Univerzita Karlova v Praze

Fakulta sociálních věd

Institut ekonomických studií

BAKALÁŘSKÁ PRÁCE

2007

Jaroslav Hlávka

Univerzita Karlova v Praze

Fakulta sociálních věd

Institut ekonomických studií

BAKALÁŘSKÁ PRÁCE

Robustní ekonometrické modely - simulační analýza the Least Weighted Squares

Vedoucí bakalářské práce: Prof. RNDr. Jan Ámos Víšek, CSc.

Vypracoval: Jaroslav Hlávka

Akademický rok: 2006/2007

Rád bych na tomto místě poděkoval profesoru J. Á. Víškovi za hodnotné rady a odborné vedení v průběhu celého mého psaní. Jsem mu vděčný za poskytnutí mnoha materiálů a programů, ze kterých jsem mohl vycházet při psaní mé práce.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

v Praze dne Jaroslav Hlávka

Obsah

Název práce:
Robustní ekonometrické modely - simulační analýza the Least Weighted Squares
Autor: Jaroslav Hlávka
Katedra (ústav): Institut ekonomických studií
Vedoucí bakalářské práce: Prof. RNDr. Jan Ámos Víšek, CSc.
e-mail vedoucího: visek@fsv.cuni.cz

Abstrakt: Tento text pojednává o dosavadních pokrocích v oblasti robustních odhadů regresních parametrů. Nejprve bude čtenář uveden do oblasti regresních odhadů. Vysvětlíme klasické metody odhadu, jejich vlastnosti a nedostatky. V druhé části práce zavedeme robustní vlastnosti odhadu, zmíníme některé robustní metody. Nakonec se zaměříme na nejnovější metodu tzv. nejmenší vážené čtverce (the Least Weighted Squares) a popíšeme její vlastnosti. V závěru práce podrobíme metodu LWS simulační analýze a okomentujeme výsledky.
Klíčová slova: regrese, robustní, ekonometrické modely, Least Weighted Squares.

Title: Robust econometric models - simulation analysis of the Least Weighted Squares
Author: Jaroslav Hlávka
Department: Institute of economic studies
Supervisor: Prof. RNDr. Jan Ámos Víšek, CSc.
Supervisor's e-mail address: visek@fsv.cuni.cz

Abstract: This text is focused on advancements in the field of robust econometrics. First we will introduce the basics of regression analysis. The Ordinary Least squares will be shown as inefficient under violations of classical assumptions. In the next part of this text, robust properties will be introduced. The most important robust methods will be described. In the end an analysis will examine the properties of the most recent method of the Least Weighted Squares.
Keywords: regression, robust, econometric models, Least Weighted Squares.

"Chybovat je lidské, odpouštět božské,
ale zahrnovat možnost výskytu chyby
do svého projektu, to je statistické."

Leslie Kish

0.1 Úvod

Robustní metody odhadu parametru jsou reakcí na kritiku klasických statistických metod. Ta se týká velmi silných podmínek, za kterých klasické odhady fungují dobře. Tyto podmínky, které podrobněji uvedeme v následujícím textu, se však ukazují v reálných případech jako stěží dosažitelné. Problémem klasických metod, které se vyvíjejí už několik století, je fakt, že fungují dobře pouze při splnění těchto předpokladů. Pokud se však vlastnosti naměřených dat od těchto předpokladů jen nepatrně odlišují, ztrácejí klasické metody požadované vlastnosti. Tento fakt vede k potřebě definovat vedle klasických vlastností dobrého odhadu navíc další vlastnosti, které zaručí jeho použitelnost i v případech, kdy se naměřená data nechovají přesně podle předpokladů. Je rozumným požadavkem, aby i v těchto případech náš odhad, když už ne nejlépe alespoň přibližně, popisoval hledaný model a při malé odchylce od předpokladů nedával (zcela) zavádějící výsledky.

