Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

THEILOVA REGRESIA

Róbert Tóth

Bratislava 2013

Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

THEILOVA REGRESIA

(Bakalárska práca)

Róbert Tóth

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	1114 aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Školiteľ:	Mgr. Ján Somorčík, PhD.

Bratislava 2013

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Róbert Tóth
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov: Theilova regresia

Cieľ: Pomocou počítačových simulácií porovnať v priamkovej regresii kvalitu Theilovej metódy s inými prístupmi. Navrhnúť podobné metódy. Pozrieť sa na implementáciu metód v štatistickom softvéri R.

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

Dátum zadania: 10.10.2012

Dátum schválenia: 03.11.2012

doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie

Ďakujem môjmu školiteľovi Jánovi Somorčíkovi za návrh skvelej témy bakalárskej práce, za pomoc pri oboznamovaní sa s problematikou a najmä za dôležité náhľady, ktoré mi pomohli pri jej tvorení.

Abstrakt

V tejto práci skúmame špeciálny druh priamkovej lineárnej regresie - Theilovej regresie. Skúmame odhad pre sklon, interval spoľahlivosti a testy. Porovnávame jej vlastnosti s metódou najmenších štvorcov pri rôznych rozdeleniach chýb merania. Druhá kapitola sa snaží dokázať, že balík mblm v softvéri R používa zlú metódu na počítanie intervalu spoľahlivosti pre sklon. Uvádzame dôkaz pre prípad $n = 5$, v ktorom je problém zredukovaný na výpočet dobre definovaného konečného integrálu. V poslednej kapitole sa snažíme zostrojiť nový odhad pre sklon, ktorý kombinuje dobré vlastnosti Theilovej metódy a metódy najmenších štvorcov.

Kľúčové slová: Theilova regresia, mblm.R, Wilcoxonova metóda

Abstract

In this paper we examine specific kind of linear regression - Theil's regression. We investigate an estimate of slope, confidence interval and tests. We compare its properties with the least squares method under different distributions of measurement error. Second chapter aims to prove that the mblm package in software R uses wrong method for computation of a confidence interval for slope. We provide a proof for case $n = 5$, in which the problem is reduced to computation of a well defined finite integral. In last chapter we attempt to construct new estimate for slope that combines good properties from both Theil's method and least squares method.

Key words: Theil's regression, mblm.R, Wilcoxon's method

Obsah

Úvod

Medzi často používané metódy štatistiky, ktoré sú bežne využívané vo všetkých vedných odboroch, patrí lineárna regresia. Najbežnejším druhom tejto regresie býva takzvaná metóda najmenších štvorcov, ktorá sa ujala najmä pre svoju jednoduchosť a ľahkú pochopiteľnosť. Pri normálnych rozdeleniach odchýlok je táto metóda dokázateľne aj najlepšou, ale ukazuje sa, že je značne nerobustná. Preto najmä pri rozdeleniach chýb s ťažkými chvostami, akými je napríklad Cauchyho alebo Laplaceovo rozdelenie, dochádza k značným nepresnostiam v odhade sklonu a priesečníku regresnej priamky. Takisto pre dáta s outliermi nie je táto metóda najvhodnejšia. Výhodou Theilovej regresie je robustnosť a takisto ľahká pochopiteľnosť, preto má zmysel sa jej venovať a porovnávať ju s inými metódami. V tejto bakalárskej práci uvedieme princíp Theilovej metódy a pokúsime sa ju vylepšiť. S odhadmi sklonu regresnej priamky úzko súvisia aj testy o skutočnej hodnote sklonu. Jedným z nich je test priamo využívajúci odhad metódy najmenších štvorcov. Theil okrem svojho odhadu uvádza aj test, ktorý je však neparametrický a dokonca ani nevyužíva príslušný odhad. Preto má takisto zmysel tieto testy porovnávať, čo tiež popisuje táto práca. A nakoniec, najčastejšie používaný matematický softvér R obsahuje jediný balík, ktorý poskytuje použitie týchto metód. Ukazuje sa však, že zdrojový kód je postavený na chybných základoch a intervaly spoľahlivosti nedosahujú požadovanú spoľahlivosť. V tejto práci sa venujeme simulačnému aj teoretickému zdôvodneniu tejto chybovosti.

Kapitola 1

Základný model

Náš model bude zostavený ako v [1]. V každej z n nezávisle fixovaných hodnôt $x_1, x_2, \cdots, x_n$ náhodnej premennej (prediktoru) x pozorujeme hodnotu náhodnej premennej Y (odpovede). Takto dostávame množinu pozorovaní $Y_1, Y_2, ..., Y_n$, kde Y_i je hodnota odpovede, keď $x = x_j$. Predpokladáme, že bez ujmy na všeobecnosti platí $x_1 < x_2 < \cdots < x_n$.

Predpoklady

Náš priamkový model má tvar

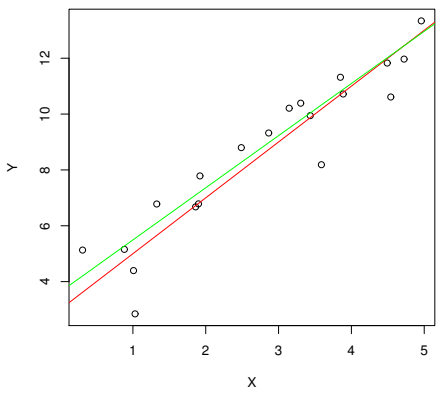
$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n,$$

kde x_i a Y_i sú známe pozorovania, α (priesečník) a β (sklon) neznáme parametre, ktoré sa pokúšame odhadnúť a ε_i je pre každé i náhodná premenná z toho istého rozdelenia so spojitou hustotou. Navyše sú tieto chyby ε_i navzájom nezávislé.

