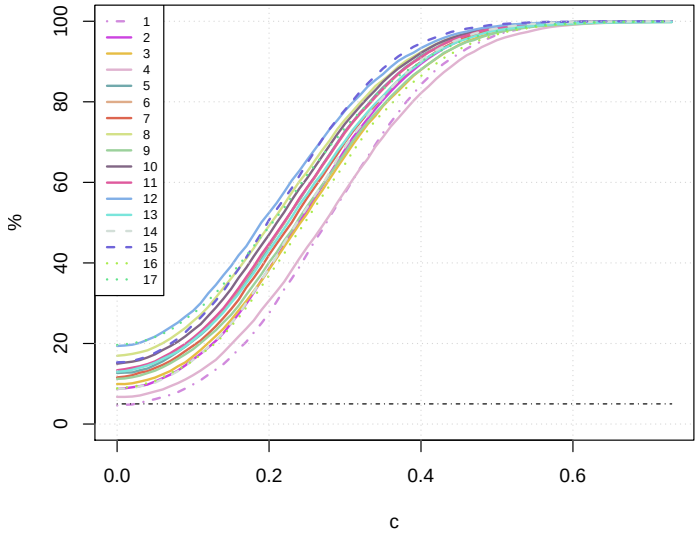
Parameter c sme zvyšovali dovtedy pokiaľ to bolo zmysluplné. To znamená, že ak už väčšina metód dosahovala mieru zamietania hypotézy H_0 na úrovni 100%, tak nemalo zmysel ďalej zvyšovať parameter c. Z toho dôvodu má vodorovná os na jednotlivých obrázkoch zobrazených nižšie rozdielne rozpätie. Celkom prirodzene platilo, že čím bolo n väčšie, tým nám stačila menšia hodnota parametra c na to, aby sme boli schopní vykresliť zmysluplne simulované silofunkcie a metódy medzi sebou porovnávať. Dôvod je jednoduchý. Platí, že čím viac dát uvažujeme v regresnom modeli, tým viac sa aj malý nárast hodnoty testovaného regresného parametra ukáže ako významný. To spôsobí, že väčšina skúmaných metód dosiahne takmer 100% mieru zamietania hypotézy H_0 už pri menšej hodnote parametra c. Tento fakt sme privítali, pretože pre $n = 200$ simulácie trvali síce najdlhšie, ale na dosiahnutie požadovaných výsledkov sme parameter c zvyšovali najmenej. Taktiež ak sme pracovali s rozdelením chýb e s ťažkými chvostami ako bolo napríklad Cauchyho rozdelenie bolo potrebné, aby sa parameter c nastavil na vyššie hodnoty v porovnaní so zvyšnými rozdeleniami, ak

sme chceli dosiahnuť, aby väčšina metód nadobudla mieru zamietaniu hypotézy H_0 na úrovni 100%. Na každom z nižšie uvedených obrázkov sme zobrazili čiernu prerušovanú čiaru reprezentujúcu ideálnu 5%-tnú úroveň podielu zamietania hypotézy H_0.

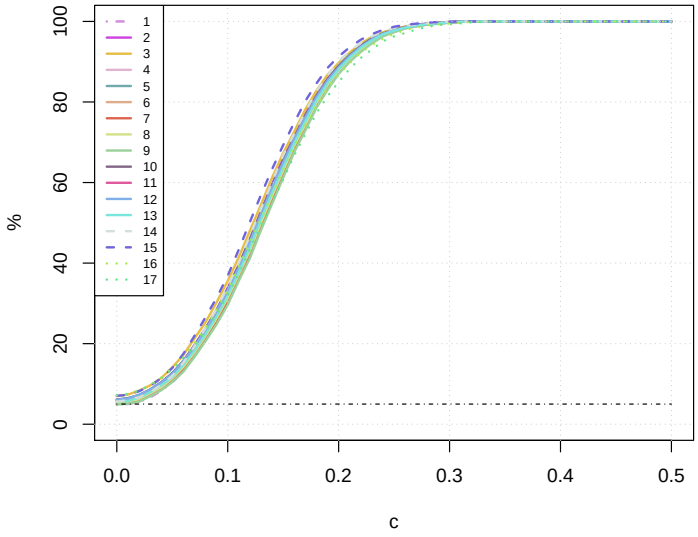
V prvom kroku sme sa pozreli na výsledky simulácií zobrazených na Obr.4.1, v ktorých sme náhodné chyby $e_1, e_2, \ldots, e_n$ generovali zo štandardizovaného normálneho rozdelenia a počet meraní n v modeli bol rovný 20. Vidíme, že len klasická parametrická metóda označená číslom 1 dosahuje pravdepodobnosť chyby 1. druhu približne na úrovni 5%, čo je požadovaná úroveň, keďže testujeme na hladine významnosti 5%. Zvyšné metódy tento ukazovateľ prevyšujú, pričom metódy ako 5,7,8,10,13,15 alebo 17 pomerne výrazne. Jediná neparametrická metóda, ktorá je porovnateľná s klasickou parametrickou metódou, je metóda s číslom 4. Z Obr.4.1 vidíme, že metóda 1 je v prvej časti grafu o trochu viac konzervatívnejšia než 4, no pre väčšie hodnoty parametra c túto iniciatívu preberá metóda s číslom 4 a hypotézu H_0 zamieta menej často ako klasická parametrická metóda 1. Ďalej si z rovnakého obrázka môžeme všimnúť, že takmer všetky neparametrické metódy zamietajú hypotézu H_0 pomerne často už pre malé hodnoty parametra c, čo nepovažujeme za príliš správne riešenie. Dôvod je veľmi jednoduchý. S veľkou silou štatistického testu ide prirodzene ruka v ruke aj vysoká pravdepodobnosť chyby 1. druhu, čo sa prejavuje aj pri väčšine neparametrických metód . Na naše prekvapenie môžeme konštatovať, že metódy 2, 14 a 16, v ktorých sme parameter τ odhadovali len pomocou smerodajnej odchýlky rezíduí $e_1^*, e_2^*, \ldots, e_n^*$ dosahujú porovnateľné dokonca aj lepšie výsledky ako iné neparametrické metódy, v ktorých sa parameter τ odhadoval oveľa zložitejším spôsobom.

Ak sa pozrieme na výsledky simulácií pre $n = 80$ a $n = 200$ na Obr.4.2 a Obr.4.3 môžeme pozorovať, že rozdiely medzi jednotlivými metódami sa zmenšili. Niektoré metódy sú dokonca medzi sebou nerozoznateľné. Vieme, že v oblasti, kde platí hypotéza H_0 t. j. pre $c = 0$ dôjde k zníženiu falošného zamietania hypotézy H_0 v porovnaní s prípadom $n = 20$. To znamená, že pri väčšine metód klesne pravdepodobnosť chyby 1. druhu k požadovanej 5%-nej úrovni, čo dokazujú Obr.4.2 resp. Obr.4.3. Na druhej strane v oblasti, v ktorej hypotéza H_0 neplatí, dôjde pri jednotlivých metódach k zvýšeniu ochoty jej zamietania v porovnaní s prípadom $n = 20$, čo korešponduje s teoretickými poznatkami. Napriek malým rozdielom medzi metódami na Obr.4.2 a

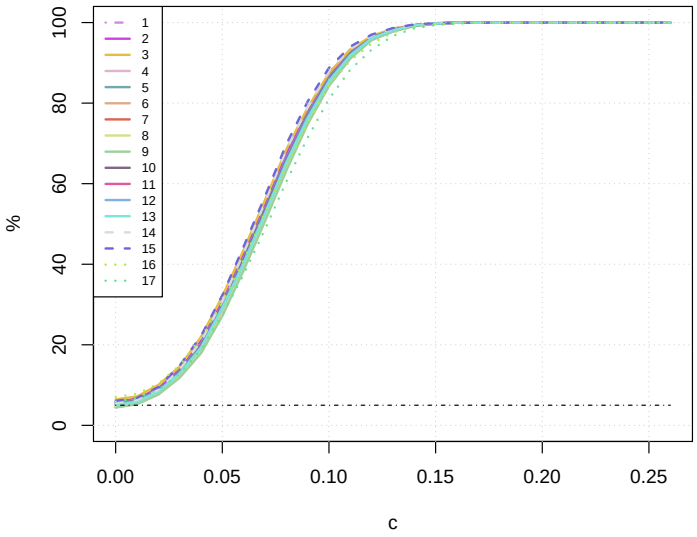
Obr.4.3 môžeme postrehnúť, že metóda 15 je v oboch prípadoch najliberálnejšia, pretože od istej hodnoty parametra c zamieta hypotézu H_0 najčastejšie. Opačnú situáciu môžeme pozorovať na Obr.4.2 a Obr.4.3 pri metóde 17, v ktorej opäť od určitého parametra c, t. j. v oblasti s neplatnou hypotézou H_0, zamietanie tejto hypotézy je o niečo menej časté v porovnaní s konkurenčnými metódami. Vzhľadom na to, že uvedené výsledky sa týkali štandardizovaného normálneho rozdelenia, tak prioritne by sme mali využívať klasický parametrický test, ktorý spoľahlivo funguje, hoci niektoré neparametrické metódy boli takmer identické v prípade $n = 80$ a $n = 200$.



Obr. 4.1: Simulácie testov pri N(0,1) a $n = 20$

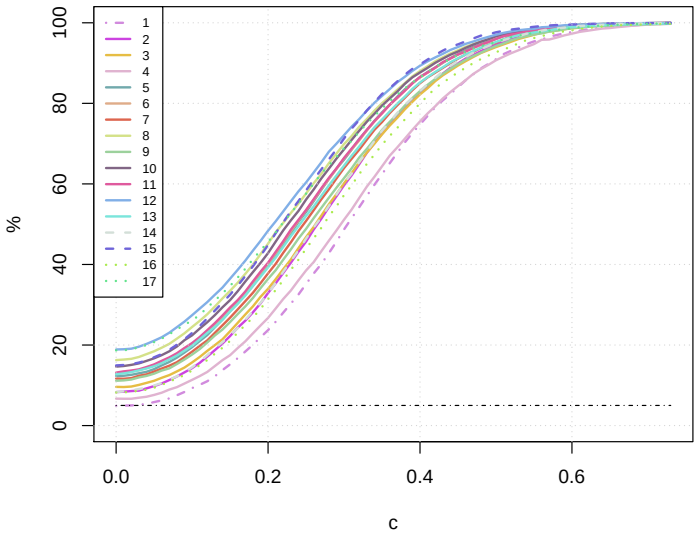


Obr. 4.2: Simulácie testov pri N(0,1) a $n = 80$

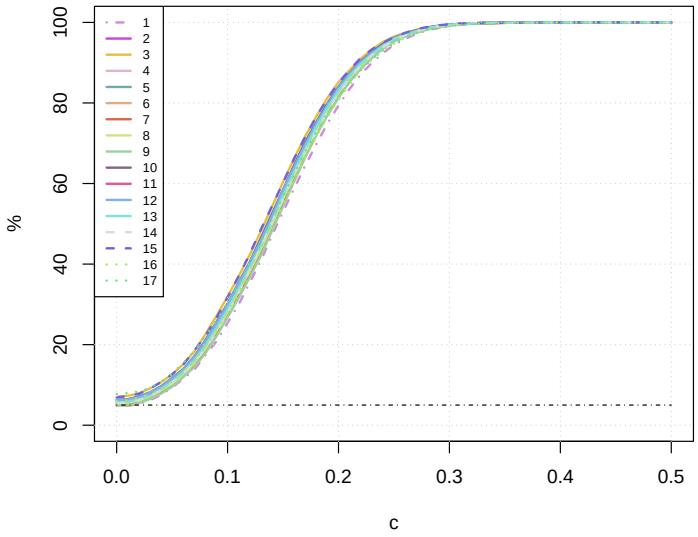


Obr. 4.3: Simulácie testov pri N(0,1) a $n = 200$

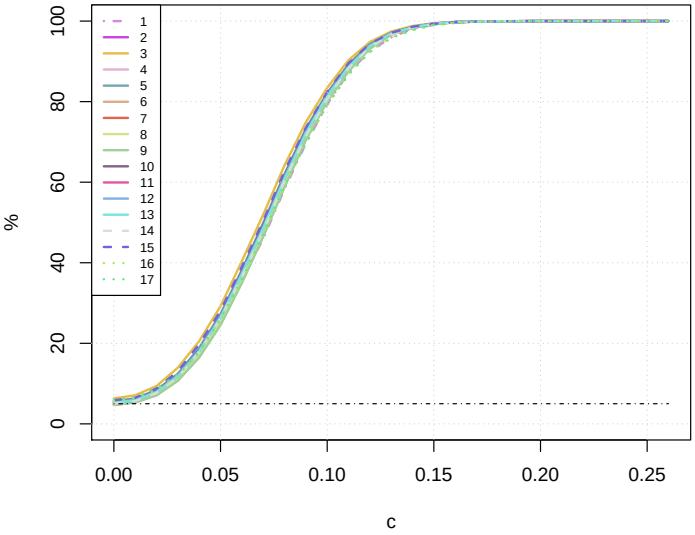
Na ďalších nižšie zobrazených obrázkoch uvádzame výsledky simulácií, v ktorých sme náhodné chyby $e_1, e_2, \ldots, e_n$ generovali zo Studentovho rozdelenia s 10 stupňami voľnosti. Z Obr.3.1 sme mohli vidieť, že rozdiel medzi Studentovym rozdelením s 10 stupňami voľnosti a štandardizovaným normálnym rozdelením nebol vôbec veľký. Studentovo rozdelenie s 10 stupňami voľnosti má len o niečo ťažšie chvosty, teda výskyt extrémnych hodnôt je len o niečo častejší v porovnaní so štandardizovaným normálnym rozdelením. To je dôvod toho, že nižšie zobrazené výsledky simulácií sú veľmi podobné ako pri predošlom rozdelení, ktoré sme opísali vyššie.



Obr. 4.4: Simulácie testov pri t_{10} a $n = 20$



Obr. 4.5: Simulácie testov pri t_{10} a $n = 80$

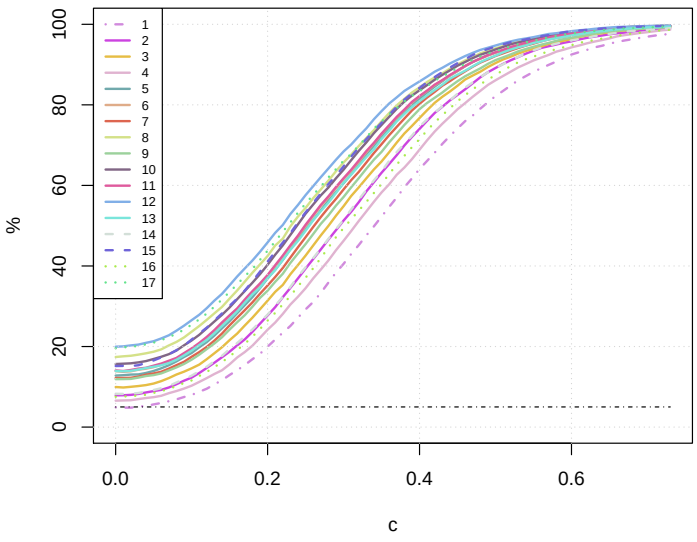


Obr. 4.6: Simulácie testov pri t_{10} a $n = 200$

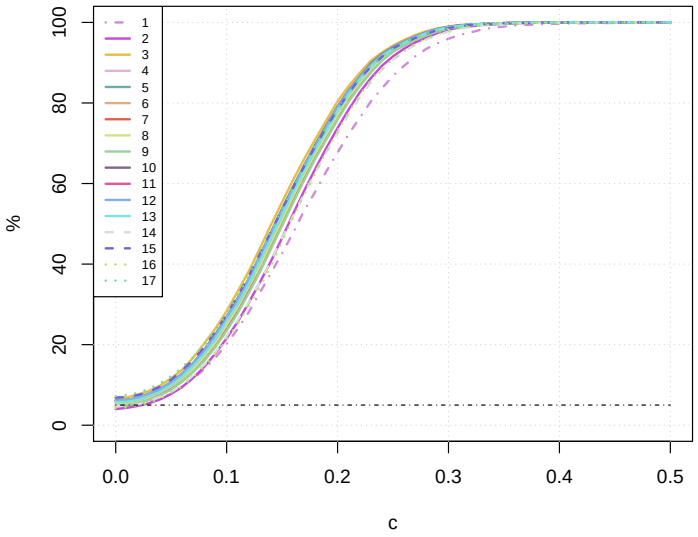
Z Obr.4.4 vidíme, že opätovne všetky neparametrické metódy zamietajú hypotézu H_0 častejšie ako klasická parametrická metóda, ktorá ako jediná má pravdepodobnosť chyby 1. druhu na úrovni 5%. Jediná neparametrická metóda porovnateľná z hľadiska silofunkcie s metódou číslo 1, je metóda 4. Medzi neparametrickými metódami dosahuje najnižšiu pravdepodobnosť chyby 1. druhu na rozdiel od iných neparametrických metód ako napríklad 8, 12, 15 alebo 17, ktoré majú pravdepodobnosť chyby 1. druhu oveľa vyššiu a neodporúčali by sme ich používať za daných podmienok. Ak sa pozrieme na Obr.4.5 resp. Obr.4.6, ktorý reprezentuje simulácie pre $n = 80$ resp. $n = 200$, platia analogické myšlienky týkajúce sa prechodu z $n = 20$ na $n = 80$ prípadne $n = 200$, ktoré

sme spomínali pri predošlom rozdelení a nebudeme ich opakovať. V tomto prípade však za najkonzervatívnejšie metódy môžeme považovať metódy 3 a 15 v prípade $n = 80$. V situácii, kde $n = 200$ je to opäť metóda 3. Môžeme si však všimnúť z Obr.4.5, že klasická parametrická metóda je stále pod všetkými neparametrickými metódami, čiže hypotézu H_0 bude zamietať menej ochotne. Pri normálnom rozdelení táto metóda s číslom 1 splynula s ostatnými neparametrickými metódami. Napriek tomu, že náhodne chyby nie sú z normálneho rozdelenia, si klasická parametrická metóda počína veľmi dobre aj v prípade rozdelenie s mierne ťažšími chvostami.