Přijde nám rozumné domnívat se, že skutečné vztahy mezi různými věcmi jsou mnohem složitější, než modely, kterými se je snažíme popsat. Dá se říci, že tyto modely mohou přinejlepším skutečnost pouze aproximovat. A to, do jaké míry se jim to daří, vypovídá o jejich kvalitě. Je bláhové věřit, že pomocí vědy a exaktních metod dokážeme objevit všechny zákonitosti našeho světa. Vždyť existují vličiny, které nedokážeme ani změřit, zvážit, či jinak kardinálně ohodnotit. Například v ekonomii je jednou ze základních veličin užitek, na kterém stojí celá neoklasická ekonomie. Abychom mohli aplikovat statistické metody na modely, ve kterých vystupuje užitek jako vysvětlující proměnná, museli bychom hodnoty užitku mít naměřené. To však není možné, a tak nám nezbývá než se smířit s tím, že můžeme přinejlepším použít nějakou jinou veličinu, která je s užitkem významně korelovaná. Tímto přistupujeme na to, že model, který se snažíme najít, bude pouhým přiblížením realitě. Toto však není nic, co by nás mělo od statistického

hledání závislostí odrazovat. Jen je třeba na tento důležitý aspekt nezapomínat, a mít ho stále na mysli při interpretaci výsledků analýzy.

Dalším důvodem pro zabývání se robustními metodami je existence chyb. Je docela běžné, že data, která dostane statistik či ekonometr k analýze, obsahují chyby. Nejčastěji se jedná o chybně naměřené hodnoty, výjimkou není chybně zapsaná desetinná čárka. Takto kontaminovaná data mohou velmi výrazně ovlivnit výsledky analýzy v špatném (ale i v dobrém) smyslu. Klasické metody, jako je například metoda nejmenších čtverců, jsou na takové chyby citlivé, a proto se začaly hledat jiné, robustní metody, které by se dokázaly s chybami mezi daty vypořádat. To znamená, že buďto jsou vůči vychýleným datům "imunní", nebo dokáží chybná data rozpoznat, a v procesu regresní analýzy s nimi jednat jinak než s dobrými daty. Například již ve starověkém Egyptě používali v jistém smyslu robustní metody při výpočtu statistik. Jsou doložené případy, kdy ze vzorku dat, ze kterých chtěli počítat průměrnou hodnotu, vyřadili před samotným výpočtem odlehlá pozorování. Tento postup je jedním z robustních postupů, kterému se budeme věnovat i v tomto textu. Otázkou však zůstává, proč se formalizace robustních metod, a její hlubší analýza objevuje teprve v 70. letech 20. století. Jedním z důvodů může být například rozvoj výpočetní techniky, který spadá do stejného období. V minulosti bylo nutné veškeré výpočty provádět ručně, na papíře. S nástupem počítačů se otevírají nové možnosti odhadů a jejich výpočtů.

Tento text si klade za cíl pokusit se shrnout dosavadní bádání v oblasti robustních odhadů do uceleného souboru a přiblížit čtenáři výhody a nevýhody jednotlivých metod. Protože za posledních 40 let bylo navrženo nepřeberné množství alternativních metod, omezíme se v této práci pouze na ty nejvýznamnější, které nějakým způsobem posunuly bádání v této oblasti dále. V první části práce zavedeme základní pojmy a nastíníme historii vývoje regresní analýzy. Zopakujeme metodu nejmenších čtverců a klasické vlastnosti, jaké by měl mít ideální odhad. V další části textu rozšíříme klasické vlastnosti odhadu o požadavky, které vycházejí z robustních teorie. Uvedeme evoluci robustních odhadů a jejich vlastnosti. V závěru práce bude pomocí simulace proveden test vybraných robustních metod.

Kapitola 1

Klasické metody odhadu

1.1 Zavedení základních pojmů

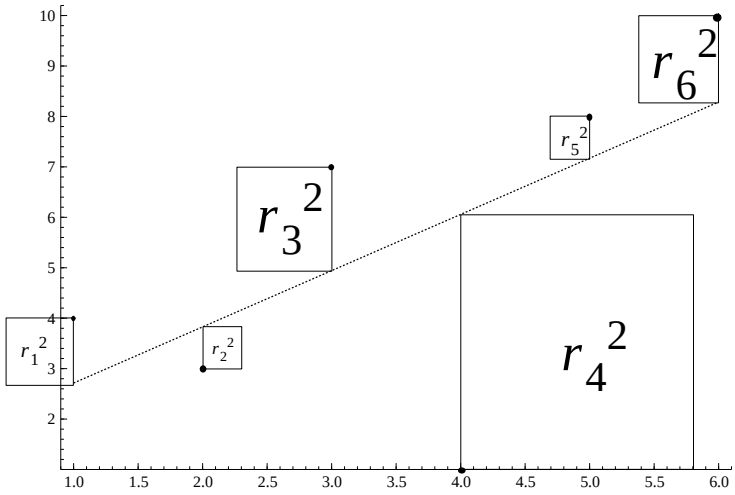
Abychom zabránili možnému nedorozumění v následujícím textu, uveďme základní značení, které budeme používat. V dalším textu budeme označovat $\mathbb{N}$ množinu všech přirozených čísel, $\mathbb{R}$ reálnou osu a $\mathbb{R}^p$ p-rozměrný Euklidovský prostor. Všechny vektory budou považovány za sloupcové, pokud nebude explicitně uvedeno jinak. V následující rovnici (1.1) tak výraz $\mathbf{X}_i^T$ považujeme za řádkový vektor (neboli transponovaný sloupcový vektor). Mějme $n \in \mathbb{N}$ pozorování, regresním modelem budeme dále rozumět rovnici