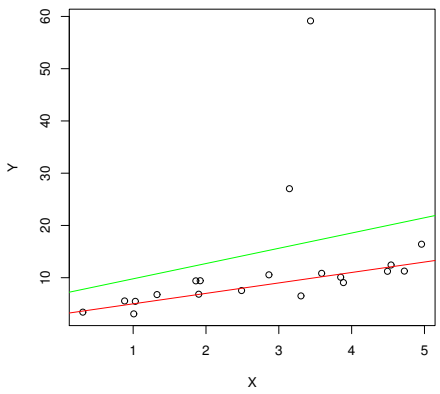
Našou úlohou je čo možno najlepšie na základe získaných dát odhadnúť veľkosť parametrov α a β. Obtiažnosť tejto úlohy závisí od hustoty rozdelenia chyby ε, či už známej, alebo nie. Štandardne sa v lineárnej regresii predpokladá, že chyby pochádzajú z rozdelenia $N(0, \sigma^2)$. V tomto prípade sa za odhad parametra β berie odhad získaný metódou najmenších štvorcov. Tento odhad je jeden z najlepších, aký môžeme v danom prípade použiť, o čom hovorí aj nasledujúca veta.

Veta 1.1 (Gauss-Markov)**.** *Nech platia všetky predpoklady zo začiatku kapitoly. Nech X je matica plánu, t.j. matica s prvým stĺpcom rovným jednotkovému vektoru dĺžky n a druhým stĺpcom rovným vektoru $(x_1, x_2, \cdots, x_n)^T$. Potom odhad $\hat{\beta} = (X^T X)^{-1} X^T Y$, ktorý získame minimalizovaním sumy reziduí $\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{2} \hat{\beta}_{ij} X_{ij})^2$, je BLUE pre $\alpha = \beta_1$ a $\beta = \beta_2$.*

Táto veta je štandardným výsledkom v pravdepodobnosti a štatistike a dá sa nájsť v mnohých učebných materiáloch ako aj v [3]. Problémom je, že v skutočnosti chyby z normálneho rozdelenia niekedy nebývajú a v drvivej väčšine prípadov ani netušíme, z akého rozdelenia pochádzajú. Okrem toho sa často stáva, že dáta obsahujú outlierov, na ktorých je metóda najmenších štvorcov obzvlášť citlivá, pretože sa radí medzi nerobustné metódy (dá sa ukázať, že tento odhad sklonu β pomocou MNŠ je váženým priemerom sklonov všetkých priamok, ktoré vieme z našich n dát vyrobiť). Na nasledujúcich obrázkoch vidno, ako sa pokazí odhad regresnej priamky, ak nemáme dáta z normálneho rozdelenia.



Obr. 1.1: Chyby z N(0,1)



Obr. 1.2: Chyby z Cauchy(0,1)

Červenou je vyznačená skutočná priamka a zelenou je odhadnutá MNŠ.
Ako si môžeme všimnúť, outliery nám značne pokazili najmä odhad sklonu. Má preto význam zaoberať sa robustnými metódami, akou je Theil-Senova metóda.

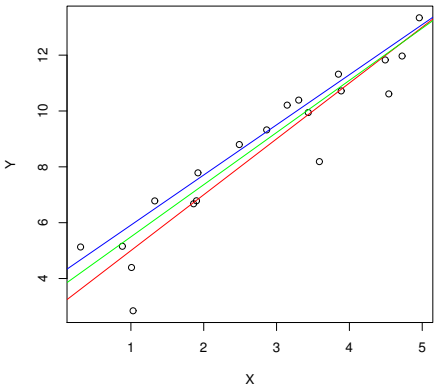
1.1 Theil-Senova metóda odhadu sklonu regresnej priamky

Prirodzený nápad, ktorý dostaneme po zistení, že MNŠ je váženým priemerom, je nebrať priemer sklonov, ale ich medián

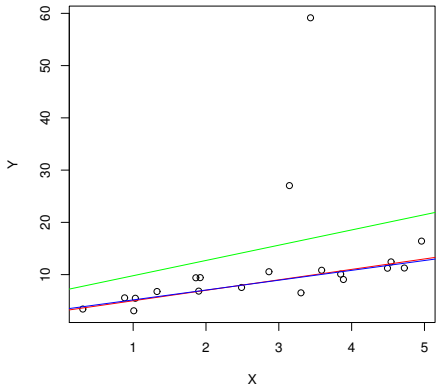
$$\hat{\beta} = med\{t_{i,j}|t_{i,j} = \frac{Y_j - Y_i}{X_j - X_i}, 1 \leq i < j \leq n\}.$$

Analogicky definujeme aj odhad pre priesečník

$$\hat{\alpha} = med\{t_{i,j}|t_{i,j} = \frac{Y_j X_j - Y_i X_j}{X_j - X_i}, 1 \leq i < j \leq n\}.$$



Obr. 1.3: Chyby z N(0,1)



Obr. 1.4: Chyby z Cauchy(0,1)

Na obrázkoch vidíme, ako by naša priamka dopadla, ak by sme použili Theilov odhad pre sklon. Červenou je vyznačená skutočná priamka, zelenou je odhadnutá MNŠ a modrou Theilov odhad.

Je vidno, že táto metóda si zachováva celkom slušnú presnosť pre chyby z normálneho rozdelenia, ale pri rozdelení s ťažkými chvostami bude na tom zrejme omnoho lepšie ako MNŠ. Nasledujúca tabuľka simulácií ukazuje priemernú hodnotu, o ktorú sa pomýlili obidve metódy pri 10 000 simuláciách

dát pre fixované $n = 20$, hodnoty $x_i = i$ a rozdielne rozdelenia chýb. Skutočný sklon bol nastavený na hodnotu $\beta = 2$, skutočný priesečník na hodnotu $\alpha = 3$. Dáta Y_i boli nasimulované podľa nášho modelu zo začiatku kapitoly a následne sa spravili odhady pomocou MNŠ aj Theila.

Tabuľka 1.1: Priemerné štvorce odchýlok sklonov β

Rozdelenie chyby	Theilov odhad	Odhad MNŠ	Pomer chýb v %
N(0,1)	0.001617324	0.001483357	91.72
Cauchy $x_0 = 0$, $\gamma = 1$	0.007178022	207.9608	2897188.11
Laplaceovo $\lambda = 1$	0.002309322	0.002956753	128.04
N(0,1), jeden outlier	0.002026732	0.01693216	835.44

Tabuľka 1.2: Priemerné štvorce odchýlok odhadov priesečníkov α

Rozdelenie chyby	Theilov odhad	Odhad MNŠ	Pomer chýb v %
N(0,1)	0.2776968	0.2269701	81.73
Cauchy $x_0 = 0$, $\gamma = 1$	0.8687903	11542.08	1328523.12
Laplaceovo $\lambda = 1$	0.3234315	0.3549759	109.75
N(0,1), jeden outlier	0.327196	0.6114294	186.87

Stĺpec "Pomer chýb v %" udáva, aká veľká je chyba MNŠ oproti Theilovej metóde. Môžeme vidieť, že Theilov odhad je ďaleko lepším odhadom, pokiaľ chyby nepochádzajú naozaj z normálneho rozdelenia. Ani v tomto prípade sa ale nemusíme báť tento odhad použiť, pretože odhad MNŠ nedosahuje až o toľko lepšie výsledky. Okrem odhadovania však častokrát potrebujeme testovať hypotézu o sklone regresnej priamky alebo vypočítať interval spoľahlivosti. Práca Theila spočíva najmä v tom, že si uvedomil súvislosť sklonov a priesečníkov s Kendallovým τ–rozdelením. V ďalšej časti sa venujeme práve týmto dvom úlohám a tiež porovnaniu úspešnosti použitia týchto metód s metódami vytvorenými na základe odhadu MNŠ.