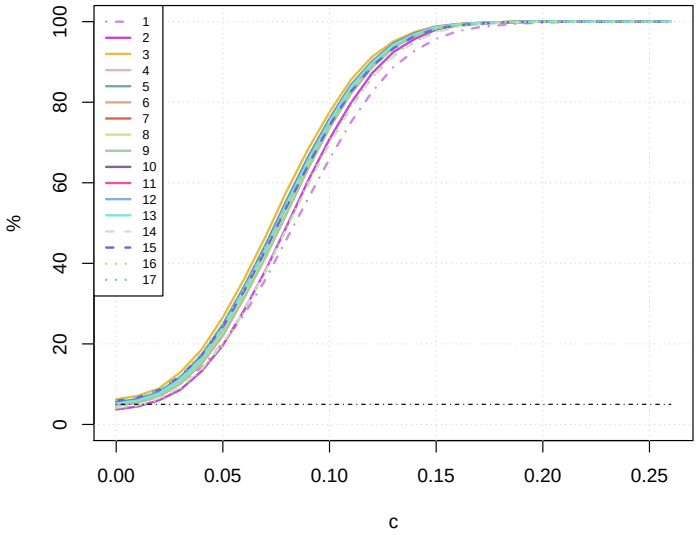
Na ďalších troch obrázkoch Obr.4.7, Obr.4.8 a Obr.4.9 dokumentujeme simulované silofunkcie v situácií, kedy náhodné chyby pochádzajú zo Studentovho rozdelenia s 5 stupňami voľnosti. Na rozdiel od predošlých dvoch prípadov môžeme vidieť, že rozdiely medzi jednotlivými metódami sú väčšie. Na Obr.4.7 reprezentujúci prípad $n = 20$ pozorujeme, že metóda 1 je opätovne najmenej ochotná zamietať hypotézu H_0. Čo sa týka pravdepodobnosti chyby 1. druhu, je metóda 1 opätovne jediná, ktorá ju dosahuje približne na úrovni 5%. Tým, že metódy sú medzi sebou viac rozlíšiteľné, z Obr.4.7 môžeme pomerne dobre vidieť usporiadanie silofunkcií jednotlivých metód. V prípade Studentovho rozdelenia s 5 stupňami voľnosti považujeme za najrozumnejšie použiť metódu s číslom 4, pretože medzi neparametrickými metódami dosahuje najmenšiu pravdepodobnosť chyby 1. druhu a v oblasti, kde hypotéza H_0 neplatí sa správa odvážnejšie a istejšie ako metóda 1.



Obr. 4.7: Simulácie testov pri t_5 a $n = 20$



Obr. 4.8: Simulácie testov pri t_5 a $n = 80$



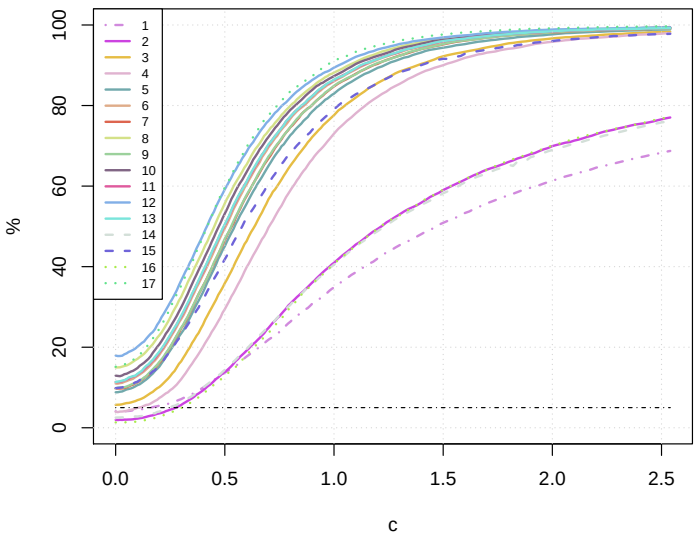
Obr. 4.9: Simulácie testov pri t_5 a $n = 200$

Na naše prekvapenie dobre si počínajú aj metódy 2, 14 a 16, v ktorých sme kľúčový parameter τ odhadli len na základe smerodajnej odchýlky rezíduí , čo je neporovnateľne jednoduchší spôsob v porovnaní s inými spôsobmi odhadu parametra τ.

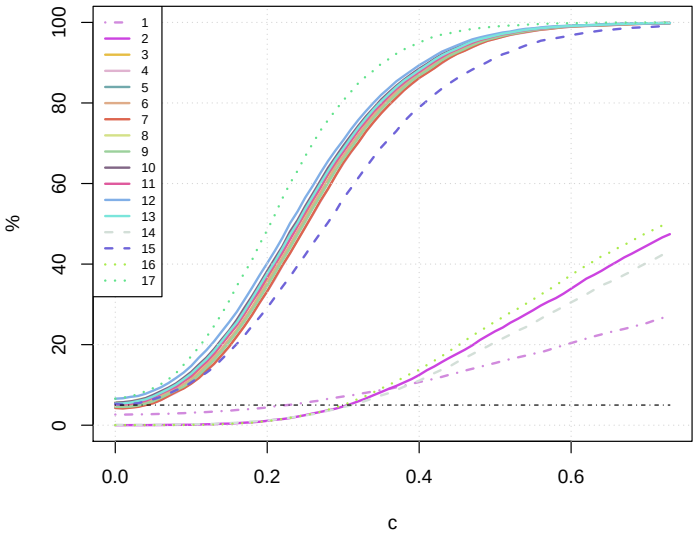
Spomínané metódy majú mierne vyššiu mieru zamietania hypotézy H_0 v porovnaní s metódou 4, teda sú menej konzervatívne alebo inak povedané viac liberálnejšie. V prípadoch $n = 80$ a $n = 200$ zobrazených na Obr.4.8 a Obr.4.9 vidíme, že miera falošného zamietania hypotézy H_0 sa znížila. Taktiež podľa očakávaní rozdiely medzi jednotlivými metódami sú oveľa menšie. V oboch prípadoch sa zdá, že parametrická metóda 1 je menej ochotná zamietať hypotézu H_0. To znamená, že sila tejto metódy

je nižšia v porovnaní s neparametrickými metódami. Vo všeobecnosti pre všetky metódy sme z našich výsledkov pozorovali menšiu mieru zamietania hypotézy H_0, keď sme postupne prechádzali od rozdelení s ľahšími chvostami k rozdeleniam s ťažšími chvostami. Povedané ľudskou rečou, častejší výskyt outlierov spôsobil, že metódy sú si menej isté zamietnutím hypotézy H_0, boja sa príliš častého zamietnutia hypotézy H_0.

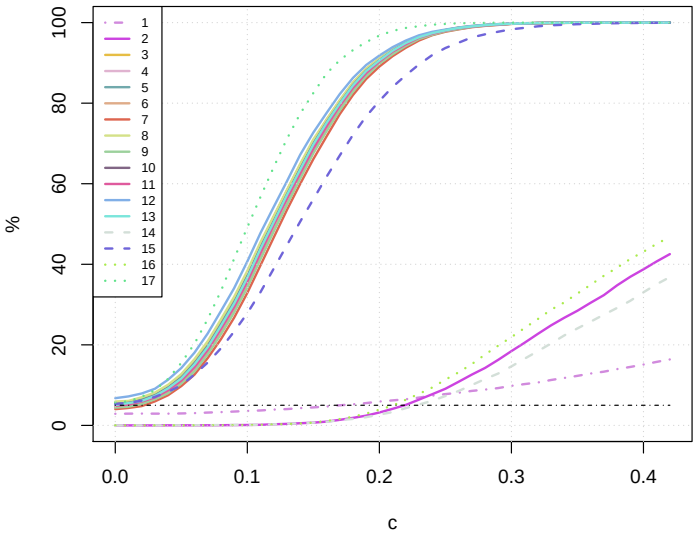
Na nižšie zobrazených troch obrázkoch Obr.4.10, Obr.4.11 a Obr.4.12 uvádzame najdôležitejšie výsledky v rámci celej kapitoly, pretože sa týkajú simulácií, v ktorých sme náhodné chyby generovali so Studentovho rozdelenia s 1 stupňom voľnosti, čo je Cauchyho rozdelenie. Z Obr.3.1 jasne vidíme, že Cauchyho rozdelenie má výrazne ťažšie chvosty v porovnaní s predošlými rozdeleniami. Výskyt extrémnych hodnôt je v tomto prípade oveľa frekventovanejší, čo sa prejavilo vo výsledkoch niektorých metód. Na rozdiel od predošlých prípadov, v ktorých rozdiel medzi parametrickým a neparametrickým prístupom nebol vôbec markantný, v prípade použitia Cauchyho rozdelenia sa celkom jasne ukazuje zo všetkých troch obrázkov veľká výhoda použitia neparametrických metód. Nech uvažujeme akékoľvek n, z obrázkov Obr.4.10, Obr.4.11 a Obr.4.12 môžeme pozorovať, že klasická parametrická metóda s číslom 1 zlyhala na celej čiare. Tento výsledok nás vôbec neprekvapuje, keďže vieme, že parametrické metódy sú citlivé na výskyt outlierov. Podobne zlé výsledky však zaznamenávajú aj niektoré neparametrické metódy s číslami 2, 14 a 16, pretože ich ochota zamietať hypotézu H_0 je na veľmi nízkej úrovni.



Obr. 4.10: Simulácie testov pri t_1 a $n = 20$



Obr. 4.11: Simulácie testov pri t_1 a $n = 80$



Obr. 4.12: Simulácie testov pri t_1 a $n = 200$

Tieto výsledky nie sú taktiež prekvapujúce, vzhľadom na to, že v metódach 2, 14 a 16 sme neznámy parameter τ odhadovali príliš jednoduchým spôsobom na základe smerodajnej odchýlky z rezíduí $e_1^*, e_2^*, \ldots, e_n^*$, ktorý sme chceli vyskúšať len zo zvedavosti. To znamená, že sila spomínaných troch neparametrických testov 2, 14 a 16 ako aj sila klasického parametrického testu je v porovnaní s ostatnými metódami oveľa nižšia, a preto tieto metódy nie je vhodné používať pri daných podmienkach. Z toho dôvodu sa im už v zvyšnej časti opisu výsledkov nebudeme venovať. Úvahy, ktoré sme spomínali v predošlých častiach týkajúce sa hodnoty parametra n a jeho vplyvu na výsledky testovacích metód platia analogicky aj pri Cauchyho rozdelení.

V prvom kroku sa pozrieme na simulované silofunkcie z Obr.4.10, ktorý reprezentuje situáciu s lineárnym regresným modelom vybudovaným z 20 dát. Za najvhodnejšie metódy v tomto prípade považujeme metódy s číslom 3 a 4. Tieto metódy sú metodikou výpočtu takmer analogické len s minimálnym rozdielom, ktorý sme opísali v úvode tejto kapitoly. Obe metódy majú pravdepodobnosť chyby 1. druhu približne na úrovni 5%. Metóda 4 zamieta hypotézu H_0 o trochu častejšie ako metóda 3 v oblasti, kde hypotéza H_0 neplatí, no v porovnaní so zvyšnými neparametrickými metódami je rozdiel malý. Zvyšné neparametrické metódy nám v rovnakej oblasti zamietajú hypotézu H_0 až príliš často, keďže uvažujeme model len s 20 meraniami, a preto ich považujeme za príliš liberálne. Navyše aj pravdepodobnosť chyby 1. druhu majú výrazne vyššiu v porovnaní s testami 3 alebo 4. Na Obr.4.11 a Obr.4.12 dokumentujeme výsledky simulácií prináležiace k $n = 80$ a $n = 200$. Vidíme, že ako vo všetkých doterajších prípadoch sa rozdiely medzi metódami zmenšili. V oboch obrázkoch si môžeme všimnúť dve metódy s číslami 15 a 17, ktoré sa odlišujú od ostatných neparametrických metód. Metóda 15 je príliš konzervatívna v porovnaní so zvyšnými neparametrickými metódami, pretože je málo ochotná zamietať hypotézu H_0, a preto by sme ju neodporučili používať. Na druhej strane metóda 17 je oveľa liberálnejšia v oblasti, kde hypotéza H_0 neplatí a navyše pravdepodobnosť chyby 1. druhu má približne rovnú 5%. Z týchto dôvodov môžeme konštatovať, že metóda 17 dosahuje veľmi dobré výsledky v porovnaní s neparametrickou konkurenciou. Zvyšné neparametrické metódy sú medzi sebou veľmi porovnateľné, niektoré sú viac ochotné zamietať hypotézu H_0, iné menej, no rozdiely nie sú vôbec veľké. Taktiež, čo sa týka pravdepodobnosti chyby 1. druhu dosahujú tieto testy pomerne dobré výsledky, keďže pri väčšine z nich je v blízkom okolí 5%. Z toho dôvodu by sme ich odporučili na praktické aplikovanie v prípade testovania významnosti regresných parametrov v situácií, keď náhodné chyby pochádzajú zo symetrického rozdelenia s ťažkými chvostami ako je napríklad Cauchyho rozdelenie. Na záver by sme ešte na upresnenie spomenuli, že na základe výsledkov našich simulácií metódy s číslami 1, 2, 14, 15 a 16 by sme v tejto situácií neodporúčali využívať.

Hlavný odkaz z tejto kapitoly je v tom, že sa potvrdili naše očakávania, pretože neparametrické metódy určené na testovanie signifikancie regresných parametrov sú robustné a vedia sa vysporiadať so situáciou, ak náhodné chyby nepochádzajú z nor-

málneho rozdelenia. Práve naopak, tieto metódy si vedia poradiť, ak rozdelenie náhodných chýb obsahuje veľký počet outlierov, čo klasický parametrický test nedokáže. To bol v podstate aj jeden z cieľov našej diplomovej práce.