$$Y_i = \beta_1^0 X_{i1} + \beta_2^0 X_{i2} + \beta_3^0 X_{i3} + \ldots + \beta_p^0 X_{ip} + \varepsilon_i = \mathbf{X}_i^T \beta^0 + \varepsilon_i, \quad i \in \{1, \ldots, n\} \tag{1.1}$$

kde $Y_i \in \mathbb{R}$ je hodnota vysvětlované proměnné a $\mathbf{X}_i \in \mathbb{R}^p$ je vektor hodnot vysvětlujících proměnných i-tého pozorování. β^0 je p-rozměrný vektor regresních koeficientů. Označením β^0 budeme chápat skutečnou hodnotu parametru β, zatímco $\hat{\beta}$ bude značit náš (nějaký) odhad. Člen ε_i je prvkem posloupnosti náhodných veličin. V tomto textu budeme vždy uvažovat model s interceptem, tedy vektor $\mathbf{X}_i^T$ bude mít na 1. místě vždy hodnotu 1 ($X_{i1} = 1, i \in \{1, \ldots, n\}$).

Když se pro jednoduchost omezíme jen na dvourozměrný prostor $\mathbb{R}^2$, můžeme přirovnat pátrání po nejlepších odhadech regresního modelu (1.1) k prokládání "nejlépe

Obrázek 1.1: Metoda nejmenších čtverců



sedící"[1] přímky oblakem dat, který vyneseme do roviny. Polohu a směr této přímky nám udávají právě koeficienty β_i^0. Co se týče hledání regresních koeficientů ve více rozměrném prostoru, je tato představa stále možná, avšak vyžaduje mnohem více fantazie (namísto přímky hledáme nadrovinu). Možností, jakým způsobem je možné hledat hodnoty těchto koeficientů, není málo. Nejpopulárnější se stal, také díky jednoduchosti výpočtu, odhad metodou nejmenších čtverců. Tuto metodu pro výpočet regresních koeficientů použil Galton[2] poprvé v roce 1886 a od té doby si získala značnou popularitu.[3] Podívejme se tedy na tuto metodu blíže.

1.2 Metoda nejmenších čtverců

Pokud opět použijeme přiblížení z minulého odstavce, můžeme si metodu nejmenších čtverců názorně ukázat na obrázku 1.1. Hledáme takovou přímku (jednoznačně určenou koeficienty), pro kterou bude součet druhých mocnin vzdáleností jednotlivých bodů od

[1] Tato vlastnost může být nahlížena z mnoha úhlů pohledu a volba způsobu měření je vlastně volba samotné regresní metody.

[2] Galton, F. (1886), *Regression towards mediocrity in herediatry stature*

[3] Název této práce dal vzniknout jménu modelu (1.1). Nutno dodat, že jako první byla o mnoho let dříve použita metoda minimalizující součet absolutních hodnot reziduí.

přímky měřených ve směru osy y nejmenší. Tuto vzdálenost pojmenujme *reziduum* a označme i-té reziduum

$$r_i = Y_i - \mathbf{X}_i^T \hat{\beta}^{(OLS,n)}, \tag{1.2}$$

kde $\hat{\beta}^{(OLS,n)}$ značí náš odhad metodou nejmenších čtverců. Je vidět, že pokud existuje přímka, která prochází všemi body, pak tento součet bude nula, v ostatních případech bude součet vždy kladný.

Aby bylo možné pohodlněji sčítat jak rezidua pro body ležící nad přímkou, tak pro body ležící pod ní, používáme druhou mocninu vzdálenosti. Tím zabezpečíme kladné sčítance pro všechna rezidua. Tento krok však, jak ukážeme v následujících kapitolách, výrazně zvyšuje vliv odlehlých pozorování, která způsobují kontaminaci dat, na hodnoty regresních koeficientů, a tím i zkreslují výsledky analýzy.