1.2 Theil-Senovo testovanie sklonu regresnej priamky

Testujme najprv hypotézu v tvare

$$H_0 : \beta = \beta_0.$$

Nech $D_k = Y_k - \beta_0 X_k = (\beta - \beta_0)X_k + \alpha + \varepsilon_k$. Za predpokladu, že H_0 platí, môžeme tvrdiť, že hodnoty D_k nezávisia od hodnôt X_k. To je zrejme to isté, ako tvrdiť, že ich korelácia je nulová. Vhodným korelačným koeficientom je Kendallov korelačný koeficient, ktorého odhad nájdeme podľa [4]:

$$T = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(D_j - D_i) sgn(X_j - X_i) =$$

$$= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(D_j - D_i).$$

Posledná rovnosť platí vďaka usporiadaniu našich X_i. Rozdelenie tejto štatistiky ale poznáme. Je ním práve Kendallovo τ–rozdelenie. Táto skutočnosť nám umožňuje určiť extrémnu kladnosť alebo extrémnu zápornosť T a kontrolovať tak chybu prvého druhu. Na základe toho je zostavený nasledujúci test pre našu hypotézu na hladine významnosti α:

Zamietni hypotézu $H_0 : \beta = \beta_0$, ak $|T| \geq k_{\alpha/2}$.

Rovnakými úvahami sa vieme dostať aj k $(1-\alpha)\%$ intervalu spoľahlivosti v tvare

$$(S^{(L)}, S^{(R)}),$$

kde $S^{(1)} \leq S^{(2)} \leq \cdots \leq S^{(N)}$ sú zoradené sklony, $N = n(n-1)/2, L = (N - k_{\alpha/2} + 2)/2$ a $R = L + k_{\alpha/2} - 1$. Znakom $k_{\alpha/2}$ v oboch prípadoch označujeme $(100 - \alpha/2)$–percentnú kritickú hodnotu Kendallovho rozdelenia. Teraz prirodzene prichádza otázka, aká je sila tohto testu v porovnaní s tradičným testom založeným na MNŠ a aké široké sú jednotlivé intervaly spoľahlivosti. V nasledujúcich tabuľkách môžeme vidieť výsledky týchto dvoch metód pre rôzne rozdelenia chyby. Testovaná hypotéza bola $\beta = 2 + x$, kde x je číslo hneď

za názvom stĺpca "Sila testu". Skutočná hodnota sklonu bola opäť $\beta = 2$ a počet simulácií bol tiež 10000. Vidíme, že ako bolo očakávané, sila testu s väčšou požadovanou presnosťou klesá pre oba testy. Taktiež sa čakalo, že pre normálne rozdelenie chyby bude MNŠ lepšia, čo sa aj potvrdilo, ale tento rozdiel nie je až taký veľký. Naopak, pre rozdelenie s ťažkými chvostami MNŠ dopadla oveľa horšie.

Tabuľka 1.3: Porovnanie síl testov

Metóda	Rozdelenie	Sila testu 0.20	Sila testu 0.10	Sila testu 0.05
Theil	Laplace	94.08	49.44	16.67
MNŠ	Laplace	92.25	44.26	14.74
Theil	N(0,1)	99.5	62.77	20.15
MNŠ	N(0,1)	99.88	67.93	22.52
Theil	Cauchy	59.84	23.07	9.32
MNŠ	Cauchy	23.54	8.85	4.5

Šírka intervalov spoľahlivosti len potvrdzuje našu domienku, že čím ťažšie chvosty má rozdelenie chýb, tým lepšie je používať Theilovu metódu.

Tabuľka 1.4: Porovnanie priemernej šírky intervalov spoľahlivosti

Rozdelenie chyby	Theil	Spoľahlivosť	MNŠ	Spoľahlivosť
Laplace	0.2051879	95.14%	0.2234207	95.19%
N(0,1)	0.1684538	95.07%	0.1612581	95.25%
Cauchy	0.3878784	95.08%	4.996909	96.75%

Kapitola 2

Balík R s Theilovou regresiou

Ešte pred niekoľkými mesiacmi bol zrejme jediným dostupným balíkom softvéru *R*, ktorý mal v sebe implementované metódy lineárnej regresie vyvinuté Theilom, balík s názvom **mblm** [2]. Tento balík ale obsahoval chybu, na ktorú sme autora balíku upozornili, na čo bol balík zanedlho stiahnutý z CRANu (oficiálne a najväčšie úložisko balíkov softvéru R na internete). Mnoho užívateľov si už medzitým tento balík stiahlo a metóda použitá na výpočet konfidenčného intervalu pre sklon regresnej priamky nedosahuje požadovanú spoľahlivosť. V ďalšej časti ukážeme, prečo je to tak.