5 Simulácie intervalov spoľahlivosti

5.1 Opis vykonaných simulácií

V tejto záverečnej kapitole sme sa zaoberali porovnaním 95%-ných intervalov spoľahlivosti skonštruovaných pomocou parametrického alebo neparametrického prístupu. Porovnanie sme vykonali rovnakým spôsobom ako v predošlých dvoch kapitolách na základe simulácií. Parametrický prístup na výpočet intervalu spoľahlivosti pre regresný parameter $\beta_l,\ l = 1, 2, \ldots, p$ sme reprezentovali intervalom spoľahlivosti daným vzťahom 1.12. Neparametrické intervaly spoľahlivosti sme konštruovali až dva. Prvý z nich bol daný vzťahom 2.42 a druhý pomocou vzťahu 2.44. Tak ako sme už spomínali v teoretickej časti oba neparametrické intervaly spoľahlivosti sú asymptotické. To znamená, že požadovanú spoľahlivosť na úrovni 95% dosahujú, ak $n \to \infty$. Tak ako v doterajších simuláciách aj pri simuláciách intervalov spoľahlivosti sme menili rôzne parametre ako napríklad n, *score function* alebo rozdelenie náhodných chýb. Tieto zmeny sme už opisovali v predošlých kapitolách, a preto ich v tejto časti podrobnejšie rozoberať nebudeme. Cieľom bolo opäť vykonať komplexné simulačné porovnanie, ktoré sa v iných prácach nevyskytovalo, čo taktiež považujeme za prínos našej diplomovej práce. Podobne čo sa týka spôsobu generovania umelých dát, tak sme ich vytvorili analogicky ako v iných simuláciách až na jeden rozdiel. Interval spoľahlivosti navrhnutý Marekom Omelkom vyžadoval, aby stĺpce matice X s jednotlivými vysvetľujúcimi premennými boli vycentrované, teda aby ich priemer bol rovný 0. V dôsledku toho sme túto maticu generovali tak, aby spomínaná požiadavka bola splnená. V simuláciách sme uvažovali lineárny model s 5 vysvetľujúcimi premennými. Bez ujmy na všeobecnosti sme sa rozhodli, že intervaly spoľahlivosti budeme počítať pre regresný parameter β_2. Počet simulácií sme taktiež zachovali rovnajúci sa 10 000.

5.2 Výsledky simulácií intervalov spoľahlivosti

V nasledujúcich štyroch tabuľkách uvedieme postupne výsledky vyššie opísaných simulácií intervalov spoľahlivosti. Intervaly spoľahlivosti sme hodnotili na základe ich spoľahlivosti a priemernej dĺžky. Je jasné, že interval spoľahlivosti je dobrý vtedy, ak dosahuje vysokú spoľahlivosť pri čo najmenšej dĺžke. V spomínaných štyroch tabuľ-

kách v riadku Dĺžka1 uvádzame explicitne priemernú dĺžku jednotlivých intervalov spoľahlivosti a v riadku Dĺžka2 zobrazujeme koľkokrát sú užšie ako klasický parametrický interval, ktorý budeme v ďalšom texte označovať ako LSE interval spoľahlivosti. Pri neparametrickom intervale danom vzťahom 2.42 sme museli odhadovať neznámy parameter τ. Čísla 2 až 17 v záhlaví každej tabuľky reprezentujú jednotlivé neparametrické metódy líšiace sa spôsobom odhadu spomínaného parametra τ. Očíslovanie týchto metód korešponduje s opisom zo 4. kapitoly a s Tab.4.1. Neparametrický interval spoľahlivosti daný vzťahom 2.44 sme označili písmenkami OM. V tejto kapitole sme opäť rozlišovali ako sme získali odhady neznámych regresných parametrov $\beta_l,\ l = 1, 2, \ldots, p$. Značenie sme ponechali analogické ako v 3. kapitole t. j. NON_1 reprezentuje odhady regresných parametrov získané pomocou balíka *Rfit* a NON_2 reprezentuje našu implementáciu neparametrickej metódy. Označenie LSE je v tomto prípade pre interval spoľahlivosti získaný parametrickým prístupom a daný vzťahom 1.12.

V Tab.5.1 zobrazujeme výsledky simulácií, v ktorých náhodné chyby pochádzali zo štandardizovaného normálneho rozdelenia. Vôbec nie je prekvapujúce, že v tomto prípade dosahuje najlepšie výsledky práve LSE interval spoľahlivosti, pretože sú splnené predpoklady na to, aby nadobúdal požadované vlastnosti. Ak sa pozrieme na neparametrické metódy, pri $n = 80$ alebo $n = 200$ sú ich výsledky pri väčšine metód porovnateľné s parametrickým prístupom. Do pozornosti však dávame, že neparametrické intervaly spoľahlivosti skonštruované pomocou *Bent score function* dosahujú v tejto situácii oproti väčšine neparametrických metód o niečo horšie výsledky, pretože zaznamenali pomerne nízku spoľahlivosť. Ak by sme chceli napriek tomu použiť z nejakých dôvodov neparametrický interval spoľahlivosti, odporúčali by sme použiť tie s číslami 4, 6, 7, 8, 9, 10 alebo 13. Taktiež ak uvažujeme $n = 80$ alebo $n = 200$, ako vhodný kandidát v rámci neparametrických intervalov spoľahlivosti sa ukazuje interval s označením OM, pri ktorom sa vyhneme odhadu neznámeho parametra τ, čo je nepochybne jeho veľká výhoda, pričom jeho výsledky sú dokonca lepšie ako pri iných neparametrických intervaloch a navyše veľmi podobné LSE intervalu spoľahlivosti

V prípade malého počtu dát v modeli, ktorý je reprezentovaný cez parameter $n = 20$, sa ešte neprejavujú asymptotické vlastnosti neparametrických intervalov spo-

N(0,1)			SCORE FUNCTION															
			Wilcoxon												Qnorm		Bent	
			Čísla metód pre odhady parametra τ (metódy NON_1 a NON_2)															
n	Metóda		2	3	4	5	6	7	8	9	10	11	12	13	OM	14	15	16	17
20	LSE	Spoľ./Dĺžka1	95,2/0,91																
	NON_1	Spoľahlivosť	90,8	90,1	93,9	86,6	86,4	87,8	82,2	88,6	84,3	86,2	79,4	86,7	98,5	90,9	83,3	89,8	79,6
		Dĺžka1	*0,79*	*0,81*	*0,94*	*0,72*	*0,72*	*0,75*	*0,66*	*0,82*	*0,68*	*0,71*	*0,62*	*0,76*	*1,25*	*0,79*	*0,64*	*0,82*	*0,67*
		Dĺžka2	*1,15*	*1,13*	*0,97*	*1,26*	*1,27*	*1,21*	*1,37*	*1,11*	*1,33*	*1,27*	*1,46*	*1,19*	*0,73*	*1,15*	*1,42*	*1,11*	*1,36*
	NON_2	Spoľahlivosť	90,9	90,7	94,3	87,3	87,0	88,7	84,3	89,4	85,2	86,9	81,7	87,4	98,6	91,0	83,8	90,2	81,2
		Dĺžka1	*0,79*	*0,81*	*0,95*	*0,72*	*0,72*	*0,76*	*0,69*	*0,86*	*0,69*	*0,72*	*0,65*	*0,80*	*1,25*	*0,78*	*0,64*	*0,81*	*0,68*
		Dĺžka2	*1,15*	*1,11*	*0,95*	*1,26*	*1,25*	*1,20*	*1,32*	*1,06*	*1,31*	*1,26*	*1,40*	*1,14*	*0,73*	*1,16*	*1,42*	*1,12*	*1,33*
80	LSE	Spoľ./Dĺžka1	95,45/0,44																
	NON_1	Spoľahlivosť	93,7	92,9	95,2	93,8	94,8	95,3	95,0	95,6	93,9	94,4	94,0	94,8	95,6	94,3	92,7	91,9	92,4
		Dĺžka1	*0,43*	*0,41*	*0,45*	*0,43*	*0,44*	*0,46*	*0,45*	*0,46*	*0,43*	*0,44*	*0,43*	*0,44*	*0,46*	*0,43*	*0,40*	*0,43*	*0,44*
		Dĺžka2	*1,03*	*1,06*	*0,97*	*1,03*	*0,99*	*0,97*	*0,98*	*0,95*	*1,03*	*1,01*	*1,02*	*1,00*	*0,96*	*1,03*	*1,09*	*1,03*	*0,99*
	NON_2	Spoľahlivosť	93,7	92,9	95,1	93,8	94,8	95,3	95,0	95,6	94,0	94,5	94,1	94,7	95,6	94,3	92,7	91,8	92,4
		Dĺžka1	*0,43*	*0,41*	*0,45*	*0,43*	*0,44*	*0,46*	*0,45*	*0,46*	*0,43*	*0,44*	*0,43*	*0,44*	*0,46*	*0,43*	*0,40*	*0,43*	*0,44*
		Dĺžka2	*1,03*	*1,06*	*0,97*	*1,03*	*0,99*	*0,97*	*0,98*	*0,95*	*1,03*	*1,01*	*1,02*	*1,00*	*0,96*	*1,03*	*1,09*	*1,03*	*0,99*
200	LSE	Spoľ./Dĺžka1	94,85/0,25																
	NON_1	Spoľahlivosť	93,8	93,2	94,7	94,1	94,9	95,3	95,2	95,4	94,3	94,6	94,5	94,7	95,0	94,4	93,8	92,3	93,6
		Dĺžka1	*0,25*	*0,25*	*0,26*	*0,26*	*0,26*	*0,27*	*0,27*	*0,27*	*0,26*	*0,26*	*0,26*	*0,26*	*0,26*	*0,25*	*0,24*	*0,25*	*0,27*
		Dĺžka2	*1,01*	*1,03*	*0,97*	*1,00*	*0,97*	*0,95*	*0,95*	*0,94*	*0,99*	*0,98*	*0,98*	*0,97*	*0,97*	*1,01*	*1,04*	*1,01*	*0,94*
	NON_2	Spoľahlivosť	93,8	93,2	94,8	94,1	94,9	95,3	95,2	95,5	94,3	94,6	94,5	94,7	95,0	94,4	93,7	92,3	93,5
		Dĺžka1	*0,25*	*0,25*	*0,26*	*0,26*	*0,26*	*0,27*	*0,27*	*0,27*	*0,26*	*0,26*	*0,26*	*0,26*	*0,26*	*0,25*	*0,24*	*0,25*	*0,27*
		Dĺžka2	*1,01*	*1,03*	*0,97*	*1,00*	*0,97*	*0,95*	*0,95*	*0,94*	*0,99*	*0,98*	*0,98*	*0,97*	*0,97*	*1,01*	*1,04*	*1,01*	*0,94*

Tab 5.1: Simulácie intervalov spoľahlivosti pri N(0,1)

ľahlivosti. Môžeme si všimnúť, že väčšina týchto intervalov spoľahlivosti nedosahuje požadovanú spoľahlivosť na úrovni 95%. Ich priemerné dĺžky sú síce o trochu kratšie ako LSE interval spoľahlivosti, no neodporúčali by sme ich používať v prípade, ak sú náhodné chyby z normálneho rozdelenia. Najlepší neparametrický interval porovnateľný s LSE intervalom spoľahlivosti je interval s číslom 4. No stále však zostáva platné, že LSE interval spoľahlivosti je lepší, pretože má vyššiu spoľahlivosť a kratšiu priemernú dĺžku. Z Tab.5.1 môžeme spozorovať, že interval spoľahlivosti s označením OM dosahuje spoľahlivosť dokonca vyššiu ako 95%, no v porovnaní s konkurenciou je až príliš široký a teda menej informatívny.

V Tab.5.2 dokumentujeme výsledky simulácií s náhodnými chybami generovanými zo Studentovho rozdelenia s 10 stupňami voľnosti. Ako už vieme na základe Obr.3.1, rozdiel medzi týmto rozdelením a rozdelením N(0,1) nie je veľmi výrazný, čo sa prejavilo aj v získaných výsledkoch. Z toho dôvodu sú závery týkajúce sa výsledkov z Tab.5.2 analogické ako sme uviedli pri predošlom rozdelení a nebudeme ich tu opätovne uvádzať, aby sme sa neopakovali. Môžeme si však všimnúť jeden zaujímavý fakt. Priemerné dĺžky intervalov spoľahlivosti sú v porovnaní so štandardizovaným normálnym rozdelením mierne väčšie pri ľubovoľnej metóde uvedenej v tabuľke. Podobne to bude platiť aj pri ďalších dvoch rozdeleniach. Zdôvodnenie je prosté. Povedané ľudskou rečou, ak

náhodné chyby pochádzajú z rozdelenia s ťažšími chvostami, konštrukcia intervalov spoľahlivosti je zložitejšia, metódy sa musia vysporiadať s náročnejšími podmienkami v porovnaní s rozdelením, ktorého chvosty sú ľahšie. Jednotlivé metódy nie sú schopné pri takto zhoršených okolnostiach vytvoriť intervaly spoľahlivosti s dĺžkou porovnateľnou so situáciou, v ktorej sú náhodne chyby z rozdelenia s ľahšími chvostami. Prejaví sa to tým, že na dosiahnutie požadovanej kvality sa intervaly spoľahlivosti stanú širšie.

Student_10			SCORE FUNCTION																
			Wilcoxon												Qnorm		Bent	
			Čísla metód pre odhady parametra τ (metódy NON_1 a NON_2)															
n	Metóda		2	3	4	5	6	7	8	9	10	11	12	13	OM	14	15	16	17
20	LSE	Spoľ./Dĺžka1	94,61/1																
	NON_1	Spoľahlivosť	91,0	90,4	93,8	86,5	86,2	88,1	82,9	88,7	84,1	85,9	80,0	86,6	98,4	91,1	83,2	90,4	79,3
		Dĺžka1	*0,88*	*0,87*	*1,02*	*0,78*	*0,77*	*0,81*	*0,71*	*0,86*	*0,73*	*0,77*	*0,67*	*0,81*	*1,37*	*0,87*	*0,70*	*0,90*	*0,71*
		Dĺžka2	*1,14*	*1,15*	*0,99*	*1,30*	*1,30*	*1,24*	*1,41*	*1,17*	*1,37*	*1,31*	*1,50*	*1,24*	*0,73*	*1,15*	*1,44*	*1,11*	*1,41*
	NON_2	Spoľahlivosť	91,1	90,7	94,3	86,7	86,7	88,3	84,2	89,0	85,0	86,6	82,0	87,1	98,5	91,3	83,6	90,7	80,9
		Dĺžka1	*0,88*	*0,88*	*1,03*	*0,78*	*0,78*	*0,82*	*0,74*	*0,91*	*0,74*	*0,78*	*0,70*	*0,85*	*1,37*	*0,87*	*0,70*	*0,90*	*0,73*
		Dĺžka2	*1,15*	*1,14*	*0,98*	*1,30*	*1,29*	*1,23*	*1,36*	*1,10*	*1,35*	*1,29*	*1,44*	*1,18*	*0,73*	*1,15*	*1,44*	*1,12*	*1,38*
80	LSE	Spoľ./Dĺžka1	95,06/0,49																
	NON_1	Spoľahlivosť	94,3	92,3	94,7	93,2	94,1	94,9	94,4	95,0	93,2	93,7	93,4	93,9	95,4	94,5	92,5	93,4	92,1
		Dĺžka1	*0,48*	*0,44*	*0,48*	*0,46*	*0,47*	*0,49*	*0,48*	*0,49*	*0,46*	*0,47*	*0,46*	*0,47*	*0,49*	*0,48*	*0,44*	*0,48*	*0,46*
		Dĺžka2	*1,03*	*1,11*	*1,01*	*1,07*	*1,04*	*1,01*	*1,03*	*1,00*	*1,07*	*1,05*	*1,07*	*1,04*	*1,00*	*1,03*	*1,11*	*1,03*	*1,06*
	NON_2	Spoľahlivosť	94,4	92,3	94,6	93,1	94,1	94,9	94,3	95,1	93,2	93,7	93,3	93,9	95,4	94,5	92,4	93,4	92,1
		Dĺžka1	*0,48*	*0,44*	*0,48*	*0,46*	*0,47*	*0,49*	*0,48*	*0,49*	*0,46*	*0,47*	*0,46*	*0,47*	*0,49*	*0,48*	*0,44*	*0,48*	*0,46*
		Dĺžka2	*1,03*	*1,11*	*1,01*	*1,07*	*1,04*	*1,01*	*1,03*	*1,00*	*1,07*	*1,05*	*1,07*	*1,04*	*1,00*	*1,03*	*1,11*	*1,02*	*1,06*
200	LSE	Spoľ./Dĺžka1	95,17/0,28																
	NON_1	Spoľahlivosť	95,4	93,7	95,1	94,6	95,4	95,7	95,5	95,8	94,8	95,1	94,9	95,2	95,2	95,3	94,3	94,4	94,1
		Dĺžka1	*0,28*	*0,26*	*0,28*	*0,27*	*0,28*	*0,29*	*0,28*	*0,29*	*0,27*	*0,28*	*0,28*	*0,28*	*0,28*	*0,28*	*0,27*	*0,28*	*0,28*
		Dĺžka2	*1,01*	*1,08*	*1,02*	*1,04*	*1,01*	*0,99*	*1,00*	*0,99*	*1,04*	*1,02*	*1,03*	*1,02*	*1,02*	*1,01*	*1,06*	*1,01*	*1,01*
	NON_2	Spoľahlivosť	95,4	93,7	95,1	94,6	95,4	95,7	95,5	95,8	94,8	95,1	95,0	95,3	95,2	95,3	94,3	94,3	94,0
		Dĺžka1	*0,28*	*0,26*	*0,28*	*0,27*	*0,28*	*0,29*	*0,28*	*0,29*	*0,27*	*0,28*	*0,28*	*0,28*	*0,28*	*0,28*	*0,27*	*0,28*	*0,28*
		Dĺžka2	*1,01*	*1,08*	*1,02*	*1,04*	*1,01*	*0,99*	*1,00*	*0,99*	*1,04*	*1,02*	*1,03*	*1,02*	*1,02*	*1,01*	*1,06*	*1,01*	*1,01*

Tab 5.2: Simulácie intervalov spoľahlivosti pri t_{10}

V Tab.5.3 zobrazujeme výsledky simulácií, v ktorých sme náhodné chyby vyberali zo Studentovho rozdelenia s 5 stupňami voľnosti. Napriek rozdeleniu s ťažšími chvostami pozorujeme, že LSE interval dosahuje stále veľmi dobré výsledky. Dokonca pri $n = 20$ sú lepšie ako ktorýkoľvek iný neparametrický interval spoľahlivosti. Pre väčšie n sú metódy medzi sebou porovnateľné a závery z Tab.5.3 sú analogické ako vyplývali z predošlých dvoch tabuliek Tab.5.1 a Tab.5.2. Na druhej strane v tomto prípade si môžeme všimnúť jednu anomáliu. V prípade $n = 20$ má interval spoľahlivosti s číslom 9 a 13 pri spôsobe odhadov regresných parametrov NON_1 priemernú dĺžku príliš širokú. Pokiaľ by spoľahlivosť bola na úrovni približne 100% bol by to jasný signál, že metóda zlyhala, pretože interval spoľahlivosti je absolútne neinformatívny. No teraz bola situácia iná. Spoľahlivosť bola porovnateľná s ostatnými metódami.