Odhadem metodou nejmenších čtverců tedy budeme rozumět takový vektor, pro který platí

$$\hat{\beta}^{(OLS,n)} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i^T \beta)^2, \quad \hat{\beta}^{(OLS,n)} \in \mathbb{R}^p. \tag{1.3}$$

Matematickými metodami, které zde nebudu rozvádět do detailů (zájemce odkazujeme na skripta John, et al. (2003), kapitola XI), dospějeme k následující soustavě normálních rovnic, které musí splňovat hledaný vektor koeficientů $\hat{\beta}^{(OLS,n)}$

$$\begin{aligned} \textstyle\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \hat{\beta}^{(OLS,n)}) X_{i1} &= 0 \\ &\vdots \\ \textstyle\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \hat{\beta}^{(OLS,n)}) X_{ip} &= 0 \end{aligned} \tag{1.4}$$

Tuto soustavu už umíme řešit a dokonce víme, že má za určitých okolností právě jedno řešení. Tím řešením bude námi hledaný odhad vektoru regresních koeficientů β. Pokud se podíváme na výše uvedenou soustavu rovnic důkladněji, je možné nahlédnout, že lze tuto soustavu přepsat pomocí maticového zápisu následovně[4]

$$\hat{\beta}^{(OLS,n)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \tag{1.5}$$

Rovnice zapsaná v tomto tvaru nám ukazuje další důležitou vlastnost, díky které si metoda OLS získala výsadní postavení, a tou je jednoduchost počítání. V 19. století

[4]Podrobný postup lze nalézt ve skriptech Víšek (1997).

byl každý výpočet spojen s tužkou a papírem, a proto se vždy hledala nějaká zjednodušení, aby byl výpočet buďto alespoň snazší, nebo dokonce vůbec možný. Není proto divu, že se metoda OLS stala nejpoužívanější metodou pro odhady parametrů. Vždyť pro výpočet odhadů koeficientů nám stačí "pouhých pár" statistik získaných z dat a s těmi provést jen (relativně) jednoduché matematické operace.

Myslíme, že je dobré se na tomto místě na chvíli zastavit a podívat se na tuto vlastnost detailně. Matice $\mathbf{X}^T\mathbf{X}$ má na ij-té souřadnici hodnotu $\sum_{k=1}^{n} X_{ik}X_{jk},\ k \in \mathbb{N}$. Obdobně matice $\mathbf{X}^T\mathbf{Y}$[5] má na pozici i člen $\sum_{l=1}^{n} X_{il}Y_l$. Tedy pro toho, kdo si vedl záznamy dat a analyzoval je, stačilo, aby měl k dispozici jen tyto statistiky. Když mu do modelu přibyla další pozorování (vzrostlo n), stačilo tyto statistiky doplnit o nové údaje a nebylo třeba přepočítávat celou soustavu rovnic znovu. To však byla v tehdejší době tak velká výhoda, že se z metody nejmenších čtverců stal základní kámen v regresní analýze. V současnosti však už díky počítačové vybavenosti není tato vlastnost takovou výhodou, i když je třeba říct, že se stále preferují odhady, které lze vypočítat jednodušší cestou.

1.3 Vlastnosti OLS

Základním teoretickým výstupem klasické regrese je následující věta, která udává podmínky, při jejichž splnění je odhad metodou OLS nejlepší mezi všemi nestrannými odhady.[6] Pro úplnost uveďme definice jednotlivých vlastností. Označme $\hat{\beta}^{(n)}$ (některý) odhad vektoru regresních koeficientů β^0 pořízených na základě dat $(Y, \mathbf{X})$.

Definice 1 *Odhad $\hat{\beta}$ nazveme* nestranný, *pokud platí* $\mathbb{E}\hat{\beta} = \beta^0$.

Definice 2 *Řekneme, že odhad $\hat{\beta}$ je* konzistentní, *pokud platí pro každé $\varepsilon > 0, \beta^0 \in \mathbb{R}^p$*

$$\lim_{n\to\infty} P(|\hat{\beta} - \beta^0| < \varepsilon) = 1. \tag{1.6}$$

[5] Matice $\mathbf{X}^T Y$ je typu $(p \times 1)$.