2.1 Metóda použitá v balíku mlbm

Zo skriptu tohto balíka uvádzame časť **confint.mblm.R**.

```
"confint.mblm" ← function (object, parm, level = 0.95, ...)
{res = c(0,0,0,0); dim(res) = c(2,2); rownames(res) =
names(object$coefficients); colnames(res) =
as.character(c((1-level)/2,1-(1-level)/2))
res[2,] = wilcox.test(object$slopes,conf.int=
TRUE,conf.level=level)$conf.int
res[1,] = wilcox.test(object$intercepts,conf.int=
TRUE,conf.level=level)$conf.int; res}
```

Ako môžeme vidieť, metóda, ktorá je použitá v tomto balíku, nie je tá uvedená v predošlej kapitole. Namiesto toho si autor vybral interval spoľahlivosti určený pomocou Walshových priemerov, ktorý môžeme nájsť napríklad aj v [4]. Tento interval spoľahlivosti môžeme použiť na iid dáta z rozdelenia so symetrickou hustotou a hovorí o strede symetrie tejto hustoty. Pri jeho tvorení najprv vytvoríme tzv. Walshove priemery, čo nie je nič iné, len podľa veľkosti vzostupne usporiadaná množina všetkých priemerov dvojíc dát $X(i)$ vrátane dát $X(i)$. Tvorbu samotného intervalu spoľahlivosti popíšeme neskôr. Táto metóda je ale použitá na objekt "slopes", ktorý nie je ničím iným, len množinou všetkých sklonov priamok vytvorených zo všetkých dvojíc dát, ktoré máme. Jednou zo základných požiadaviek funkčnosti Wilcoxonovho testu (t.j. aby intervaly spoľahlivosti dosahovali aspoň požadovanú spoľahlivosť) je požiadavka nezávislosti objektov, na ktoré test používame. Ľahko sa môžeme presvedčiť, že táto podmienka pre sklony splnená nie je a že je medzi nimi dokonca silná lineárna závislosť.

Veta 2.1. *Predpokladajme, že máme $n \geq 2$ dát. Týchto n nám určuje $n(n-1)/2$ rôznych dvojíc, teda sklonov priamok v tvare $(Y_i - Y_j)/(X_i - X_j), i, j \in \{1, 2, ...n\}, i > j$. Potom na jednoznačné určenie týchto $n(n-1)/2$ sklonov potrebujeme poznať práve $n-1$ z nich.*

Dôkaz. Je jasné, že menej ako $n-1$ nám nebude stačiť (ľahko ukážeme indukciou). Ak znakom s_{ij} označíme taký zo sklonov, ktorý vznikol použitím dát (X_i, Y_i) a (X_j, Y_j), potom výberom sklonov $s_{1,2}, s_{1,3}, \cdots, s_{1.n-1}$ a $s_{2,3}$ (v prípade $n = 3$ nevyberáme $s_{2,3}$) nie sme schopní popísať žiadny zo sklonov v tvare $s_{k,n}$. Ak si ale zvolíme $n-1$ sklonov v tvare $s_{1,2}, s_{1,3}, \cdots, s_{1,n}$, potom ľubovoľný sklon s_{ij} sme schopní vyjadriť nasledujúcim spôsobom.

$$s_{ij} = \frac{s_{1,i}(X_i - X_1) - s_{1,j}(X_j - X_1)}{X_i - X_j}$$

Toto zároveň dokazuje silnú závislosť medzi sklonmi. □

Z toho však nemusí hneď vyplývať, že takýmto spôsobom vypočítané intervaly spoľahlivosti budú mať menšiu spoľahlivosť. V zásade sa môže ešte stať,

že ich spoľahlivosť ostane v poriadku, ale zbytočne sa rozšíria, či dokonca, že budú dosahovať požadovanú spoľahlivosť pri menšej šírke intervalu, akú by mali intervaly zostrojené Theilovou metódou.

V nasledujúcej tabuľke sú uvedené spoľahlivosti konfidenčných intervalov balíka **mblm** pre simulácie s rôznymi počtami dát. Požadovaná spoľahlivosť bola zakaždým 95%. Menili sme hodnotu n, dáta $X(i)$ boli pevne dané konštanty náhodne vybrané z rovnomerného rozdelenia na intervale $(0, 5)$. Hodnota priesečníku bola 3 a hodnota sklonu bola 2. Počet simulácií bol 10000.

Tabuľka 2.1: Spoľahlivosť Wilcoxonovej metódy pre rozdelenie ε z $N(0, 1)$

Počet dát n	5	6	7	10	15	20	40
Spoľahlivosť	88%	87.8%	86.2%	81%	72.6%	66.8%	53.1%

Táto tabuľka nás utvrdzuje v tom, že použitá metoda nie je na mieste. Teoreticky dokážeme, prečo je to tak.

2.2 Dôkaz pre $n = 5$

Pri dôkaze sa obmedzíme na nefunkčnosť 95%–ného intervalu, ale rovnako tak sa týmito úvahami dá postupovať pre inú požadovanú spoľahlivosť.

Značenie

Aby sme dostali Theilov 95% interval spoľahlivosti pre sklon regresnej priamky, musíme najprv tieto sklony usporiadať podľa veľkosti od najmenšieho sklonu po najväčší - nazvime túto usporiadanú množinu T a jej prvky t_i, kde index i udáva pozíciu v T. Ľavú hranicu intervalu spoľahlivosti dáva hodnota t_l, kde $l = (n(n-1)/2 - C_\alpha)/2$ a $C_\alpha = \binom{n}{2} k_{0.025;n}$ ($k_{0.025;n}$ udáva 2.5% kritickú hodnotu Kendallovho rozdelenia s n stupňami voľnosti). Pravá hranica tohto intervalu je symetricky hodnota t_r, kde $r = n(n-1)/2 - l + 1$.

Poznámka
Definícia Kendallovho tau rozdelenia sa dá nájsť napríklad v [4].

Wilcoxonov 95% interval spoľahlivosti je z našich dát vytvorený nasledujúcim spôsobom: z množiny T vytvoríme množinu $W' = \{(t_i + t_j)/2 | t_i, t_j \in T, i \leq j\}$. Za W potom zoberieme usporiadanú množinu W' a jej prvky nazveme podobne ako predtým w_i, kde index i udáva pozíciu vo W. Špeciálne znakom $\overline{w_{ij}}, i \leq j$ budeme označovať ten prvok množiny W, ktorý vznikol ako priemer $(t_i + t_j)/2$. Ak za $N = \binom{n}{2}$ označíme počet prvkov množiny T, počet prvkov W je potom $P = N(N+1)/2$. Za ľavú hranicu intervalu spoľahlivosti berieme w_l, kde $l = s_{0.025;N}$ ($s_{0.025;N}$ udáva 2.5% kvantil Wilcoxonového znamienkového rozdelenia). Pravá hranica tohto intervalu je potom symetricky hodnota w_r, kde $r = N - l + 1$.