Začali sme pátrať po príčine, ktorú sme našťastie veľmi rýchlo objavili. Medzi 10 000 simuláciami došlo v dvoch prípadoch k zlyhaniu spomínaných dvoch metód, pretože šírka intervalov spoľahlivosti bola úplne uletená, rádovo v tisícoch. To spôsobilo, že aj priemer bol vychýlený a nezodpovedal celkom reálnym výsledkom. Z toho dôvodu sme sa pozreli na výsledky týchto dvoch intervalov spoľahlivosti pomocou robustnejšieho ukazovateľa, mediánu, ktorý určil mediánovu dĺžku intervalu spoľahlivosti s číslom 9 resp. 13 približne rovnú 0.878 resp. 0.829, čo sú už hodnoty porovnateľné so zvyšnými metódami. Prekvapujúci bol pre nás fakt, že ak sme odhadovali regresné parametre spôsobom NON_2, tak táto anomália nenastala.

Student_5			SCORE FUNCTION																
			Wilcoxon												Qnorm		Bent	
			Čísla metód pre odhady parametra τ (metódy NON_1 a NON_2)															
n	Metóda		2	3	4	5	6	7	8	9	10	11	12	13	OM	14	15	16	17
20	LSE	Spoľ./Dĺžka1	95,09/1,15																
	NON_1	Spoľahlivosť	92,3	90,1	94,0	86,8	85,7	87,5	81,6	87,8	83,7	85,5	78,9	86,0	98,4	92,0	83,5	91,4	79,4
		Dĺžka1	*1,01*	*0,95*	*1,10*	*0,84*	*0,83*	*0,87*	*0,76*	*4,02*	*0,79*	*0,82*	*0,72*	*3,16*	*1,53*	*1,00*	*0,77*	*1,04*	*0,75*
		Dĺžka2	*1,14*	*1,21*	*1,04*	*1,37*	*1,39*	*1,32*	*1,51*	*0,29*	*1,46*	*1,39*	*1,60*	*0,36*	*0,75*	*1,15*	*1,49*	*1,11*	*1,53*
	NON_2	Spoľahlivosť	92,3	90,5	94,1	86,9	86,1	87,8	83,2	88,3	84,2	85,9	81,0	86,3	98,5	91,7	83,3	91,5	80,6
		Dĺžka1	*1,01*	*0,95*	*1,11*	*0,84*	*0,84*	*0,88*	*0,79*	*0,98*	*0,80*	*0,83*	*0,75*	*0,91*	*1,53*	*1,00*	*0,77*	*1,03*	*0,77*
		Dĺžka2	*1,14*	*1,20*	*1,03*	*1,36*	*1,37*	*1,31*	*1,46*	*1,17*	*1,44*	*1,38*	*1,54*	*1,26*	*0,75*	*1,15*	*1,49*	*1,11*	*1,50*
80	LSE	Spoľ./Dĺžka1	94,99/0,56																
	NON_1	Spoľahlivosť	96,1	92,8	95,2	93,7	94,7	95,3	94,7	95,4	93,8	94,3	93,7	94,3	95,4	95,4	92,8	95,7	92,2
		Dĺžka1	*0,55*	*0,47*	*0,52*	*0,49*	*0,51*	*0,52*	*0,51*	*0,52*	*0,49*	*0,50*	*0,49*	*0,50*	*0,53*	*0,55*	*0,48*	*0,55*	*0,48*
		Dĺžka2	*1,03*	*1,19*	*1,09*	*1,15*	*1,11*	*1,08*	*1,11*	*1,08*	*1,15*	*1,13*	*1,15*	*1,13*	*1,07*	*1,03*	*1,16*	*1,02*	*1,17*
	NON_2	Spoľahlivosť	96,1	92,8	95,1	93,7	94,7	95,3	94,7	95,4	93,7	94,3	93,7	94,4	95,4	95,4	92,8	95,7	92,2
		Dĺžka1	*0,55*	*0,47*	*0,52*	*0,49*	*0,51*	*0,52*	*0,51*	*0,52*	*0,49*	*0,50*	*0,49*	*0,50*	*0,53*	*0,55*	*0,48*	*0,55*	*0,48*
		Dĺžka2	*1,03*	*1,19*	*1,09*	*1,15*	*1,12*	*1,08*	*1,11*	*1,08*	*1,15*	*1,13*	*1,15*	*1,13*	*1,07*	*1,03*	*1,16*	*1,02*	*1,17*
200	LSE	Spoľ./Dĺžka1	94,61/0,33																
	NON_1	Spoľahlivosť	96,4	93,6	94,9	94,3	95,2	95,6	95,4	95,6	94,6	94,9	94,7	94,9	95,1	95,8	93,8	96,1	93,9
		Dĺžka1	*0,32*	*0,28*	*0,30*	*0,29*	*0,30*	*0,31*	*0,30*	*0,31*	*0,29*	*0,30*	*0,29*	*0,30*	*0,30*	*0,32*	*0,29*	*0,32*	*0,29*
		Dĺžka2	*1,01*	*1,17*	*1,10*	*1,13*	*1,09*	*1,07*	*1,08*	*1,07*	*1,12*	*1,10*	*1,11*	*1,11*	*1,10*	*1,01*	*1,11*	*1,01*	*1,12*
	NON_2	Spoľahlivosť	96,4	93,6	94,9	94,4	95,2	95,6	95,4	95,6	94,6	94,9	94,7	94,9	95,1	95,8	93,9	96,2	93,9
		Dĺžka1	*0,32*	*0,28*	*0,30*	*0,29*	*0,30*	*0,31*	*0,30*	*0,31*	*0,29*	*0,30*	*0,29*	*0,30*	*0,30*	*0,32*	*0,29*	*0,32*	*0,29*
		Dĺžka2	*1,01*	*1,17*	*1,10*	*1,13*	*1,09*	*1,07*	*1,08*	*1,07*	*1,12*	*1,10*	*1,11*	*1,11*	*1,10*	*1,01*	*1,11*	*1,01*	*1,12*

Tab 5.3: Simulácie intervalov spoľahlivosti pri t_5

V Tab.5.4 sme zobrazili výsledky zo simulácií, ktoré boli reprezentované chybami so Studentovho rozdelenia s 1 stupňom voľnosti, ktorého chvosty sú oveľa ťažšie ako pri predošlých troch rozdeleniach. V tomto prípade už LSE interval spoľahlivosti podľa očakávaní zlyháva. Priemerná dĺžka je príliš široká a interval je pre nás nepoužiteľný. Podobne nedostatočné výsledky zaznamenávajú vo všetkých prípadoch aj neparametrické intervaly spoľahlivosti 2, 14 a 16, v ktorých sme neznámy parameter τ odhadovali len na základe smerodajnej odchýlky rezíduí. Tieto intervaly spoľahlivosti by sme neodporúčali používať, ak náhodné chyby pochádzajú z nejakého rozdelenia s ťažkými chvostami bez ohľadu na voľbu parametra n. Pozrime sa bližšie na prípady, v ktorých

sa počet meraní v lineárnom regresnom modeli n rovná 20. Ako prvé nás zaujalo, že interval spoľahlivosti s označením OM dosahuje prekvapujúco veľmi zlé výsledky, pretože je vďaka jeho šírke neinformatívny. Opäť to však bolo spôsobené tým, že v pár prípadoch z 10 000 boli hranice intervalu spoľahlivosti veľmi široké, čo vychýlilo priemernú šírku. Vieme, že Cauchyho rozdelenie je dané ako $(N(0,1))/(N(0,1))$. To znamená, že menovateľ sa často pohybuje blízko nuly, čo spôsobuje, že rozdelenia nadobúdajú občas príliš extrémne hodnoty. Vďaka tomu nastala situácia, že spomínaná metóda určila v niektorých prípadoch hranice intervalu spoľahlivosti, ktoré boli až príliš vzdialené od seba. Mediánová šírka bola rovná približne 8, čo je oveľa menej v porovnaní s 28, ale v porovnaní s inými neparametrickými metódami je to stále výrazne viac. Tento interval spoľahlivosti by sme teda pri veľmi malom n neodporúčali používať. Zaujímavosťou je, že ak sme n zvýšili na 34, tak ako bolo v článku [4], na základe ktorého sme interval spoľahlivosti vybudovali, priemerná dĺžka intervalu spoľahlivosti už bola porovnateľná so zvyšnými neparametrickými metódami. Najlepšie výsledky zaznamenáva interval spoľahlivosti 4, ktorý ako jediný odporúčame používať pri spomínanej konštelácií parametrov. Pri zvyšných neparametrických metódach sa prejavilo to, že požadovanú spoľahlivosť dosahujú len asymptoticky.

Student_1			SCORE FUNCTION																
			Wilcoxon												Qnorm		Bent	
			Čísla metód pre odhady parametra τ (metódy NON_1 a NON_2)															
n	Metóda		2	3	4	5	6	7	8	9	10	11	12	13	OM	14	15	16	17
20	LSE	Spoľ./Dĺžka1	95,12/37,40																
	NON_1	Spoľahlivosť	97,5	93,0	95,4	88,4	85,8	87,8	80,8	87,7	83,1	85,2	77,3	85,0	99,7	97,0	86,1	98,0	81,5
		Dĺžka1	*36,43*	*3,05*	*3,47*	*2,54*	*2,28*	*2,42*	*2,04*	*2,42*	*2,13*	*2,25*	*1,89*	*2,25*	*27,92*	*36,33*	*2,86*	*36,56*	*1,87*
		Dĺžka2	*1,03*	*12,27*	*10,79*	*14,71*	*16,43*	*15,48*	*18,34*	*15,46*	*17,59*	*16,65*	*19,75*	*16,63*	*1,34*	*1,03*	*13,06*	*1,02*	*20,00*
	NON_2	Spoľahlivosť	97,4	92,8	95,3	88,2	85,6	87,7	81,5	87,3	83,3	85,2	78,2	85,0	99,7	96,7	85,7	98,0	81,8
		Dĺžka1	*36,43*	*3,05*	*3,46*	*2,54*	*2,28*	*2,42*	*2,07*	*2,42*	*2,14*	*2,25*	*1,93*	*2,25*	*27,92*	*36,33*	*2,86*	*36,55*	*1,88*
		Dĺžka2	*1,03*	*12,28*	*10,81*	*14,70*	*16,37*	*15,43*	*18,02*	*15,48*	*17,50*	*16,59*	*19,33*	*16,62*	*1,34*	*1,03*	*13,07*	*1,02*	*19,87*
80	LSE	Spoľ./Dĺžka1	96,15/23,82																
	NON_1	Spoľahlivosť	100,0	94,1	95,5	93,8	94,7	95,5	93,5	95,0	93,5	94,1	92,1	93,8	94,7	100,0	93,6	100,0	92,2
		Dĺžka1	*23,76*	*0,85*	*0,91*	*0,85*	*0,87*	*0,90*	*0,84*	*0,88*	*0,83*	*0,85*	*0,80*	*0,84*	*0,93*	*23,75*	*0,98*	*23,76*	*0,70*
		Dĺžka2	*1,00*	*27,92*	*26,31*	*28,17*	*27,35*	*26,34*	*28,32*	*26,92*	*28,78*	*28,00*	*29,76*	*28,44*	*25,61*	*1,00*	*24,34*	*1,00*	*33,83*
	NON_2	Spoľahlivosť	100,0	94,2	95,4	93,9	94,7	95,5	93,5	95,1	93,6	94,2	92,1	93,9	94,7	100,0	93,7	100,0	92,2
		Dĺžka1	*23,76*	*0,85*	*0,90*	*0,85*	*0,87*	*0,90*	*0,84*	*0,88*	*0,83*	*0,85*	*0,80*	*0,84*	*0,93*	*23,75*	*0,98*	*23,76*	*0,70*
		Dĺžka2	*1,00*	*27,92*	*26,32*	*28,18*	*27,36*	*26,35*	*28,33*	*26,93*	*28,79*	*28,01*	*29,77*	*28,45*	*25,62*	*1,00*	*24,35*	*1,00*	*33,84*
200	LSE	Spoľ./Dĺžka1	96,3/20,66																
	NON_1	Spoľahlivosť	100,0	94,3	95,0	94,4	95,4	96,0	93,6	95,5	94,6	94,9	92,4	94,7	94,5	100,0	94,2	100,0	93,8
		Dĺžka1	*20,66*	*0,47*	*0,49*	*0,47*	*0,50*	*0,51*	*0,47*	*0,50*	*0,48*	*0,49*	*0,45*	*0,48*	*0,49*	*20,65*	*0,55*	*20,66*	*0,40*
		Dĺžka2	*1,00*	*44,12*	*42,56*	*43,64*	*41,61*	*40,43*	*44,27*	*41,44*	*43,32*	*42,50*	*46,07*	*43,21*	*42,20*	*1,00*	*37,23*	*1,00*	*51,32*
	NON_2	Spoľahlivosť	100,0	94,4	95,1	94,5	95,4	96,1	93,6	95,5	94,8	95,0	92,5	94,8	94,5	100,0	94,4	100,0	94,0
		Dĺžka1	*20,66*	*0,47*	*0,49*	*0,47*	*0,50*	*0,51*	*0,47*	*0,50*	*0,48*	*0,49*	*0,45*	*0,48*	*0,49*	*20,66*	*0,55*	*20,66*	*0,40*
		Dĺžka2	*1,00*	*44,14*	*42,57*	*43,65*	*41,62*	*40,44*	*44,28*	*41,45*	*43,33*	*42,51*	*46,08*	*43,22*	*42,21*	*1,00*	*37,24*	*1,00*	*51,33*

Tab 5.4: Simulácie intervalov spoľahlivosti pri t_1

Ďalej sa pozrieme na simulácie, v ktorých sme parameter n nastavili na hodnotu 80

alebo 200. Našťastie v týchto prípadoch už interval spoľahlivosti s označením OM dosahuje veľmi dobré výsledky tak ako aj iné neparametrické metódy. V prípade, že $n = 80$ odporúčame použiť metódy 3, 4, 6, 7, 9, 11. To znamená, že použitie *Wilcoxon score function* sa vo všeobecnosti ukazuje ako rozumnejšia voľba v porovnaní s ostatnými dvomi *score function*. V situácií ak $n = 200$ je vhodných viac intervalov spoľahlivosti ako napr. 3, 4, 5, 6, 7, 9, 10, 11, 13 a 15. Priemerné šírky uvedených intervalov spoľahlivosti sú medzi sebou viac menej porovnateľné. Aj v týchto záverečných simuláciách sa potvrdilo, že neparametrické intervaly spoľahlivosti fungujú lepšie ako klasický parametrický interval spoľahlivosti, ak rozdelenie náhodných chýb má príliš ťažké chvosty, čo bol jeden z cieľov našej diplomovej práce. Podarilo sa nám preukázať, že použitie neparametrických metód má význam.