[6] Udává také podmínky, kdy je metoda OLS nejlepší mezi všemi *lineárními* nestrannými odhady, ale o tomto (drastickém) omezení jsem se již zmiňoval výše v textu.

Definice 3 *Řekneme, že odhad* $\hat{\beta}$ *je* nejlepší *mezi všemi odhady, pokud pro všechny ostatní odhady* $\hat{\mathbf{b}}^{(n)} \in \mathbb{R}^p$ *a pro všechna* $n \in \mathbb{N}$ *platí*

$$\mathrm{var}\hat{\beta}^{(n)} \leq \mathrm{var}\hat{\mathbf{b}}^{(n)}. \tag{1.7}$$

Věta 1 *Nechť* $\{\varepsilon_i\}_{i=1}^{\infty}$ *je posloupnost náhodných veličin, pro které platí:*

$$\mathbb{E}\varepsilon_i = 0 \quad a \quad \mathbb{E}\varepsilon_i\varepsilon_j = \delta_{ij}\sigma^2, kde \quad \begin{cases} \delta_{ij} = 0 & \Leftrightarrow j \neq i \\ \delta_{ij} = 1 & \Leftrightarrow j = i \end{cases} \quad \forall i, j \in \{1, \ldots, n\} \tag{1.8}$$

Pak $\hat{\beta}^{(OLS,n)}$ *je nejlepší nestranný lineární odhad.*

- *Pokud navíc platí*

$$(\mathbf{X}^T\mathbf{X}) = \mathcal{O}(n), \quad (\mathbf{X}^T\mathbf{X})^{-1} = \mathcal{O}\left(\frac{1}{n}\right) \; a \; \varepsilon_i \; jsou \; nezávislé, \tag{1.9}$$

pak $\hat{\beta}^{(OLS,n)}$ *je konzistentní.*

- *Pokud navíc platí*

$$\lim_{n\to\infty} \frac{1}{n}\left(\mathbf{X}^T\mathbf{X}\right) = \mathbf{Q} \; je \; regulární \; matice, \tag{1.10}$$

pak $\mathfrak{L}(\sqrt{n}(\hat{\beta}^{(OLS,n)} - \beta^0)) \to_{n\to\infty} N(0, \Sigma)$, *kde* $\Sigma = cov(\sqrt{n}(\hat{\beta} - \beta^0)) = \sigma^2 Q^{-1}$.

- *Pokud navíc platí*

$$\mathfrak{L}(\varepsilon_i) = N(0, \sigma^2), \tag{1.11}$$

pak $\hat{\beta}^{(OLS,n)}$ *je nejlepší nestranný odhad mezi všemi odhady (tedy i nelineárními).*

Důkaz viz. Víšek (1997).

Jak už jsme zmínili výše, tato věta nám jasně definuje, za jakých podmínek je nejvhodnější použití metody OLS. To by však neznamenalo mnoho, protože jako každá matematická věta i tato nám neříká nic v případě, kdy nejsou splněny všechny předpoklady. Proto bylo pro klasickou regresi vymyšleno mnoho testů - obecně nástrojů, které testují splnění požadovaných předpokladů a také nabízejí řešení, pokud tyto předpoklady splněny nejsou. Jedná se například o testy na normalitu reziduí, testy homoskedasticity,

testy multikolinearity, Whiteova kovarianční matice (opravuje výsledky při porušení homoskedasticity). Říkejme jim doprovodné nástroje regrese. Bez těchto nástrojů by byla možnost používání regrese buďto velmi omezená, nebo by nebyla jistota správnosti jejich výsledků.

Celou tuto kapitolu můžeme shrnout do následujícího tvrzení. Existuje celá řada možností, jak zkonstruovat odhad parametru. Abychom uměli rozhodnout, který z nich je nejlepší, definujeme si následující seznam vlastností, které by měl tento odhad splňovat:[7]

- nestrannost,
- $(\sqrt{n})$ konzistence,
- eficience,
- asymptotická normalita,
- existence doprovodných nástrojů regrese.

[7] Víšek, (2000b).

Kapitola 2

Robustní metody odhadu

Jak jsme již zmínili v předcházející kapitole, klasická metoda nejmenších čtverců je velmi efektivní při splnění velmi silných podmínek, avšak ztrácí velmi rychle svou kvalitu, pokud tyto podmínky nejsou splněny. Proto se už v polovině 70. let minulého století začíná rozvíjet nová oblast ekonometrie, která se snaží objevit nové možnosti, jak odhadovat parametry v reálnějších (horších) podmínkách. V následující kapitole se ve stručnosti zmíníme o dalších typech odhadů, které budou v určitém slova smyslu robustními. Definujeme nové vlastnosti odhadů, které nám dají informaci o míře robustnosti.