Poznámka
Pod Wilcoxonovým znamienkovým rozdelením rozumieme nasledovnú konštrukciu. Nech x je náhodný výber veľkosti N zo spojitej a symetrickej distribúcie. Potom Wilcoxonova znamienková štatistika je sumou rankov absolútnych hodnôt x_i, pričom x_i je kladné. Táto štatistika nadobúda hodnoty od 0 do $N(N+1)/2$ a je symetricky rozdelená okolo hodnoty $N(N+1)/4$.

Pozrime sa teraz na niektoré základné vlastnosti množiny W. Z definície $w_i \leq w_j$ práve vtedy, keď $i \leq j$ (rovnako $t_i \leq t_j$ práve vtedy, keď $i \leq j$). Ďalej tiež z definície $\overline{w_{ii}} = t_i$ a zjavne $w_1 = \overline{w_{1,1}} = t_1$ a $w_P = \overline{w_{N,N}} = t_N$. O niečo menej triviálna vlastnosť je nasledujúca implikácia, ktorá síce priamo v dôkaze nevystupuje, ale veľmi pomohla pri jeho tvorbe.

Veta 2.2. *Ak $w_q = \overline{w_{i,j}}$, potom $q \geq ij - i(i-1)/2$.*

Dôkaz. Ukážeme, že pre dané $i, j \in \{1, 2, ...N\}, i \leq j$ platí

$$\forall m, n \in \mathbb{N}, m \leq i, n \leq j, m \leq n : \overline{w_{i,j}} \geq \overline{w_{m,n}}. \qquad (*)$$

Táto nerovnosť sa dá z definície prepísať na $(t_i + t_j)/2 \geq (t_m + t_n)/2$, ktorá

zjavne platí, lebo $t_i \geq t_m$ a $t_j \geq t_n$. Keďže takýchto dvojíc indexov m, n je $ij - i(i-1)/2$, existuje aspoň $ij - i(i-1)/2$ prvkov W, ktoré sú menšie alebo rovné ako $\overline{w_{i,j}}$, a teda index q poradia tohto prvku musí byť aspoň $ij - i(i-1)/2$. □

Ukážeme, že pre $n = 5$ je hodnota t_l vždy menšia ako hodnota w_l (hodnoty označujú ľavé hranice príslušných intervalov spoľahlivosti), a preto je interval spoľahlivosti použitý v balíku **mblm** vždy užší ako ten Theilov.
Príslušné hodnoty jednotlivých rozdelení nám určujú Theilov interval spoľahlivosti ako (t_1, t_{10}), pričom množina T má 10 prvkov a Wilcoxonov interval spoľahlivosti ako (w_9, w_{47}), pričom množina W má 55 prvkov.

Veta 2.3. *Platí* $t_1 < w_9$.

Dôkaz. Tvrdenie je zjavné, lebo platí $t_1 = \overline{w_{1,1}} = w_1 < w_9$. □

Tento fakt samotný ale nestačí na to, aby sme mohli prehlásiť, že druhý interval má menšiu spoľahlivosť ako 95%. Môžeme si však všimnúť, že veľmi často by sa malo stávať, že dokonca $w_9 \geq t_2$. Pozrime sa teda, akej spoľahlivosti zodpovedá interval (t_2, t_9) - označme ju p. Musí platiť $2 = (n(n-1)/2 - C_\alpha)/2$, t.j. $C_\alpha = 6$. To znamená, že $k_{(1-p)/2;5} = 0.6$. Z toho plynie, že $(1-p)/2 = 1 - D(0.6, 5) = 0.041667$ (D označuje distribučnú funkciu Kendallovho rozdelenia), čo dáva $p = 91,67\%$ (stačí nám aj odhad 92%, ktorým môžeme prehliadnuť aj chyby, čo vznikli pri zaokrúhľovaní). Je preto možné, že sa nám podarí ukázať, že úspešnosť intervalu v mblm.R je naozaj menšia ako 95%. Naším cieľom bude odhadnúť pravdepodobnosť úspechu intervalu (w_9, w_{47}) zhora tak, aby sme výsledný výraz vedeli ľahko spočítať a ukázať, že je menší ako 95%.

Lema 2.1. *Udalosť* $t_2 > w_9$ *nastáva práve vtedy, keď udalosť* $2t_2 > t_1 + t_9$.

Dôkaz. Uvažujme najprv, že deviaty prvok W, t.j. w_9, je menší ako $t_2 = \overline{w_{2,2}}$. To však znamená, že žiaden z deväťice najmenších prvkov W nie je v tvare $\overline{w_{k,l}}, k > 1, l \geq k$, pretože ak by tomu tak bolo, tak potom $\overline{w_{k,l}} \leq w_9 < \overline{w_{2,2}}$, čo je spor s **(*)**. To znamená, že deviaty najmenší prvok vo W je $w_9 = \overline{w_{1,9}}$,

z čoho už vyplýva $t_2 > w_9 = (t_1 + t_9)/2$.
Naopak, ak platí $t_2 > (t_1 + t_9)/2 = \overline{w_{1,9}}$, potom existuje aspoň deväťica prvkov v tvare $\overline{w_{1,l}}, 9 \geq l \geq 1$, z ktorých každý je menší ako prvok $t_2 = \overline{w_{2,2}}$, teda deviaty najmenší prvok množiny W musí byť tiež menší ako t_2, z čoho plynie $t_2 > w_9$. □

Dôsledok: Vďaka obmene tohto tvrdenia dostávame $t_2 \leq w_9 \Leftrightarrow 2t_2 \leq t_1 + t_9$. Zo symetrie dôkazu ďalej vyplýva $t_9 < w_{47} \Leftrightarrow 2t_9 < t_2 + t_{10}$. Potom obmenou platí aj $t_9 \geq w_{47} \Leftrightarrow 2t_9 \geq t_2 + t_{10}$.

Veta 2.4. *Označme doleuvedené pravdepodobnosti nasledovne:*

$$p_1 = P(\beta_1 \in (t_2, t_9) \wedge 2t_2 \leq t_1 + t_9 \wedge 2t_9 \geq t_2 + t_{10})$$

$$p_2 = P(\beta_1 \in (t_2, t_{10}) \wedge 2t_2 \leq t_1 + t_9 \wedge 2t_9 < t_2 + t_{10})$$

$$p_3 = P(\beta_1 \in (t_1, t_{10}) \wedge 2t_2 > t_1 + t_9 \wedge 2t_9 < t_2 + t_{10})$$

$$p_4 = P(\beta_1 \in (t_1, t_9) \wedge 2t_2 > t_1 + t_9 \wedge 2t_9 \geq t_2 + t_{10})$$

Potom $P(\beta_1 \in (w_9, w_{47})) \leq p_1 + p_2 + p_3 + p_4$.