5.3 Simulácie intervalov spoľahlivosti metódou bootstrap

V tejto časti v krátkosti rozoberieme výsledky simulácií intervalov spoľahlivosti pre regresný parameter, ktoré sme vypočítali na základe metódy bootstrap. Konštrukcie takýchto bootstrapových intervalov spoľahlivosti sme opísali v 2. kapitole. Hlavný stavebný kameň takýchto intervalov spoľahlivosti je bodový odhad daného parametra, ktorý vypočítame veľmi veľakrát na umelo vytvorených dátach. Myšlienka týchto metód je preto veľmi jednoduchá, no výpočtová náročnosť oveľa zložitejšia. V dnešnej dobe vyspelej výpočtovej techniky však nie je problém toto obmedzenie prekonať a porovnať výsledky simulácií intervalov spoľahlivosti s výsledkami z predošlej časti, čo je hlavným cieľom v tejto záverečnej časti. Kvôli prehľadnosti sme sa rozhodli uviesť výsledky týchto metód v ďalších štyroch tabuľkách, pretože by bol problém uviesť taký veľký počet metód len do štyroch tabuliek. V nasledujúcich tabuľkách v riadku Dĺžka2 opäť informujeme o tom, koľkokrát je priemerná dĺžka jednotlivých bootrstrapových intervalov menšia v porovnaní s LSE intervalom spoľahlivosti z časti 5.2. Parametre simulácií sme menili tak ako doteraz, taktiež generovanie umelých dát sme zachovali. Znížili sme však počet simulácii z 10 000 na 1000, pričom v každej simulácií sme vytvorili 1000 nových bootstrapových sád z pôvodnej dátovej sady, z ktorých sme zrátali 1000 odhadov potrebného regresného parametra. Bootstrapove sady pozostávali z pôvodných n dátových záznamov, pričom záznamy do nich sa vyberali s opakovaním.

Odhady regresných parametrov sme v tejto časti realizovali už len pomocou funkcie z balíka *Rfit*, pretože ako už vieme z 3. kapitoly, rozdiely medzi touto funkciou a našou vlastnou funkciou neboli vôbec výrazne, a preto sme v použití oboch funkcií nevideli hlbší zmysel. Taktiež pri týchto simuláciách sme využili aj funkciu *boot* z balíka *Rfit*, ktorá mala v sebe zahrnutý aj výpočet BCA intervalu spoľahlivosti, v ktorom bolo potrebné odhadnutie ďalších dvoch dodatočných parametrov. Dôvod bol jednoduchý, a síce zvýšená výpočtová náročnosť týchto simulácií. Predpokladali sme, že v balíku *boot* sú spôsoby odhadnutia týchto dvoch dodatočných parametrov efektívne implementované a jeho použitím ušetríme potrebný čas. V jednotlivých tabuľkách označenie IS_1 reprezentuje bootstrapový interval spoľahlivosti daný vzťahom 2.45, IS_2 predstavuje bootstrapový interval spoľahlivosti zo vzťahu 2.46, IS_3 zodpovedá percentilovému intervalu spoľahlivosti a IS_4 prináleží interval spoľahlivosti s označením BCA z 2. kapitoly. Pri metóde LSE zobrazujeme taktiež 4 typy bootstrapových intervalov v poradí IS_1/IS_2/IS_3/IS_4.

V Tab.5.5 zobrazujeme výsledky simulácií intervalov spoľahlivosti, ak sme náhodné chyby generovali zo štandardizovaného normálneho rozdelenia. Podľa očakávaní aj medzi týmito metódami víťazí metóda založená na odhade regresného parametra pomocou metódy najmenších štvorcov, pretože oproti zvyšným metódam sú jej intervaly spoľahlivosti najinformatívnejšie. Potešujúci je pre nás fakt, že väčšina uvedených intervalov spoľahlivosti dosahuje požadovanú spoľahlivosť. Medzi neparametrickými metódami najlepšie výsledky zaznamenávajú intervaly spoľahlivosti s *Kvantil score function*, pretože dosahujú najmenšiu šírku. Naším prioritným cieľom je však tieto parametrické a neparametrické intervaly spoľahlivosti založené na bootstrap metóde porovnať s intervalmi spoľahlivosti z Tab.5.1. Môžeme si všimnúť, že intervaly spoľahlivosti v prípade $n = 80$ alebo $n = 200$ z oboch tabuliek dosahujú porovnateľné výsledky aj čo sa týka spoľahlivosti, tak aj priemernej šírky. Teda môžeme konštatovať, že v týchto prípadoch sú intervaly spoľahlivosti založené na metóde bootstrap zaujímavou a aj vhodnou alternatívou k intervalom uvedeným v Tab.5.1. V prípade $n = 20$ pozorujeme, že intervaly spoľahlivosti z Tab.5.5 dosahujú na rozdiel od väčšiny intervalov z Tab.5.1 požadovanú spoľahlivosť. Na druhej strane sa táto skutočnosť prejavila v tom, že ich priemerná dĺžka je väčšia. To znamená, že sú menej informatívne, ale spo-

ľahlivejšie. V tomto prípade stále ostáva platné, že najvhodnejší interval spoľahlivosti považujeme klasický parametrický interval z Tab.5.1 a v rámci neparametrického prístupu prvenstvo ostáva intervalu spoľahlivosti s číslom 4.

N(0,1)			SCORE FUNCTION											
			Wilcoxon				Qnorm				Bent			
n	Metóda		IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4
20	LSE	Spoľahlivosť	96,5/96,2/95,4/95,4											
		Dĺžka1	*1,07/1,07/1,07/1,09*											
		Dĺžka2	*0,85/0,85/0,85/0,83*											
	NON_1	Spoľahlivosť	98,3	97,3	98,1	97,2	97,6	96,8	97,6	96	99,3	97,3	98,9	97,4
		Dĺžka1	*1,26*	*1,26*	*1,25*	*1,28*	*1,19*	*1,19*	*1,17*	*1,24*	*1,44*	*1,44*	*1,42*	*1,48*
		Dĺžka2	*0,72*	*0,72*	*0,73*	*0,71*	*0,76*	*0,76*	*0,77*	*0,73*	*0,63*	*0,63*	*0,64*	*0,61*
80	LSE	Spoľahlivosť	94,8/95/94,4/94,8											
		Dĺžka1	*0,43/0,43/0,43/0,44*											
		Dĺžka2	*1,02/1,02/1,02/1,01*											
	NON_1	Spoľahlivosť	95,2	94,8	95,8	95,7	94,5	94,3	94,7	95,2	96,2	96	96,5	96,4
		Dĺžka1	*0,46*	*0,46*	*0,46*	*0,47*	*0,44*	*0,44*	*0,44*	*0,45*	*0,52*	*0,52*	*0,51*	*0,52*
		Dĺžka2	*0,95*	*0,95*	*0,96*	*0,95*	*1,00*	*1,00*	*1,00*	*0,99*	*0,86*	*0,86*	*0,86*	*0,85*
200	LSE	Spoľahlivosť	95,1/95,1/95,2/94,9											
		Dĺžka1	*0,25/0,25/0,25/0,25*											
		Dĺžka2	*1,01/1,01/1,02/1*											
	NON_1	Spoľahlivosť	95,6	95,7	95,6	95,5	95	95,1	94,9	95,2	95,3	95,4	95,9	95,8
		Dĺžka1	*0,26*	*0,26*	*0,26*	*0,26*	*0,25*	*0,25*	*0,25*	*0,25*	*0,29*	*0,29*	*0,29*	*0,29*
		Dĺžka2	*0,97*	*0,97*	*0,97*	*0,97*	*1,01*	*1,01*	*1,01*	*1,00*	*0,89*	*0,89*	*0,89*	*0,89*

Tab 5.5: Simulácie bootstrapových intervalov spoľahlivosti pri N(0,1)

V Tab.5.6 sme uviedli výsledky ďalších simulácií, v ktorých náhodné chyby pochádzali zo Studentovho rozdelenia s 10 stupňami voľnosti. Opäť vidíme, že bootstrapové intervaly spoľahlivosti vychádzajúce z metódy LSE dosahujú veľmi dobré výsledky. Čo sa týka neparametrického prístupu, môžeme si všimnúť, že použitím *Kvantil score function* získame najužšie bootstrapové intervaly spoľahlivosti v porovnaní so zvyšnými dvomi *score function*, čo sa ale prejavilo v nižšej spoľahlivosti týchto intervalov najmä v prípade $n = 200$. Pri pohľade na Tab.5.6 a Tab.5.2 vidíme, že platia analogické úvahy a závery, aké sme vyvodili opise predchádzajúcich výsledkov a bolo by zbytočné ich opätovne uvádzať.

V Tab.5.7 dokumentujeme výsledky simulácií, v ktorých sa jednotlivé metódy museli vysporiadať s rozdelením náhodných chýb, ktoré sme reprezentovali Studentovým rozdelením s 5 stupňami voľnosti. Rovnako ako pri metódach nevyužívajúc bootstrap prístup aj teraz platí, že akonáhle prechádzame z rozdelenia s ľahšími chvostami na rozdelenie s ťažšími chvostami, prejaví sa tým, že šírky bootstrapových intervalov spoľahlivosti sa zväčšujú. Zdôvodnenie sme uviedli už viackrát v predošlých častiach, a preto ho uvádzať už nebudeme. Z Tab.5.7 vidíme, že v prípadoch $n = 80$ a $n = 200$

Student_10			SCORE FUNCTION											
			Wilcoxon				Qnorm				Bent			
n	Metóda		IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4
20	LSE	Spoľahlivosť	97,4/97/96,6/96,7											
		Dĺžka1	*1,20/1,20/1,20/1,22*											
		Dĺžka2	*0,84/0,84/0,84/0,82*											
	NON_1	Spoľahlivosť	98,5	97,7	98,5	97,6	98,1	97,8	97,9	97,4	99,1	98	99,2	98
		Dĺžka1	*1,39*	*1,39*	*1,38*	*1,42*	*1,32*	*1,32*	*1,31*	*1,38*	*1,58*	*1,58*	*1,56*	*1,63*
		Dĺžka2	*0,72*	*0,72*	*0,73*	*0,71*	*0,76*	*0,76*	*0,77*	*0,73*	*0,63*	*0,63*	*0,64*	*0,62*
80	LSE	Spoľahlivosť	95,2/95/94,9/94,7											
		Dĺžka1	*0,48/0,48/0,48/0,49*											
		Dĺžka2	*1,02/1,02/1,02/1,01*											
	NON_1	Spoľahlivosť	95,5	95	95,6	95,5	95	94,5	94,8	94,6	95,1	94,7	95,9	95,5
		Dĺžka1	*0,50*	*0,50*	*0,50*	*0,50*	*0,49*	*0,49*	*0,49*	*0,49*	*0,54*	*0,54*	*0,54*	*0,54*
		Dĺžka2	*0,98*	*0,98*	*0,99*	*0,98*	*1,01*	*1,01*	*1,01*	*1,00*	*0,90*	*0,90*	*0,91*	*0,90*
200	LSE	Spoľahlivosť	94,2/94,4/93,9/94,1											
		Dĺžka1	*0,28/0,28/0,28/0,28*											
		Dĺžka2	*1,01/1,01/1,02/1*											
	NON_1	Spoľahlivosť	94,2	93,7	94,5	94,1	94	93,9	93,9	93,8	95	94,4	95,5	94,9
		Dĺžka1	*0,28*	*0,28*	*0,28*	*0,28*	*0,28*	*0,28*	*0,28*	*0,28*	*0,30*	*0,30*	*0,30*	*0,30*
		Dĺžka2	*1,01*	*1,01*	*1,02*	*1,01*	*1,02*	*1,02*	*1,03*	*1,02*	*0,95*	*0,95*	*0,96*	*0,95*

Tab 5.6: Simulácie bootstrapových intervalov spoľahlivosti pri t_{10}

takmer všetky metódy dosahujú spoľahlivosť na úrovni 95%. Môžeme spozorovať, že v týchto prípadoch už metóda LSE nedosahuje najužšie bootstrapové intervaly spoľahlivosti, ale prekonávajú ju metódy NON_1 s *Wilcoxon score function* a *Kvantil score function*, čo sa do istej miery prejavuje práve rozdelenie s častejším výskytom outlierov.

Student_5			SCORE FUNCTION											
			Wilcoxon				Qnorm				Bent			
n	Metóda		IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4
20	LSE	Spoľahlivosť	97,3/96,4/96,3/95,5											
		Dĺžka1	*1,34/1,34/1,34/1,38*											
		Dĺžka2	*0,85/0,85/0,86/0,83*											
	NON_1	Spoľahlivosť	98,5	97,4	98,2	97,7	98,2	97,4	97,9	96,4	99,2	98	98,9	97,5
		Dĺžka1	*1,54*	*1,54*	*1,52*	*1,58*	*1,46*	*1,46*	*1,45*	*1,52*	*1,74*	*1,74*	*1,72*	*1,79*
		Dĺžka2	*0,75*	*0,75*	*0,75*	*0,73*	*0,78*	*0,78*	*0,79*	*0,75*	*0,66*	*0,66*	*0,67*	*0,64*
80	LSE	Spoľahlivosť	94,8/94,2/93,4/93,3											
		Dĺžka1	*0,55/0,55/0,55/0,56*											
		Dĺžka2	*1,03/1,03/1,03/1,01*											
	NON_1	Spoľahlivosť	94,9	94,3	95,1	94,1	94,9	94,8	94,4	94	94,8	94,4	95,4	94,8
		Dĺžka1	*0,54*	*0,54*	*0,54*	*0,54*	*0,54*	*0,54*	*0,54*	*0,54*	*0,58*	*0,58*	*0,58*	*0,58*
		Dĺžka2	*1,05*	*1,05*	*1,05*	*1,04*	*1,05*	*1,05*	*1,05*	*1,04*	*0,98*	*0,98*	*0,98*	*0,98*
200	LSE	Spoľahlivosť	95,6/95,6/94,9/94,4											
		Dĺžka1	*0,32/0,32/0,32/0,33*											
		Dĺžka2	*1,01/1,01/1,01/1*											
	NON_1	Spoľahlivosť	95,2	95,4	95,5	95,9	96,2	95,6	95,8	95,8	94,8	94,4	95,7	94,9
		Dĺžka1	*0,30*	*0,30*	*0,30*	*0,30*	*0,31*	*0,31*	*0,31*	*0,31*	*0,31*	*0,31*	*0,31*	*0,32*
		Dĺžka2	*1,08*	*1,08*	*1,09*	*1,08*	*1,07*	*1,07*	*1,07*	*1,06*	*1,04*	*1,04*	*1,05*	*1,04*

Tab 5.7: Simulácie bootstrapových intervalov spoľahlivosti pri t_5

Dokonca v prípade $n = 200$ aj použitie *Bent score function* nám zabezpečí informatívnejšie bootstrapové intervaly spoľahlivosti ako metóda LSE. V prípade $n = 20$ najužšie

bootstrapové intervaly zabezpečuje metóda LSE nasledovaná neparametrickou metódou s *Kvantil score function*. Pri porovnaní výsledkov s výsledkami z Tab.5.3 vidíme, že v prípade $n = 80$ a $n = 200$ sú intervaly spoľahlivosti založené na metóde bootstrap vhodnou a zaujímavou alternatívou, pretože sú porovnateľné na základe spoľahlivosti a priemernej šírky. V prípade $n = 20$ stále platí, že priemerná šírka bootstrapových intervalov z Tab.5.7 je o niečo väčšia, no kompenzujú to vyššou mierou spoľahlivosti.