Než se pustíme do analýzy metody LWS[1], kterou zavedeme později v této kapitole, zmínímě alespoň ve stručnosti další metody odhadů, které této metodě předcházely. Heuristika klasických čtverců vychází vlastně ze statistického průměru, protože součet členů posloupnosti (pokud ho navíc vydělíme počtem členů této posloupnosti) je statistikou průměru. Víme již, že klasický průměr není vůbec robustním odhadem střední hodnoty, dokonce stačí pouze jediné dost odlehlé pozorování, aby vzdálilo průměr od střední hodnoty nad všechny meze. Proto se jako jeden z prvních robustních odhadů regresních koeficientů objevil tzv. *repeated median*, který však byl pouhým teoretickým konstruktem a pokud je nám známo, nebyl nikdy prakticky aplikován. Jeho hlavním přínosem bylo, že byl založen na mediánu (narozdíl od LS, které pracují s obyčejným průměrem). Tento postup byl záhy využit v další robustní metodě regresních modelů,

[1]LWS - značí the Least Weighted Squares.

kterou je metoda minimalizace mediánu čtverců reziduí (*the Least Median Squares* - LMS). Další výzkum v této oblasti však objevil slabinu metody LMS. Zatím nejnadějnějším modelem, kterým se budeme zabývat v této práci, je metoda nejmenších vážených čtverců (LWS), kterou navrhl J.Á. Víšek (2000b).

Ještě před tím je však třeba definovat nějaké užitečné míry robustnosti. My použijeme vlastnosti, které zavedl v 70. letech Frank R. Hampel. Popíšeme zde jeho přístup založený na infinitezimálním počtu - influenční funkci.[2]

2.1 Chyby v datech

2.1.1 Outliery a Leverage pointy

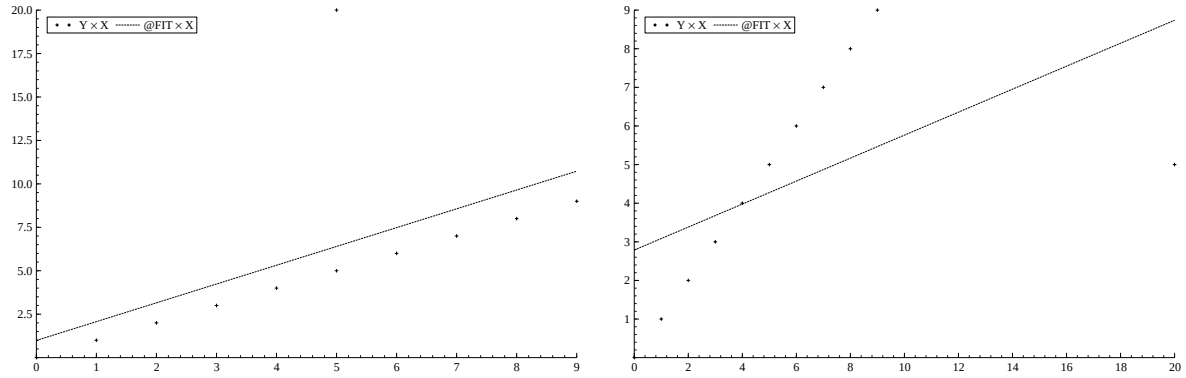
Nejčastější formou porušení klasických předpokladů jsou chyby v datech, na která aplikujeme regresní analýzu. Jedná se o tzv. odlehlá pozorování. Ty mohou být zapříčiněny mnoha faktory např. špatným přepisem čísla a jeho desetinné čárky. V dalším textu budeme rozlišovat mezi odlehlým pozorováním ve směru X (chyba v datech regresorů) a pozorováním ve směru Y (chyba v závislé proměnné). První z předchozích dvou označíme *leverage point*, druhý druh nazvěme *outlier*. Oba dva mají vliv na výsledek klasické regrese, avšak je nutno podotknout, že zatímco outliery nemusí mít až tak velký vliv, o to nebezpečnější mohou být leverage pointy. K vysvětlení tohoto tvrzení použijeme obrázek 2.1.