Dôkaz. Celý pravdepodobnostný priestor vieme rozdeliť na tieto štyri disjunktné podmnožiny (môžu byť aj prázdne):

$$A : t_2 \leq w_9 \wedge w_{47} \leq t_9$$

$$B : t_2 \leq w_9 \wedge t_9 < w_{47}$$

$$C : w_9 < t_2 \wedge t_9 < w_{47}$$

$$D : w_9 < t_2 \wedge w_{47} \leq t_9$$

Potom ak udalosť $\beta_1 \in (w_9, w_{47})$ nazveme U, podľa vety o úplnej pravdepodobnosti dostávame

$$P(U) = P(A)P(U|A) + P(B)P(U|B) + P(C)P(U|C) + P(D)P(U|D).$$

Venujme sa teraz jednotlivým sčítancom osobitne. Zjavne

$$P(A)P(U|A) \leq P(A)P(\beta_1 \in (t_2, t_9)|A) = P(\beta_1 \in (t_2, t_9) \wedge A).$$

Použitím Lemy 2.1 dostávame

$$P(A)P(U|A) \leq p_1.$$

Ďalej keďže t_1 a t_{10} sú najmenším a najväčším prvkom v T aj vo W, platia nasledovné odhady:

$$P(B)P(U|B) \leq P(B)P(\beta_1 \in (t_2, t_{10})|B) = P(\beta_1 \in (t_2, t_{10}) \wedge B).$$

Z Lemy 2.1 dostávame

$$P(B)P(U|B) \leq p_2.$$

Ďalej

$$P(C)P(U|C) \leq P(C)P(\beta_1 \in (t_1, t_{10})|C) = P(\beta_1 \in (t_1, t_{10}) \wedge C).$$

Opäť z Lemy 2.1 dostávame

$$P(C)P(U|C) \leq p_3.$$

A nakoniec

$$P(D)P(U|D) \leq P(D)P(\beta_1 \in (t_1, t_9)|D) = P(\beta_1 \in (t_1, t_9) \wedge D),$$

z čoho použitím Lemy 2.1 dostávame

$$P(C)P(U|C) \leq p_4,$$

čo spolu dokazuje naše tvrdenie. □

Teraz nám stačí len vyčísliť pravdepodobnosť $p_1 + p_2 + p_3 + p_4$ očistenú od hodnôt w. Vypočítať túto pravdepodobnosť tiež nie je veľmi jednoduché, ale ide o značné zjednodušenie. Uvedieme spôsob, akým sa to dá.
Sklony t sa dajú napísať ako $(Y_i - Y_j)/(X_i - X_j), i, j \in \{1, 2, ...n\}, i > j$, čo môžeme využitím $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ prepísať do tvaru $\beta_1 + (\varepsilon_i - \varepsilon_j)/(X_i - X_j)$. Keďže t_1, t_2, t_9 a t_{10} sú zakaždým iné sklony (v zmysle z ktorej dvojice dát vznikli), potrebujeme rozlíšiť všetky možné permutácie sklonov, ktorých je 10!. Tieto permutácie nám potom deviatimi nerovnosťami určujú všetky

možné pätice $(\varepsilon_1, \varepsilon_2, ...\varepsilon_5)$, z ktorých za vyhovujúce považujeme len tie spĺňajúce všetky príslušné podmienky obsiahnuté v jednotlivých p_i. Takýmto spôsobom sú hranice integrovania jednoznačne určené, a preto by nemal byť problém jednotlivé integrály vypočítať pre akékoľvek konštanty X_i a rozdelenie chyby ε. Simuláciou pre chybu z rozdelenia $N(0,1)$ a $X_i = i$ naozaj dostávame výsledok menší ako 95%, konkrétne 93.77%.
Počet potrebných výpočtov sa dá ešte zmenšiť pomocou nasledujúceho tvrdenia. To vraví, že nepotrebujeme uvažovať o všetkých permutáciách, lebo niektoré nikdy nemôžu nastať.

Veta 2.5. *Nech* $X_1 < X_2 < ... < X_n$ *určujú sklony* $s_{ij} = (Y_i - Y_j)/(X_i - X_j), i, j \in \{1, 2, ...n\}, i > j$. *Potom najväčší aj najmenší z nich sú v tvare* $s_{i+1,i}$ *a pre ľubovoľné tri za sebou idúce indexy* $i, i+1, i+2$, *kde* $i \in \{1, 2, ...n(n-1)/2 - 2\}$, *platí, že* $s_{i+2,i}$ *nie je najväčší ani najmenší z trojice* $s_{i+1,i}, s_{i+2,i+1}$ *a* $s_{i+2,i}$.

Dôkaz. Nech $a < b < c$. Ukážeme, že až na prípad, keď všetky tri body $[X_a, Y_a], [X_b, Y_b]$ a $[X_c, Y_c]$ ležia na priamke, platí buď $s_{ba} < s_{ca} < s_{cb}$, alebo $s_{ba} < s_{ca} < s_{cb}$. Sporom, nech s_{ca} je najväčší z nich. Potom platí

$$\beta_1 + (\varepsilon_c - \varepsilon_a)/(X_c - X_a) > \beta_1 + (\varepsilon_b - \varepsilon_a)/(X_b - X_a)$$

a zároveň

$$\beta_1 + (\varepsilon_c - \varepsilon_a)/(X_c - X_a) > \beta_1 + (\varepsilon_c - \varepsilon_b)/(X_c - X_b).$$

Ekvivalentne

$$(\varepsilon_c - \varepsilon_a)(X_b - X_a) > (\varepsilon_b - \varepsilon_a)(X_c - X_a)$$

$$(\varepsilon_c - \varepsilon_a)(X_c - X_b) > (\varepsilon_c - \varepsilon_b)(X_c - X_a).$$

Sčítaním oboch nerovností dostávame $(\varepsilon_c - \varepsilon_a)(X_c - X_a) > (\varepsilon_c - \varepsilon_a)(X_c - X_a)$, čo je spor. Analogicky pre prípad, že by bol s_{ca} je najmenší z nich. Preto ak pre dané $a, c \in \{1, 2, ...n\}, a < c$ existuje b z tejto množiny väčšie ako a a menšie ako c, potom s_{ca} zaručene nie je najväčším ani najmenším sklonom, z čoho už plynie naše tvrdenie. □

Poznámka
V každom z dôkazov ignorujeme fakt, že sa teoreticky môže stať, aby sa nejaké sklony alebo Walshove priemery rovnali aj napriek inému indexu. Takéto situácie ale nastávajú s nulovou pravdepodobnosťou, a preto náš cieľ odhadovať pravdepodobnosti zhora nijako negatívne neovplyvňujú.