V poslednej tabuľke s označením Tab.5.8 uvádzame výsledky simulácií, v ktorých náhodne chyby pochádzali z Cauchyho rozdelenia. V prvom kroku si rozoberieme výsledky vzťahujúce sa na prípad $n = 20$. Z nižšie uvedenej tabuľky vidíme, že všetky bootstrapové intervaly spoľahlivosti sú veľmi široké. To znamená, že nie sú vhodné na praktické použitie. V tomto prípade odporúčame použiť intervaly spoľahlivosti, ktoré neboli vypočítané na základe metódy boostrap. Konkrétne ide o interval s číslom 4, ktorý sme spomínali pri opise výsledkov z Tab.5.4. V situácií, kde $n = 80$ resp. $n = 200$, bootstrapové intervaly spoľahlivosti založené na odhade regresného parametra pomocou metódy najmenších štvorcov podľa očakávaní zlyhali. Z toho dôvodu sa ďalej pozrieme už len na bootstrapové intervaly spoľahlivosti vzťahujúce k metóde NON_1. Vidíme, že na rozdiel od predošlých výsledkov teraz dosahujú najlepšie štatistiky bootstrapové intervaly spoľahlivosti s *Bent score function*. Ak porovnáme tieto výsledky s Tab.5.4 všimneme si, že v prípade $n = 80$ sú intervaly z Tab.5.4 o niečo užšie, t. j. viac informatívne. Síce majú o trochu nižšiu spoľahlivosť, no bootstrapové intervaly spoľahlivosti z Tab.5.8 v prípade $n = 80$ považujeme zbytočne za príliš široké. Situácia sa však trošku mení, ak porovnávame intervaly spoľahlivosti z Tab.5.4 a Tab.5.8 v prípade $n = 200$.

Vidíme, že priemerné dĺžky intervalov spoľahlivosti sú menej odlišné. Hlavne ak sa zameriame na bootstrapové intervaly spoľahlivosti z Tab.5.8, pri ktorých sme použili *Bent score function* alebo *Wilcoxon score function*. Obe tieto možnosti sú veľmi dobrou alternatívou k intervalom spoľahlivosti, ktoré sme odporučili používať pri danej konštelácii parametrov v opise výsledkov z Tab.5.4. Najmä použitie ktoréhokoľvek bootstrapového intervalu spoľahlivosti z Tab.5.8 pri *Bent score function* považujeme za excelentnú voľbu, pretože tieto bootstrapové intervaly spoľahlivosti majú v porovnaní s konkurenciou z Tab.5.4 a Tab.5.8 najmenšiu šírku, teda sú najinformatívnejšie

pre používateľa a súčasne ich spoľahlivosť prevyšuje 95%. Z týchto dôvod považujeme práve ktorýkoľvek z týchto bootstrapových intervalov za najlepšie použitie v danej situácií.

Student_1			SCORE FUNCTION											
			Wilcoxon				Qnorm				Bent			
n	Metóda		IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4	IS_1	IS_2	IS_3	IS_4
20	LSE	Spoľahlivosť	99,1/99/97,4/94,7											
		Dĺžka1	***22,43/22,43/21,35/27,11***											
		Dĺžka2	***1,67/1,67/1,75/1,38***											
	NON_1	Spoľahlivosť	99,8	99,5	99	98,4	99,6	99,5	98,7	98,1	100	99,9	99,8	99,4
		Dĺžka1	***18,06***	***18,06***	***14,42***	***19,91***	***18,01***	***18,01***	***15,69***	***21,17***	***18,26***	***18,26***	***13,07***	***20,61***
		Dĺžka2	***2,07***	***2,07***	***2,59***	***1,88***	***2,08***	***2,08***	***2,38***	***1,77***	***2,05***	***2,05***	***2,86***	***1,81***
80	LSE	Spoľahlivosť	98,3/98,1/93/84,3											
		Dĺžka1	***20,40/20,40/18,13/25,93***											
		Dĺžka2	***1,17/1,17/1,31/0,92***											
	NON_1	Spoľahlivosť	98,6	99	95,2	96,2	98,4	98,8	95	95,2	98,3	98,5	96,7	97,4
		Dĺžka1	***1,06***	***1,06***	***1,06***	***1,07***	***1,24***	***1,24***	***1,25***	***1,26***	***0,94***	***0,94***	***0,94***	***0,94***
		Dĺžka2	***22,49***	***22,49***	***22,39***	***22,33***	***19,18***	***19,18***	***19,08***	***18,89***	***25,33***	***25,33***	***25,22***	***25,27***
200	LSE	Spoľahlivosť	98,7/98,8/92,1/79,5											
		Dĺžka1	***12,80/12,80/11,60/17,27***											
		Dĺžka2	***1,61/1,61/1,78/1,20***											
	NON_1	Spoľahlivosť	97,4	98	96,7	96,9	97,8	98,2	95,9	96,8	96,7	96,9	96,2	96,3
		Dĺžka1	***0,51***	***0,51***	***0,51***	***0,51***	***0,60***	***0,60***	***0,60***	***0,60***	***0,45***	***0,45***	***0,45***	***0,45***
		Dĺžka2	***40,57***	***40,57***	***40,53***	***40,23***	***34,54***	***34,54***	***34,48***	***34,21***	***46,39***	***46,39***	***46,37***	***46,12***

Tab 5.8: Simulácie bootstrapových intervalov spoľahlivosti pri t_1

V tejto záverečnej časti sme mohli vidieť, že metóda bootstrap sa vo väčšine prípadov ukázala ako vhodná alternatíva k vybudovaniu potrebných intervalov spoľahlivosti. Vďaka kvalitnej výpočtovej technike nie je v dnešnej dobe problém aplikovať uvedené postupy založené na metóde bootstrap, ktorých daň je vyššia výpočtová náročnosť, no na druhej strane myšlienka ich odvodenia je veľmi jednoduchá, priamočiara a zrozumiteľná.

Záver

Naša diplomová práca sa venovala porovnaniu parametrických a neparametrických metód v lineárnej regresii s viacerými vysvetľujúcimi premennými. Prioritným cieľom bolo prezentovať výhodu použitia neparametrickej lineárnej regresie v prípade, že nie je splnený predpoklad normálneho rozdelenia náhodných chýb v klasickej t. j. parametrickej lineárnej regresii.

V získaných výsledkov môžeme jednoznačne vidieť, že spomínaná výhoda neparametrických metód sa naozaj prejavila. Napríklad pri Cauchyho rozdelení, ktoré zabezpečuje oveľa častejší výskyt outlierov ako normálne rozdelenie, sme mohli vidieť, že parametrické metódy sú oveľa horšie ako neparametrické metódy bez ohľadu na to, aký veľký dátový súbor, akú *score function* alebo aký spôsob odhadu parametra τ, vystupujúci v neparametrických metódach, sme uvažovali. Prínos a výhody neparametrikcých metód sme badali v záveroch v 3. kapitole pri počítaní odhadov regresných parametrov, taktiež aj v 4. kapitole pri testovaní ich signifikantnosti a v neposlednom rade aj v záverečnej 5. kapitole, v ktorej sme konštruovali intervalové odhady pre regresné parametre. V 5. kapitole sme zistili, že v prípade rozdelenie náhodných chýb s ťažkými chvostami a pri malej dátovej vzorke zaznamel interval OM z článku [4] nie veľmi uspokojivé výsledky v porovaní s konkurenciou. Vidíme, že má zmysel poznať aj neparametrické metódy lineárnej regresie, ktoré sú vhodnou alternatívou ku klasickej lineárnej regresii. Taktiež z uvedeného vyplýva, že je veľmi dôležité pred samotným aplikovaním niektorej z klasických regresných metód overiť predpoklad normálneho rozdelenia náhodných chýb. Pokiaľ tak neurobíme, môže zbytočne dôjsť k nesprávnej interpretácii výsledkov bez toho, aby sme si to uvedomili.

Prínosom našej diplomovej práce sú vykonané simulácie v softwari R, na základe ktorých sme vyhodnocovali kvalitu jednotlivých metód. Oproti inými prácam je naše porovnanie oveľa komplexnejšie, pretože sme menili okrem rozdelenia náhodných chýb aj veľkosť dátovej vzorky a pri neparametrických metódach aj tzv. *score function*, spôsob odhadu neznámeho parametra τ a skúmali sme vplyv týchto zmien. Taktiež sme porovnávali neparametrickú metódu z balíka *Rfit* s našou základnou neparametrickou metódou. Pri práci so spomínaným balíkom sme zistili isté nezrovnalosti medzi popisom niektorých funkcií a ich reálnou implementáciou v balíku. Tento fakt prispel k

tomu, že sme napísali jednému z autorov balíka *Rfit*, ktorý nám veľmi ochotne odpovedal na naše otázky a objasnil nezhody, ktoré sme našli. Aj túto skutočnosť môžeme považovať za prínos našej práce, pretože autor balíka J.W. McKean priznal, že manuál k balíku je zastaralý a potreboval by aktualizáciu. Je teda pravdepodobné, že aj vďaka našim pripomienkam dôjde pri nasledujúcej aktualizácii balíka *Rfit* k úprave popisu niektorých funkcií. Veríme, že by to prispelo k eliminácii situácií pre ďalších používateľov balíka *Rfit*, v ktorých by museli pátrať po tom, čo v skutočnosti niektoré funkcie z balíka robia. Niekedy takéto hľadanie môže zbytočne zabrať pomerne veľa času.

O komplexnosti našej práce svedčí aj fakt, že porovnanie parametrických a neparametrických metód sme vykonali z viacerých pohľadov. Ako sme už spomínali najprv sme kvalitu metód merali na základe odhadov neznámych regresných parametrov. V ďalšej fáze sme sa pozreli na test hypotézy o významnosti regresných parametrov uskutočnený pomocou parametrického a neparametrického prístupu a nakoniec sme sa zaoberali konštrukciou parametrického a neparametrického intervalu spoľahlivosti pre regresné parametre. Takéto komplexné porovnanie je podľa nášho názoru oveľa prínosnejšie a objektívnejšie pre vyvodenie záverov pri hodnotení parametrickej a neparametrickej lineárnej regresie.

Zoznam použitej literatúry

[1] Bhattacharyya, G. K., Roussas G. G.: *Estimation of a certain functional of a probability density function*, Scandinavian Actuarial Journal 3-4 (1969), 201-206

[2] George, K. J., McKean, J. W., Schucany, W. R., and Sheather, S. J.: *A comparison of condence intervals from R-estimators in regression*, Journal of Statistical Computation and Simulation 53 (1995), 13-22

[3] Hall, P, Sheather, S.J., Jones, M.C., Marron, J.S.: *An optimal data-based bandwidth selection in kernel density estimation*, Biometrika 78 (1991), 263-269

[4] Hettmansperger, T. J., McKean, J.W.: *Robust Nonparametric Statistical Methods*, CRC Press, New York, 2011

[5] Hollander, M., Wolfe, D.A., Chicken, E.: *Nonparametric Statistical Methods*, WILEY, New Yersey, 2014

[6] Jaeckel, L. A.: *Estimating regression coefficients by minimizing the dispersion of the residuals*, Ann. Math. Statist. 43 (1972), 1449-1458

[7] Johnson, J., DiNardo, J.: *Econometric Methods*, McGraw Hill, New York, 1997

[8] Kloke, J.D., Mckean, J.W.: *Rfit: Rank-based estimation for linear models*, The R Journal 4 (2012), 57-64

[9] Koul, H.L., Sievers, G., McKean, J.W.: *An Estimator of the Scale Parameter for the Rank Analysis of Linear Models under General Score Functions*, Scandinavian Journal of Statistics, 14 (1987), 131-141

[10] McKean, J. W., Hettmansperger, T. P.: *Tests of Hypothesis in the General Linear Model Based on Ranks*, Communications in Statistics 8 (1976), 693-709

[11] McKean, J. W., Sheather, S. J.: *Small sample properties of robust analyses of linear models based on R-estimates: a survey. Directions in Robust Statistics and Diagnostics*, Springer-Verlag (1991), 1-20

[12] Omelka, M.: *Comparison of two types of confidence intervals based on Wilcoxon-type R-estimators*, Statistics and Probability Letters 78 (2008), 3366–3372

[13] Silverman, B.W.: *Density Estimation for Statistics and Data Analysis*, Champan and Hall, London, 1986

[14] Sheather, S. J., Jones, M.C.: *A reliable data-based bandwidth selectionmethod for kernel density estimation*, Journal of Royal Statistical Society 53 (1991), 683-690

[15] Schuster, E.: *On the rate of convergence of an estimate of a functional of a probability density*, Scandinavian Actuarial Journal 1 (1974), 103-107

[16] Schweder, T.: *Window estimation of the asymptotic variances of rank estimators of location*, Scandinavian Journal of Statistics 2 (1975), 113-126