Z grafu je vidět, že v našem případě outlier ovlivnil pouze hodnotu interceptu a zachoval správný sklon regresní přímky. Na druhé straně leverage point změnil nejen hodnotu interceptu, ale i hodnotu koeficietu udávajícího sklon. Je třeba podotknout, že toto je pouze akademický případ. V reálném světě i outliery mohou ovlivnit sklon regresní nadroviny. Důvodem velkého vlivu leverage pointů na klasickou regresi je ten fakt, že metoda OLS spočívá na minimalizaci součtu reziduí, což je vzdálenost bodu od regresní přímky měřená podle osy y.[3] Jinými slovy klasická metoda OLS zveličuje

[2] Hampel, (1986).

[3] Ve vícerozměrném případě by tomu bylo obdobně, pouze zaměníme regresní přímku za regresní nadrovinu.

Obrázek 2.1: Outlier x Leverage Point



důležitost X-ových hodnot na úkor Y-ových.

2.2 Hamplův program

Tuto sekci zahájíme definicí influenční funkce (IF), ze které potom v dalších částech textu vyvodíme vhodné vlastnosti, kterými budeme popisovat robustní metody odhadu. V roce 1974 Frank Hampel přestavil svou teorii influenční funkce, která byla jakousi obdobou Taylorova rozvoje funkce v bodě. Avšak namísto obyčejné funkce vystupuje v Hamplově IF funkcionál (statistika) T, která je odhadem nějakého parametru. Dále zmíníme vlastnosti, kterými lze porovnávat metody odhadů. Mezi ně patří Gross Error Sensitivity, Local Shift Sensitivity, Rejection Point a Breakdown Point. Nakonec popíšeme robustní metodu, kterou navrhl Hampel jako východisko z "krize".

2.2.1 Influenční funkce

Základním kamenem Hamplova programu a jeho přínosem do robustní analýzy je nástroj, který označil influenční funkce. Ústřední myšlenkou byl následující vztah: pokud máme vektor dat (řekněme posloupnost $\{x_i\}_{i=1}^{n}$), pak existuje empirická distribuční

funkce, která je dána vztahem

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i \leq x}, \text{ kde } I \text{ je indikátor stavu, který nabývá hodnot } 0, 1. \tag{2.1}$$

Statistika T se dá vlastně popsat jako funkcionál (funkce, která přiřazuje proměnným - funkcím - reálné číslo) této empirické distribuční funkce. Neboli $T(\{x_i\}_{i=1}^n) = T(F_n)$. Tento empirický vztah v mnoha případech má i svůj teoretický protějšek, jinými slovy $T(x) = T(F(x))$, kde $F(x)$ je teoretická distribuční funkce veličiny x.[4] Nyní už můžeme uvést následující definici, ze které je na první pohled patrná podobnost s klasickou definicí derivace funkce.[5] IF pak udává, jak se změní hodnota statistiky (odhadu parametru), když se nepatrně (limitně) změní zdrojová data.

Definice 4 *Influenční funkce statistiky T v bodě F je dána vztahem*

$$IF(x; T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t} \tag{2.2}$$

pro všechna x, kde limita existuje.[6]

Z influenční funkce posléze Hampel vyvozuje nové vlastnosti robustních odhadů, které si na následujících řádcích zavedeme a ve stručnosti popíšeme.

2.2.2 Vlastnosti vycházející z IF

Hamplova myšlenka byla následující: aby byl odhad dobrý, měl by být imunní vůči chybám v datech. Tyto chyby mohou být vícero druhů, jedním z nich je malá část velmi odlehlých pozorování. Připomeňme si vliv jednoho odlehlého pozorování na statistiku průměru, který je odhadem střední hodnoty. Jedno dostatečně odlehlé pozorování může ovlivnit hodnotu odhadu přes všechny meze. Proto Hampel zavádí následující vlastnost:

Definice 5 Gross error sensitivity *je dána vztahem*

$$\gamma^*(T, F) = \sup_{x \in \mathbb{R}} |IF(x, F, T)| \tag{2.3}$$

[4] Viz Hampel (1997), str.82-83.

[5] Viz např. John (1997).

[6] Viz. Hampel (1997), str. 84.

Řečeno slovy, GES udává největší možný vliv "nejhoršího možného" pozorování x z distribuční funkce F na hodnotu statistiky T. Z toho jednoznačně vyplývá, že cílem pro dobrý odhad bude, aby měl pokud možno co nejnižší nebo alespoň konečný GES.