2.3 Prípady $n = 3$ a $n = 4$

Ak skúmame len tri dáta, metóda v balíku mblm.R bude spoľahlivou, lebo za interval spoľahlivosti vyhlási interval $(-\infty, \infty)$. Tento interval však dostaneme aj použitím Theilovej metódy (bude určený ako "nultý" a "N+prvý" sklon), teda sa o ňom nedá povedať, že by bol zbytočne široký.

V prípade $n = 4$ sa da triviálne preukázať, že metóda v mblm.R je zlá. Za 95%–ný interval spoľahlivosti máme totiž zobrať interval (w_1, w_{21}), čo je samozrejme interval totožný s intervalom (t_1, t_6), ktorého spoľahlivosť vieme ľahko vyjadriť integrálom ako v predošlej kapitole a dostávame číslo menšie ako 93%.

Kapitola 3

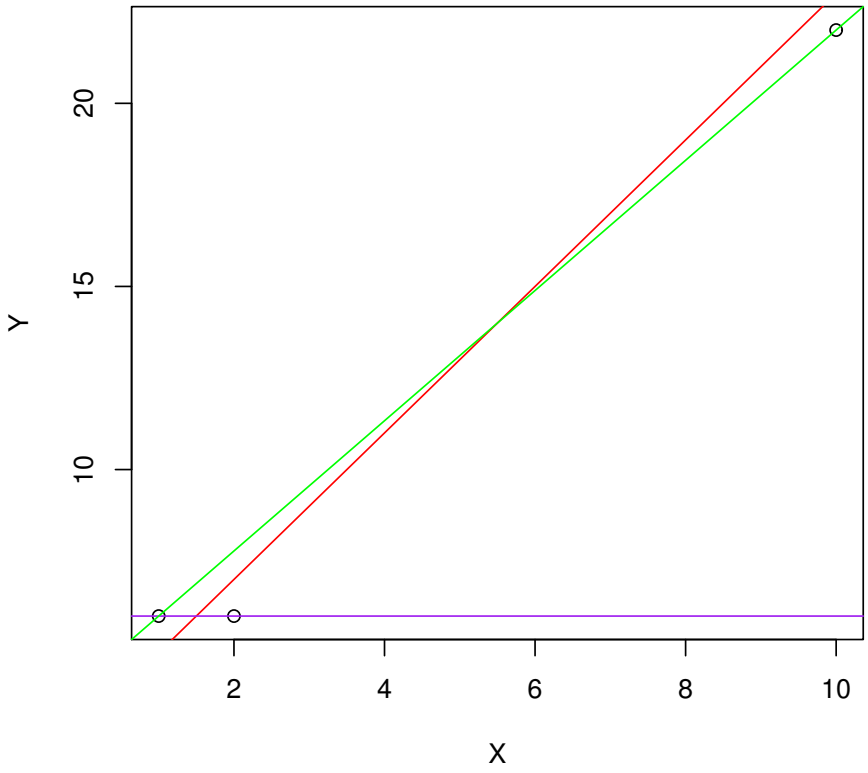
Odhad pre sklon regresnej priamky

V tejto kapitole sa pokúsime vytvoriť nový odhad pre sklon regresnej priamky, ktorý bude lepší ako Theilov alebo odhad pomocou MNŠ v niektorých špeciálnych prípadoch. Je preto logické, že sa budeme snažiť vytvoriť ho tak, aby neniesol nedostatky týchto dvoch spomínaných odhadov.

3.1 Nedostatky Theilovho odhadu a odhadu MNŠ

Ako je spomenuté už v prvej kapitole, odhad MNŠ je síce najlepším možným lineárnym odhadom, ak chyby sú z normálneho rozdelenia, ale je veľmi nerobustný. Je to spôsobené tým, že je váženým priemerom sklonov (odvodenie váh sa dá nájsť aj v [1]). Najjednoduchší spôsob, ako odstrániť problémy spôsobené nerobustnosťou priemeru, je použiť namiesto neho medián. To samozrejme spĺňa už Theilov odhad, preto sa pozrime, kde sú jeho nedostatky. Pripomeňme si, že Theilov odhad dostaneme ako medián sklonov všetkých priamok, ktoré vieme vytvoriť z dvojíc dát. Otázka ale je, či dvojica dát nesie dostatočne dôveryhodnú informáciu o skutočnom sklone regresnej priamky. Odpoveď na túto otázku môžeme nájsť po zhliadnutí nasledujúceho obrázku.

Záleží na tom, o akú dvojicu dát ide.



Obr. 3.1: Porovnanie vierohodnosti dvojíc dát

Jasne vidíme, že malá výchylka dát, ktoré sa líšia v X–ovej súradnici len o málo, spôsobí veľkú zmenu sklonu (fialová priamka). Naopak, ak sú tieto dáta vzdialené dostatočne, malá chyba neovplyvní sklon až tak veľmi (zelená priamka). Preto by sme mali viac dôverovať takým sklonom, ktoré vznikli z dát s väčším rozdielom v X–ovej súradnici. Námietka voči takémuto myšlienkovému postupu by mohla byť, že aj tak používame medián, ale predstavme si situáciu, kde máme desať dát s ekvidištantne rozdelenými X. Potom ak za málo dôveryhodné dáta budeme pokladať také, ktorých rozdiel v X je nanajvýš o dve jednotky, zo všetkých 45 sklonov sme vybrali 17, čo je viac

ako tretina. Tretina dát, ktoré nebudú až tak dôveryhodné, nám môže značne vychýliť aj medián, pokiaľ sa stane, že viac z nich bude pod skutočnou hodnotou sklonu ako nad ňou (na rozdiel od prípadu, keby to bolo opačne, medián sa tak zaručene zmení).