[17] Zvára, K.: *Regrese*, MatfyzPress, Praha, 2008

Príloha

V prílohe zobrazujeme zoznam funkcií, ktoré sme vytvorili a následne využívali v softwari R pri vypracovaní našej diplomovej práce.

```
simulacie_1 <- function(N,p,n,SCORES,epsilon,epsilon_grad)
{
  set.seed(5)
  rozsah_vyberu <- seq(from=-5,to=5,by=0.01)
  koef <- sample(rozsah_vyberu,size = p+1,replace = FALSE)
  set.seed(6)
  X <- matrix(runif(p*n,min = 0,max = 5),ncol=p)
  INTERCEPT <- rep(koef[1],n)
  X_pomocna <- cbind(rep(1,times=n),X)
  v1<-NA
  v2<-NA
  v3<-NA
  set.seed(3)
  for (i in (1:N))
  {
    e <- rnorm(n,mean = 0,sd = 1)
    Y <- INTERCEPT + X%*%koef[2:(p+1)] + e
    koef_LS_estimation <- solve(t(X_pomocna)%*%X_pomocna)%*%t(X_pomocna)
    %*%Y
    koef_nonparametric_estimation1 <- rfit(Y ~ .,scores = SCORES,
    data = as.data.frame(X),symmetric = TRUE)$coefficients
    koef_nonparametric_estimation2 <- tryCatch(BFGS(Y,X,epsilon,
    epsilon_grad,
    koef_LS_estimation[2:(p+1)],SCORES)[2:(p+1)],error=function(e)
    BFGS(Y,X,epsilon,epsilon_grad,koef_nonparametric_estimation1[2:(p+1)
    ],SCORES)[2:(p+1)])
    e_star <- Y - X%*%koef_nonparametric_estimation2
    koef_nonparametric_estimation2 <- c(median(walsh(e_star)),
    koef_nonparametric_estimation2)
    v1[i] <- mean((koef-koef_LS_estimation)^2)
    v2[i] <- mean((koef-koef_nonparametric_estimation1)^2)
    v3[i] <- mean((koef-koef_nonparametric_estimation2)^2)
  }
  my_list <- list("LSE" = v1, "nonE1_RFIT" = v2, "nonE1" = v3, "N" = N)
  return(my_list)
}

simulacie_2 <- function(N,p,q,n,SCORES,epsilon1,epsilon_grad1,epsilon2,
epsilon_grad2,Num_of_tau,pomocna)
{
  set.seed(5)
  rozsah_vyberu <- seq(from=-5,to=5,by=0.01)
  Koef_full <- sample(rozsah_vyberu,size = p+1,replace = FALSE)
  Koef_full[(q+1)] <- rep(C,times = length(q))
  set.seed(6)
  X_full <- matrix(runif(p*n,min = 0,max = 5),ncol=p)
  X_rest <- X_full[,-q]
  X_rest <- as.matrix(X_rest)
  INTERCEPT <- rep(Koef_full[1],n)
  X_pomocna_full <- cbind(rep(1,times=n),X_full)
  X_pomocna_rest <- cbind(rep(1,times=n),X_rest)
  v2 <- NA
  v3 <- NA
  V1 <- NA
  V2 <- matrix(NA,ncol = Num_of_tau,nrow = N)
  V3 <- matrix(NA,ncol = Num_of_tau,nrow = N)
  set.seed(3)
  for (i in 1:N)
  {
    e <- rnorm(n,mean=0,sd=1)
    Y <- INTERCEPT + X_full%*%Koef_full[2:(p+1)] + e
    koef_full_LSE <- solve(t(X_pomocna_full)%*%X_pomocna_full)%*%
    t(X_pomocna_full)%*%Y
    koef_rest_LSE <- solve(t(X_pomocna_rest)%*%X_pomocna_rest)%*%
    t(X_pomocna_rest)%*%Y
    RSS_full <- sum((Y - X_pomocna_full%*%koef_full_LSE)^2)
    RSS_rest <- sum((Y - X_pomocna_rest%*%koef_rest_LSE)^2)
    V1[i] <- ((RSS_rest-RSS_full)/length(q))/(RSS_full/(n-p-1))
    koef_full <- rfit(Y ~ .,scores = SCORES,data = as.data.frame(X_full)
    ,symmetric = TRUE)$coefficients
    Jaeckel_dispersion_full <- JD_fun(Y,X_full,koef_full[2:length
    (koef_full)], SCORES)
    if (length(q)<p)
    {
      koef_rest <- rfit(Y ~ .,scores = SCORES,data = as.data.frame
      (X_rest),symmetric = TRUE)$coefficients
      Jaeckel_dispersion_rest <- JD_fun(Y,X_rest,koef_rest[2:length
      (koef_rest)],SCORES)
    } else {
      koef_rest <- median(walsh(Y))
      Jaeckel_dispersion_rest <- JD_fun_ubohy_model(Y,SCORES)
    }
    drop_dispersion_2 <- Jaeckel_dispersion_rest - Jaeckel_dispersion_
    full
    res_2 <- Y - X_full%*%koef_full[2:length(koef_full)]
    koef_full <- tryCatch(BFGS(Y,X_full,epsilon1,epsilon_grad1,koef_
    full_LSE
    [2:(p+1)],SCORES)[2:(p+1)],error=function(e) BFGS(Y,X_full,epsilon1,
    epsilon_grad1,koef_full[2:(p+1)],SCORES)[2:(p+1)])
    Jaeckel_dispersion_full <- JD_fun(Y,X_full,koef_full,SCORES)
    if (length(q)<p)
    {
      koef_rest <- tryCatch(BFGS(Y,X_rest,epsilon2,epsilon_grad2,
      koef_rest_LSE
      [2:length(koef_rest_LSE)],SCORES)[2:(p-length(q)+1)],error=
      function(e) BFGS(Y,X_rest,epsilon2,epsilon_grad2,koef_rest[2:
      length(koef_rest)],SCORES)[2:(p-length(q)+1)])
      Jaeckel_dispersion_rest <- JD_fun(Y,X_rest,koef_rest,SCORES)
    } else {
      koef_rest <- median(walsh(Y))
      Jaeckel_dispersion_rest <- JD_fun_ubohy_model(Y,SCORES)
    }
    drop_dispersion_3 <- Jaeckel_dispersion_rest - Jaeckel_dispersion_
    full
    res_3 <- Y - X_full%*%koef_full
    if (pomocna == 1)
    {
      tau_2ndGROUP <- c(sd(res_2),tau1(res_2,SCORES,0.8,p),gettau(res_2,
      p,SCORES,0.8),tau2(res_2,SCORES,1),tau3(res_2,SCORES,rule_of_
      thumb1(res_2)),tau3(res_2,SCORES,rule_of_thumb2(res_2)),tau3(res_2
      ,SCORES,bw.SJ(res_2)),tau3(res_2,SCORES,bandwidth_est2(res_2,
      1/(2*sqrt(pi)),1,3)),tau4(res_2,SCORES,rule_of_thumb1(res_2)),
```

```
    tau4(res_2,SCORES,rule_of_thumb2(res_2)),tau4(res_2,SCORES,bw.SJ
    (res_2)),tau4(res_2,SCORES,bandwidth_est2(res_2,1/(2*sqrt(pi)),1,
    3)))
    tau_3rdGROUP <- c(sd(res_3),tau1(res_3,SCORES,0.8,p),gettau(res_3,
    p,SCORES,0.8),tau2(res_3,SCORES,1),tau3(res_3,SCORES,rule_of_
    thumb1(res_3)),tau3(res_3,SCORES,rule_of_thumb2(res_3)),tau3(res_3
    ,SCORES,bw.SJ(res_3)),tau3(res_3,SCORES,bandwidth_est2(res_3,
    1/(2*sqrt(pi)),1,3)),tau4(res_3,SCORES,rule_of_thumb1(res_3)),
    tau4(res_3,SCORES,rule_of_thumb2(res_3)),tau4(res_3,SCORES,bw.SJ
    (res_3)),tau4(res_3,SCORES,bandwidth_est2(res_3,1/(2*sqrt(pi)),1,
    3)))
  } else if (pomocna == 3) {
    tau_2ndGROUP <- c(sd(res_2),tau2(res_2,SCORES,1))
    tau_3rdGROUP <- c(sd(res_3),tau2(res_3,SCORES,1))
  } else {
    tau_2ndGROUP <- c(sd(res_2),tau2(res_2,SCORES,1))
    tau_3rdGROUP <- c(sd(res_3),tau2(res_3,SCORES,1))
  }
  for (j in 1:length(tau_2ndGROUP))
  {
    v2[j] <- 2*drop_dispersion_2/tau_2ndGROUP[j]
    v3[j] <- 2*drop_dispersion_3/tau_3rdGROUP[j]
  }
  V2[i,] <- v2
  V3[i,] <- v3
 }
 my_list <- list("LSE" = V1, "non_RFIT" = V2, "non" = V3, "p" = p, "q"
 = q, "n" = n, "N" = N)
 return(my_list)
}

simulacie_3 <- function(N,p,l,n,SCORES,epsilon,epsilon_grad,Num_of_tau,
pomocna)
{
 set.seed(5)
 rozsah_vyberu <- seq(from=-5,to=5,by=0.01)
 Koef <- sample(rozsah_vyberu,size = p+1,replace = FALSE)
 set.seed(6)
 X <- matrix(NA,nrow = n,ncol = p)
 for (i in 1:p)
 {
   X[,i] <- runif(n,min = -2,max = 2)
 }
 X <- center_colmeans(X)
 X <- as.matrix(X)
 INTERCEPT <- rep(Koef[1],n)
 X_pomocna <- cbind(rep(1,times=n),X)
 M1 <- solve(t(X_pomocna)%*%X_pomocna)
 M2 <- solve(t(X)%*%X)
 T_nl <- sum((X[,l]^2))/n
 c_n <- T_nl*sqrt(diag(M2)[l])*sqrt(n/(n-p-1))*sqrt(n)/sqrt(12)
 set.seed(3)
 v1 <- matrix(NA,nrow = N,ncol = 2)
 v2 <- matrix(NA,nrow = N,ncol = Num_of_tau*2)
 v3 <- matrix(NA,nrow = N,ncol = Num_of_tau*2)
 if (pomocna == 1)
 {
   v4 <- matrix(NA,nrow = N,ncol = 2)
   v5 <- matrix(NA,nrow = N,ncol = 2)
 }
 for (i in 1:N)
 {
   e <- rnorm(n,mean=0,sd=1)
   Y <- INTERCEPT + X%*%Koef[2:(p+1)] + e
   koef_LSE <- M1%*%t(X_pomocna)%*%Y
   RSS <- sum((Y - X_pomocna%*%koef_LSE)^2)
   S <- sqrt(RSS/(n-p-1))
   v1[i,1] <- koef_LSE[l+1] - qt(0.975,df=n-p-1)*S*sqrt(diag(M1)[l+1])
   v1[i,2] <- koef_LSE[l+1] + qt(0.975,df=n-p-1)*S*sqrt(diag(M1)[l+1])
   koef1 <- rfit(Y ~ .,scores = SCORES,data = as.data.frame(X),
   symmetric = TRUE)$coefficients
   res_2 <- Y - X%*%koef1[2:length(koef1)]
   koef2 <- tryCatch(BFGS(Y,X,epsilon,epsilon_grad,koef_LSE[2:(p+1)],
   SCORES)[2:(p+1)],error=function(e) BFGS(Y,X,epsilon,epsilon_grad,
   koef1[2:(p+1)],SCORES)[2:(p+1)])
   res_3 <- Y - X%*%koef2
   if (pomocna == 1)
   {
     tau_2ndGROUP <- c(sd(res_2),tau1(res_2,SCORES,0.8,p),gettau(res_2,
     p,SCORES,0.8),tau2(res_2,SCORES,1),tau3(res_2,SCORES,rule_of_thumb
     1(res_2)),tau3(res_2,SCORES,rule_of_thumb2(res_2)),tau3(res_2,
     SCORES,bw.SJ(res_2)),tau3(res_2,SCORES,bandwidth_est2(res_2,1/(2*
     sqrt(pi)),1,3)),tau4(res_2,SCORES,rule_of_thumb1(res_2)),tau4(res_
     2,SCORES,rule_of_thumb2(res_2)),tau4(res_2,SCORES,bw.SJ(res_2)),
     tau4(res_2,SCORES,bandwidth_est2(res_2,1/(2*sqrt(pi)),1,3)))
     tau_3rdGROUP <- c(sd(res_3),tau1(res_3,SCORES,0.8,p),gettau(res_3,
     p,SCORES,0.8),tau2(res_3,SCORES,1),tau3(res_3,SCORES,rule_of_thumb
     1(res_3)),tau3(res_3,SCORES,rule_of_thumb2(res_3)),tau3(res_3,
     SCORES,bw.SJ(res_3)),tau3(res_3,SCORES,bandwidth_est2(res_3,1/(2*
     sqrt(pi)),1,3)),tau4(res_3,SCORES,rule_of_thumb1(res_3)),tau4(res_
     3,SCORES,rule_of_thumb2(res_3)),tau4(res_3,SCORES,bw.SJ(res_3)),
     tau4(res_3,SCORES,bandwidth_est2(res_3,1/(2*sqrt(pi)),1,3)))
   } else if (pomocna == 3) {
     tau_2ndGROUP <- c(sd(res_2),tau2(res_2,SCORES,1))
     tau_3rdGROUP <- c(sd(res_3),tau2(res_3,SCORES,1))
   } else {
     tau_2ndGROUP <- c(sd(res_2),tau2(res_2,SCORES,1))
     tau_3rdGROUP <- c(sd(res_3),tau2(res_3,SCORES,1))
   }
   for (j in 1:length(tau_2ndGROUP))
   {
     v2[i,(2*j-1)] <- koef1[l+1] - qt(0.975,df=n-p-1)*sqrt(diag(M2)
     [l])*tau_2ndGROUP[j]
     v2[i,(2*j)] <-     koef1[l+1] + qt(0.975,df=n-p-1)*sqrt(diag(M2)
     [l])*tau_2ndGROUP[j]
     v3[i,(2*j-1)] <- koef2[l] - qt(0.975,df=n-p-1)*sqrt(diag(M2)
     [l])*tau_3rdGROUP[j]
     v3[i,(2*j)] <- koef2[l] + qt(0.975,df=n-p-1)*sqrt(diag(M2)
     [l])*tau_3rdGROUP[j]
   }
   if (pomocna == 1)
   {
     t <- 0
     while (S_nl(t,res_2,X[,l])<c_n*qt(0.975,df=n-p))  t <- t-1
     delta_minus <- bisekcia1(t,t+1,c_n*qt(0.975,df=n-p),res_2,X[,l])
     t <- 0
     while (S_nl(t,res_2,X[,l])>(-1*c_n*qt(0.975,df=n-p))) t <- t+1
     delta_plus <- bisekcia2(t-1,t,c_n*qt(0.975,df=n-p),res_2,X[,l])
     v4[i,1] <- koef1[l+1] + delta_minus
     v4[i,2] <- koef1[l+1] + delta_plus
     t <- 0
     while (S_nl(t,res_3,X[,l])<c_n*qt(0.975,df=n-p)) t <- t-1
     delta_minus <- bisekcia1(t,t+1,c_n*qt(0.975,df=n-p),res_3,X[,l])
     t <- 0
     while (S_nl(t,res_3,X[,l])>(-1*c_n*qt(0.975,df=n-p))) t <- t+1
     delta_plus <- bisekcia2(t-1,t,c_n*qt(0.975,df=n-p),res_3,X[,l])
     v5[i,1] <- koef2[l] + delta_minus
     v5[i,2] <- koef2[l] + delta_plus
```

```
    }
    print(i)
  }
  if (pomocna == 1) {
    return(cbind(v1,v2,v3,v4,v5))
  } else {
    return(cbind(v1,v2,v3))
  }
}

simulacie_4 <- function(N,B,n,p,l,SCORES)
{
  l <<- l
  SCORES <<- SCORES
  set.seed(5)
  rozsah_vyberu <- seq(from=-5,to=5,by=0.01)
  Koef <- sample(rozsah_vyberu,size = p+1,replace = FALSE)
  set.seed(6)
  X <- matrix(NA,nrow = n,ncol = p)
  for (i in 1:p)
  {
    X[,i] <- runif(n,min = -2,max = 2)
    X[,i] <- X[,i] - mean(X[,i])
  }
  X <- as.matrix(X)
  INTERCEPT <- rep(Koef[1],n)
  X_pomocna <- cbind(rep(1,times=n),X)
  M1 <- solve(t(X_pomocna)%*%X_pomocna)%*%t(X_pomocna)
  vektor <- seq(from=1,to=n,by=1)
  v1 <- matrix(NA,nrow = N,ncol = 8)
  v2 <- matrix(NA,nrow = N,ncol = 8)
  set.seed(3)
  for (i in 1:N)
  {
    e <- rnorm(n,mean=0,sd=1)
    Y <- INTERCEPT + X%*%Koef[2:(p+1)] + e
    koef_LSE <- M1%*%Y
    koef1 <- rfit(Y ~ .,scores = SCORES,data = as.data.frame(X),
    symmetric = TRUE)$coefficients
    sada <- cbind(Y,X_pomocna)
    sada_non <- cbind(Y,X)
    koef_LSE_boot <- boot(data = sada, statistic = f_LSE,R = B)
    koef1_boot <- boot(data = sada_non, statistic = f_non,R = B)
    v1[i,1] <- koef_LSE[l+1] - qnorm(0.975)*sd(koef_LSE_boot$t)
    v1[i,2] <- koef_LSE[l+1] + qnorm(0.975)*sd(koef_LSE_boot$t)
    v1[i,3] <- 2*koef_LSE[l+1] - mean(koef_LSE_boot$t) - qnorm(0.975)*
    sd(koef_LSE_boot$t)
    v1[i,4] <- 2*koef_LSE[l+1] - mean(koef_LSE_boot$t) + qnorm(0.975)*
    sd(koef_LSE_boot$t)
    v1[i,5] <- sort(koef_LSE_boot$t)[(round(0.025*B)+1)]
    v1[i,6] <- sort(koef_LSE_boot$t)[(B-round(0.025*B))]
    IS_bca <- boot.ci(koef_LSE_boot,type = "bca")
    v1[i,7] <- IS_bca$bca[4]
    v1[i,8] <- IS_bca$bca[5]
    v2[i,1] <- koef1[l+1] - qnorm(0.975)*sd(koef1_boot$t)
    v2[i,2] <- koef1[l+1] + qnorm(0.975)*sd(koef1_boot$t)
    v2[i,3] <- 2*koef1[l+1] - mean(koef1_boot$t) - qnorm(0.975)*
    sd(koef1_boot$t)
    v2[i,4] <- 2*koef1[l+1] - mean(koef1_boot$t) + qnorm(0.975)*
    sd(koef1_boot$t)
    v2[i,5] <- sort(koef1_boot$t)[(round(0.025*B)+1)]