3.2 Vylepšený Theilov odhad

Zmeníme algoritmus nasledujúcim spôsobom. Namiesto všetkých sklonov budeme brať množinu iba tých sklonov, ktoré vznikli z dát, pre ktoré platí, že ich index je vzdialený aspoň o hodnotu parametra p. Potom vyberieme medián z tejto ochudobnenej množiny.
Takto potrebujeme už len určiť veľkosť parametra p, ktorý nám udáva, ktoré dáta už sú dostatočne vzdialené, aby sme ich brali do úvahy. Testovanie sme robili na modeli s ekvidištantnými $X(i) = i, i = 1, 2, \cdots, n$, sklonom $b = 2$ a priesečníkom $a = 3$. Rozdelenie chyby sme menili, rovnako tak n.

3.3 Porovnanie s odhadom Theila

Po skúšaní rôznych hodnôt p sme si zvolili hodnotu 5. V nasledujúcich tabuľkách vidíme, ako sa darilo nášmu odhadu v porovnaní s odhadom Theila. Druhý stĺpec hovorí, koľkokrát bol náš odhad aspoň taký dobrý (to znamená, že absolútna hodnota výchylky bola menšia alebo rovná), v stĺpci Odchýlka 1 je priemer odchýlok Theilovho odhadu, v ďalšom toho nášho.

Tabuľka 3.1: Porovnanie nášho odhadu s odhadom Theila $n = 10$

Rozdelenie chyby	Úspešnosť	Odchýlka 1	Odchýlka 2
N(0,1)	60.8%	0.09407183	0.09455158
Cauchy $x_0 = 0$, $\gamma = 1$	54.9%	0.2069137	0.242797
Laplaceovo $\lambda = 1$	59.1%	0.1111466	0.1145499

Tabuľka 3.2: Porovnanie nášho odhadu s odhadom Theila $n = 20$

Rozdelenie chyby	Úspešnosť	Odchýlka 1	Odchýlka 2
N(0,1)	54.1%	0.03255317	0.03276727
Cauchy $x_0 = 0$, $\gamma = 1$	54.6%	0.06436257	0.06542679
Laplaceovo $\lambda = 1$	55.1%	0.03889127	0.03918769

Tabuľka 3.3: Porovnanie nášho odhadu s odhadom Theila $n = 50$

Rozdelenie chyby	Úspešnosť	Odchýlka 1	Odchýlka 2
N(0,1)	55.5%	0.007921133	0.007911898
Cauchy $x_0 = 0$, $\gamma = 1$	52.6%	0.01513806	0.01512285
Laplaceovo $\lambda = 1$	56.0%	0.009138342	0.009131859

Tabuľka 3.4: Porovnanie nášho odhadu s odhadom Theila $n = 100$

Rozdelenie chyby	Úspešnosť	Odchýlka 1	Odchýlka 2
N(0,1)	53.1%	0.002826436	0.002822261
Cauchy $x_0 = 0$, $\gamma = 1$	51.5%	0.004991336	0.004996031
Laplaceovo $\lambda = 1$	51.4%	0.003298931	0.003298308

Môžeme vidieť, že pre každé skúmané n sme mali lepší odhad vo viac ako polovici prípadov. Pre väčšie n sme dokonca dosiahli aj menšiu priemernú odchýlku. Je teda možné, že vhodnou voľbou parametra a (v závislosti od hodnôt $X(i)$ a variancie rozdelenia chyby) môžeme dosiahnuť lepšie výsledky, ako dosahuje Theilov odhad a zároveň si zachovať robustnosť, čo nám zaručí lepšie výsledky ako odhad pomocou MNŠ pre chyby z rozdelení s ťažkými chvostami.

Ďalším návrhom by mohlo byť použitie spojenia MNŠ a Theila nasledujúcim spôsobom: medián neaplikujeme na sklony priamok vytvorených z dvojíc bodov, ale z k–tic vytvorených pomocou MNŠ.

Záver

V práci skúmame špeciálny druh priamkovej lineárnej regresie – Theilovej regresie. Prvá kapitola jasne ukazuje, že Theilova regresia má oproti klasickej metóde najmenších štvorcov mnoho výhod. Kým pri chybách z normálneho rozdelenia zaostáva v kvalite odhadu sklonu len o veľmi málo, pri všetkých iných skúmaných rozdeleniach alebo pri výskyte outlierov je omnoho lepšia. Najmä pri rozdeleniach s najťažšími chvostami je metóda najmenších štvorcov úplne nepoužiteľná. Theilovu metódu odporúčame použiť vždy vtedy, ak si nie sme úplne istí normálnosťou chýb. Na druhej strane, táto metóda sa dá použiť len pri priamkovej regresii.

V druhej kapitole sa nám podarilo simulačne preukázať nespoľahlivosť balíka mblm.R a dokonca toto tvrdenie podporiť aj dôkazom pre $n = 5$. Tento dôkaz sa síce spoliehal na výpočet veľkého počtu konečných integrálov, a preto bolo lepšie rátať ich Monte Carlo simuláciou, ale v konečnom čase je možné ich jednoducho vypočítať. Tieto dôkazy sa nepodarilo úplne zovšeobecniť, ale ich konštrukcia je zrejme použiteľná aj pre ľubovoľné iné n. Balík už momentálne v obehu nie je a na ďalší so správne implementovanými Theilovymi metódami sa ešte čaká.

V tretej kapitole sa pokúšame o vytvorenie nového odhadu, ktorý by bol dobrý v situáciách, keď Theil a MNŠ nefungujú až tak dobre. Dostávame sa k tomu, že Theilovu metódu je možné jemne upraviť tak, aby v istých situáciách dosahovala lepšie výsledky. Pre konkrétne prípady ale treba túto úpravu najprv empiricky zistiť.

Literatúra

[1] Hollander M., Wolfe D. A. 1973. *Nonparametric statistical methods* John Wiley & Sons, Inc., New York.

[2] Komsta Lukasz 2005. *mblm: Median-Based Linear Models* R package version 0.1.

[3] Pázman A., Lacko V. 2012. *Prednášky z regresných modelov* vysokoškolské skriptá, Univerzita Komenského, Bratislava.

[4] Rublík František 1993. *Neparametrické metódy a štatistická kontrola akosti* vysokoškolské skriptá, Univerzita Komenského, Bratislava.