    v2[i,6] <- sort(koef1_boot$t)[(B-round(0.025*B))]
    IS_bca <- boot.ci(koef1_boot,type = "bca")
    v2[i,7] <- IS_bca$bca[4]
    v2[i,8] <- IS_bca$bca[5]
  }
  return(cbind(v1,v2))
}

library(Rfit)

wilcox_function <- function(u) sqrt(12)*(u-0.5)
wilcox_function_d <- function(u) rep(sqrt(12),length(u))
wilcox_score_function <- new("scores",phi = wilcox_function, Dphi =
 wilcox_function_d)

kvantil_function <- function(u) qnorm(u)
kvantil_function_d <- function(u) 1/dnorm(qnorm(u))
kvantil_score_function <- new("scores",phi = kvantil_function, Dphi =
 kvantil_function_d)

bent_function <- function(u) ifelse(u<0.25,-sqrt(3/2),ifelse(u>0.75,
sqrt(3/2),sqrt(3/2)*(4*u-2)))
bent_function_d <- function(u) ifelse(u<0.25,0,ifelse(u>0.75,0,
sqrt(3/2)*4))
bent_score_function <- new("scores",phi = bent_function, Dphi = bent_
function_d)

pomocna_funkcia <- function()
{
  p_value_LSE <- 1-pf(sim2$LSE,df1 = length(sim2$q), df2 = sim2$n -
  sim2$p -1)
  pocet_zamietnuti <- length(p_value_LSE[p_value_LSE<=0.05])
  v1 <- NA
  v1 <- 100*pocet_zamietnuti/sim2$N
  p_value_non_chi <- function(V)
  {
    p_value <- 1-pchisq(V,df = length(sim2$q))
    pocet_zamietnuti <- length(p_value[p_value<=0.05])
    return(100*pocet_zamietnuti/sim2$N)
  }
  v2 <- NA
  for (i in 1:dim(sim2$non_RFIT)[2])
  {
    v2[i] <- p_value_non_chi(sim2$non_RFIT[,i])
  }
  v3 <- NA
  for (i in 1:dim(sim2$non)[2])
  {
    v3[i] <- p_value_non_chi(sim2$non[,i])
  }
  p_value_non_F <- function(V)
  {
    p_value <- 1-pf(V/length(sim2$q),df1 = length(sim2$q),df2 = sim2$n -
    sim2$p -1)
    pocet_zamietnuti <- length(p_value[p_value<=0.05])
    return(100*pocet_zamietnuti/sim2$N)
  }
  v4 <- NA
  for (i in 1:dim(sim2$non_RFIT)[2])
  {
    v4[i] <- p_value_non_F(sim2$non_RFIT[,i])
  }
  v5 <- NA
  for (i in 1:dim(sim2$non)[2])
  {
    v5[i] <- p_value_non_F(sim2$non[,i])
  }
```

```
  v1 <- rep(v1,times = length(v5))
  return(cbind(v1,v2,v4,v3,v5))
}
JD_fun <- function(Y,X,beta,score)
{
  score_fun <-function(u)
  {
    a<-score@phi(u)
    ac<-a-mean(a)
    sigma<-sum(ac*ac)/(length(ac)+1)
    return(ac/sqrt(sigma))
  }
  at_fun <- function(t) score_fun(t/(length(Y)+1))
  pseudo_norm <- function(u) at_fun(rank(u,ties.method = "average"))
  %*%u
  return(pseudo_norm(Y-X%*%beta))
}
JD_fun_ubohy_model <- function(Y,score)
{
  score_fun <-function(u)
  {
    a<-score@phi(u)
    ac<-a-mean(a)
    sigma<-sum(ac*ac)/(length(ac)+1)
    return(ac/sqrt(sigma))
  }
  at_fun <- function(t) score_fun(t/(length(Y)+1))
  pseudo_norm <- function(u) at_fun(rank(u,ties.method = "average"))
  %*%u
  return(pseudo_norm(Y))
}
grad_JD_fun<-function(Y,X,beta,score)
{
  score_fun <-function(u)
  {
    a<-score@phi(u)
    ac<-a-mean(a)
    sigma<-sum(ac*ac)/(length(ac)+1)
    return(ac/sqrt(sigma))
  }
  at_fun <- function(t) score_fun(t/(length(Y)+1))
  e<-Y-X%*%beta
  return(t(-X)%*%at_fun(rank(e,ties.method="average")))
}
korekcna_matica <- function(H,p,y)
{
  a<-as.numeric((1+(t(y)%*%H%*%y)/(t(p)%*%y))/(t(p)%*%y))
  b<-as.numeric(t(p)%*%y)
  return(a*(p%*%t(p)) - (H%*%y%*%t(p) + p%*%t(y)%*%H)/b)
}
BFGS<-function(Y,X,eps,eps_grad,beta0,score)
{
  fun_hod <- function(b) JD_fun(Y,X,b,score)
  g_hod <- function(b) grad_JD_fun(Y,X,b,score)
  g <- g_hod(beta0)
  H <- diag(ncol(X))
  beta_stare<-beta0
  pomocne_beta<-beta_stare
  beta_nove <- beta_stare*10
  k<-0
  while (((fun_hod(pomocne_beta)-fun_hod(beta_nove))^2)>eps)
  {
    if (norm(g,type = "2")<eps_grad) break
    smer<- -H%*%g
    fun_phi <- function(lambda) fun_hod(beta_stare + lambda*smer)
    opt_krok <- optimize(fun_phi,interval = c(-20,20))$minimum
    pomocne_beta<- beta_stare
    beta_nove<-beta_stare + opt_krok*smer
    g_nove<-g_hod(beta_nove)
    y<- g_nove-g
    p<- beta_nove-beta_stare
    dH <- korekcna_matica(H,p,y)
    H <- H + dH
    g<-g_nove
    beta_stare<-beta_nove
    k<-k+1
  }
  return(c(k,beta_nove))
}
tau1 <- function(residua,score,delta,p)
{
  score_fun <-function(u)
  {
    a<-score@phi(u)
    ac<-a-mean(a)
    sigma<-sum(ac*ac)/(length(ac)+1)
    return(ac/sqrt(sigma))
  }
  stand_score_fun <- function(u) (score_fun(u)-score_fun(u)[1])/(score_
  fun(u)[length(u)]-score_fun(u)[1])
  n <- length(residua)
  U <- score_fun(seq(from=0,to=1,by=1/n))
  H_n <- function(y,residua)
  {
    n<-length(residua)
    usp_residua <- sort(residua)
    p <- stand_score_fun(seq(from=0,to=1,by=1/n))
    p1 <- p[2:(n+1)]-p[1:n]
    matica1 <- t(matrix(usp_residua,ncol = n,nrow = n))
    matica1 <- abs(matica1-usp_residua)
    pomocna <- matica1
    matica1[pomocna<=y] <- 1
    matica1[pomocna>y] <- 0
    return(sum(p1%*%matica1)/n)
  }
  horna_hranica <- function(default_hranica,re,qua)
  {
    h <- default_hranica
    while (H_n(h,re) < (qua+0.1)) h<-h+0.5
    return(h)
  }
  kvantil_of_H_n <- function(residua,qua)
  {
    pomocna_funkcia1<- function(y) H_n(y,residua) - qua
    vysledok<- uniroot(pomocna_funkcia1,c(0,horna_hranica(0.5,residua,
    qua)),tol= 0.0001)
    return(vysledok)
  }
  gama_est <- function(residua,qua,t_n)
  {
    n<-sqrt(length(residua))
    gama<-(U[length(U)]-U[1])*(H_n(t_n/n,residua)/(2*t_n/n))
    return(gama)
  }
  kvan1<-kvantil_of_H_n(residua,delta)$root
  return((sqrt((n)/(n-p-1)))/gama_est(residua,delta,kvan1))
}
tau2 <- function(residua,score,A)
```

```
{
  score_fun_plus <- function(u)
  {
    A <- score@phi((u+1)/2)
    a <- score@phi(u)
    ac<-a-mean(a)
    sigma<-sum(ac*ac)/(length(ac)+1)
    return((A-mean(a))/sigma)
  }
  a_plus <- function(k,N) score_fun_plus(k/(N+1))
  N <- length(residua)
  Z_N <- function(A,e,alpha)
  {
    N<-length(e)
    matica <- matrix(e, ncol = length(alpha), nrow = N)
    matica <- matica - t(matrix(alpha,ncol = N,nrow = length(alpha)))
    v<-diag(t(sign(matica))%*%a_plus(apply(abs(matica),MARGIN = 2,FUN =
    rank),N)
    v<-v/(A*sqrt(N))
    return(v)
  }
  pod_hodnotou <- function(index,WA,hod)
  {
    while (Z_N(A,residua,WA[index]) < hod)
    {
      index_stary <- index
      index <- index - round(length(WA)*0.005)
    }
    return(c(index,index_stary))
  }
  nad_hodnotou <- function(index,WA,hod)
  {
    while (Z_N(A,residua,WA[index]) > -hod)
    {
      index_stary <- index
      index <- index + round(length(WA)*0.005)
    }
    return(c(index_stary,index))
  }
  WA <- sort(walsh(residua))
  ind <- round(length(WA)/2)
  if (Z_N(A,residua,WA[ind]) < qnorm(0.975))
  {
    v <- pod_hodnotou(ind,WA,qnorm(0.975))
  } else
  {
    v <- nad_hodnotou(ind,WA,-qnorm(0.975))
  }
  if (Z_N(A,residua,WA[ind]) > -qnorm(0.975))
  {
    w <- nad_hodnotou(ind,WA,qnorm(0.975))
  } else
  {
    w <- pod_hodnotou(ind,WA,-qnorm(0.975))
  }
  p1 <- Z_N(A,residua,WA[v[1]:v[2]])
  l <- p1[p1<qnorm(0.975)]
  l_ind <- which(p1==max(l))
  l_ind <- l_ind[1] + v[1] - 1
  p2 <- Z_N(A,residua,WA[w[1]:w[2]])
  u <- p2[p2>-qnorm(0.975)]
  u_ind <- which(p2==min(u))
  u_ind <- u_ind[length(u_ind)] + w[1] - 1
  return((WA[u_ind]-WA[l_ind])*sqrt(N)/(2*qnorm(0.975)))
}
rule_of_thumb1 <- function(residua)
{
  n <- length(residua)
  0.9*min(sd(residua),(quantile(residua)[4]-quantile(residua)[2])/1.34)*
  (n^(-1/5))
}
rule_of_thumb2 <- function(residua)
{
  n <- length(residua)
  1.06*min(sd(residua),(quantile(residua)[4]-quantile(residua)[2])/1.34)
  *(n^(-1/5))
}
bandwidth_est1<-function(residua,R_k,sigma_k_nadruhu)
{
  n<-length(residua)
  lambda<-as.numeric(quantile(residua)[4]-quantile(residua)[2])
  a<-0.920*lambda*(n^(-1/7))
  b<-0.912*lambda*(n^(-1/9))
  norm_density_4 <- function(x) (1/sqrt(2*pi))*exp(-0.5*x^2)*(x^4 -
  6*x^2 + 3)
  norm_density_6 <- function(x) (1/sqrt(2*pi))*exp(-0.5*x^2)*(x^6 -
  15*x^4 + 45*x^2 - 15)
  S_D <- function(alp)
  {
    matica <- t(matrix(residua,ncol = n,nrow = n))
    matica <- (matica - as.vector(residua))/alp
    p <- sum(norm_density_4(matica))
    return(p/((alp^5)*n*(n-1)))
  }
  T_D <- function(b)
  {
    matica <- t(matrix(residua,ncol = n,nrow = n))
    matica <- (matica - as.vector(residua))/b
    r <- sum(norm_density_6(matica))
    return(-r/((b^7)*n*(n-1)))
  }
  nthroot <- function(a,b) ifelse( b %% 2 == 1 | a >= 0,sign(a)*abs(a)^
  (1/b),NaN)
  alpha <- function(h)
  {
    m<-(S_D(a))/(T_D(b))
    v<-nthroot(m,7)
    v<-(h^(5/7))*1.357*v
    return(v)
  }
  pomocna_funkcia <- function(h) (nthroot(((R_k)/((sigma_k_nadruhu^2)*
  S_D(alpha(h)))),5))*(n^(-1/5)) - h
  vysledok<- uniroot(pomocna_funkcia,c(0.00001,10),tol = 0.001)
  return(vysledok$root)
}
bandwidth_est2 <- function(res,R_k,mu_2,mu_4)
{
  n<-length(res)
  lambda<-as.numeric(quantile(res)[4]-quantile(res)[2])
  a<-4.29*lambda*(n^(-1/11))
  b<-0.91*lambda*(n^(-1/9))
  L_4 <- function(condition,x) ifelse(condition,(135135/16384)*(-184756
  *x^10 + 504900*x^8 - 491400*x^6 + 200200*x^4 - 29700*x^2 + 756),0)
  norm_density_6 <- function(x) (1/sqrt(2*pi))*exp(-0.5*x^2)*(x^6 -
  15*x^4 +45*x^2 - 15)
  nthroot <- function(a,b,c) ifelse( b %% 2 == 1 | a >= 0,sign(a)*abs(a)
^(c/b),NaN)
  matica <- t(matrix(res,ncol = n,nrow = n))
```

```
  matica1 <- (matica - as.vector(res))/a
  i2 <- sum(L_4(abs(matica1[abs(matica1)<=1]<=1),matica1[abs(matica1)
  <=1]))
  matica2 <- (matica - as.vector(res))/b
  i3 <- sum(norm_density_6(matica2))
  i2<-i2/(n*(n-1)*(a^5))
  i3<- -i3/(n*(n-1)*(b^7))
  J2<-(mu_4*i3)/(20*mu_2*i2)
  J1<-R_k/((mu_2^2)*i2)
  v<-nthroot(J1/n,5,1) + J2*nthroot(J1/n,5,3)
  return(v)
}
tau3 <- function(residua,score,h)
{
  N <- length(residua)
  funkcional <- function(res,h)
  {
    n <- length(res)
    k <- 1/(sqrt(2)*h*(n^2))
    matica <- t(matrix(res,ncol = n,nrow = n))
    matica <- (matica - as.vector(res))/(sqrt(2)*h)
    p <- sum(dnorm(matica))
    return(k*p)
  }
  return(1/(getScoresDeriv(score,1:N/(N+1))[1]*funkcional(residua,h)))
}
tau4 <- function(residua,score,h)
{
  w<- function(x) dnorm(x)
  w_n<-function(DIF,h) w((DIF)/h)/h
  der_score<- function(x)
  {
    aP <- score@Dphi(x)
    a <- score@phi(x)
    ac<-a-mean(a)
    sigma<-sum(ac*ac)/(length(ac)+1)
    return(aP/sqrt(sigma))
  }
  gama <- function(residua,h)
  {
    n <- length(residua)
    p <- seq(from=1/n,to=1,by=1/n)
    matica <- t(matrix(sort(residua),ncol = n,nrow = n))
    matica <- matica - as.vector(sort(residua))
    return(sum(der_score(p)*t(w_n(matica,h)))/(n^2))
  }
  return(1/gama(residua,h))
}
S_nl <- function(t,e,x)
{
  R <- rank(e - t*x)
  n <- length(e)
  v <- (t(x)%*%R)/(n^(3/2))
  return(v)
}
bisekcia1 <- function(a,b,podmienka,e,x) #a<b
{
  while (abs(b-a)>(10^(-9)))
  {
    novy <- (a+b)/2
    if (S_nl(novy,e,x)>=podmienka) {
      a <- novy
    } else {
      b <- novy
    }
  }
  return(a)
}
bisekcia2 <- function(a,b,podmienka,e,x) #a<b
{
  while (abs(b-a)>(10^(-9)))
  {
    novy <- (a+b)/2
    if (S_nl(novy,e,x)<=(-1*podmienka)) {
      b <- novy
    } else {
      a <- novy
    }
  }
  return(b)
}
spolahlivost <- function(vysledok,p,l,s)
{
  set.seed(s)
  rozsah_vyberu <- seq(from=-5,to=5,by=0.01)
  Koef <- sample(rozsah_vyberu,size = p+1,replace = FALSE)
  bodovy_odhad <- Koef[l+1]
  P <- dim(vysledok)[2]/2
  V <- matrix(NA,nrow = 3,ncol = P)
  for (i in 1:P)
  {
    lava <- vysledok[,(2*i-1)]<=bodovy_odhad
    prava <- vysledok[,(2*i)]>=bodovy_odhad
    pom <- lava+prava
    V[1,i] <- 100*length(pom[pom==2])/dim(vysledok)[1]
    V[2,i] <- mean(vysledok[,(2*i)]-vysledok[,(2*i-1)])
    V[3,i] <- sd(vysledok[,(2*i)]-vysledok[,(2*i-1)])
  }
  return(V)
}
center_colmeans <- function(x)
{
  xcenter <- colMeans(x)
  return(x - rep(xcenter, rep.int(nrow(x), ncol(x))))
}
f_LSE <- function(udaje,indexy)
{
  d <- udaje[indexy,]
  y <- d[,1]
  x <- d[,2:dim(d)[2]]
  v <- solve(t(x)%*%x)%*%t(x)%*%y
  return(v[l+1])
}
f_non <- function(udaje,indexy)
{
  d <- udaje[indexy,]
  y <- d[,1]
  x <- d[,2:dim(d)[2]]
  v <- rfit(y ~ .,scores = SCORES,data = as.data.frame(x),symmetric =
  TRUE)$  coefficients
  return(v[l+1])
}
```