Druhým pólem je reakce na velmi malé pohybování s daty. Bylo by dobré, kdyby odhad, který použijeme, málo reagoval na malé změny ve velkém počtu dat, neboli aby nebyl senzitivní, např. na velké množství zaokrouhlení dat. Tato vlastnost nám mimo jiné říká, jak se odhad chová při nepřesnostech v záznamu dat (ve fyzice tak zvaná chyba měření).

Definice 6 Local shift sensitivity *je dána vztahem*

$$\lambda^* = \sup_{x,y \in \mathbb{R}} \left| \frac{IF(x, F, T) - IF(y, F, T)}{x - y} \right| \tag{2.4}$$

Ze samotného názvu vyplývá, že se jedná o citlivost na lokální posun (local shift) většího množství dat.

Definice 7 Rejection point *je dán vztahem*

$$\rho^* = \inf_{r \in \mathbb{R}} \{ r > 0; \forall x \in \mathbb{R}^p,\ |x| > r :\ IF(x, T, F) = 0 \} \tag{2.5}$$

Tato vlastnost vyplývá z potřeby zcela vyjmout velmi odlehlá pozorování před výpočtem odhadů. Ve světle influenční funkce se jedná o ta data, která mají nulový vliv na hodnotu statistiky, neboli jde o takovou část křivky IF, která má nulovou hodnotu. Hodnota ρ^* pak udává nejmenší velikost x, která už nemá vliv na hodnotu statistiky. Pokud taková není, říkáme, že $\rho^* = +\infty$ (což plyne z definice infima).

Poslední vlastností je Breakdown Point, neboli bod zlomu, jehož matematickou definici zde nebudeme uvádět[7], neboť je technicky komplikovaná a navíc používá matematické pojmy přesahující rámec této práce. Co však uvedeme, je slovní definice, která mnohem jasněji a velmi jednoduše vysvětlí její podstatu.

Definice 8 Breakdown point *značíme ϵ^* a považujeme za něj nejmenší počet odlehlých pozorování, která dokáží pozměnit hodnotu statistiky přes všechny meze, vydělený počtem všech pozorování.*

[7] Zájemce ji může nalézt např. v textu Hampel (1997), str. 97

Jinými slovy jde o nejmenší poměr špatných a dobrých dat, který už stačí k totálnímu znehodnocení odhadu. Z této definice jednoznačně vyplývá, že hodnota ρ^* může nabývat hodnoty od 0% do 50%. Pro vyšší míru kontaminace se z většiny, kterou tvoří kontaminovaná data, "stávají" čistá data a z čistých dat, kterých je méně se "stávají" odlehlá pozorování. Výše uvedené uvozovky naznačují, že se nejedná o transformaci skutečného modelu, který stojí za vygenerovanými daty, avšak naznačují, že je nemožné rozpoznat, která data jsou ta pravá (podporující generující model).

Shrneme-li výše uvedené požadavky na robustní odhad, dostaneme následující seznam[8]:

- přiměřeně nízká gross-error sensitivity,
- malá local shift sensitivity,
- konečný rejection point,
- rozumně vysoký breakdown point.

Samozřejmě stále platí požadavky, které jsme uvedli na konci první kapitoly. Tento seznam ještě není úplný, a proto toho, kdo má zájem dozvědět se více, odkazujeme na text Víšek (2000b), kde jsou podrobně rozebrány i další vlastnosti robustních odhadů.

2.3 The Least Median Squares, The Least Trimmed Squares

Ještě před tím, než se dostaneme k analýze metody LWS, která bude podrobně rozebrána v následující kapitole, je třeba uvést dvě důležité robustní metody odhadu regresních koeficientů, které přímo předcházely návrhu metody LWS a jsou v nich vyvinuty postupy, na kterých staví i metoda LWS.

Obě tyto metody představil v krátkém intervalu v roce 1983 Peter Rousseeuw . Časově první se objevila metoda the Least Median of Squares, tak se na ni podívejme nejdříve.

[8] Víšek (2000b), str. 12.

2.3.1 LMS - definice a vlastnosti

Ačkoliv tato metoda nebyla první, která v oblasti regresní analýzy použila robustní statistiku mediánu namísto klasické sumy čtverců reziduí, stala se první metodou, která se začala v praxi používat. Tou metodou, která získala zmíněné prvenství, byla čistě teoretická metoda zvaná *repeated median*. Pokud je nám známo, nebyla nikdy použita v praxi, avšak její přínos je jednoznačný - otevřela alternativní přístup k odhadům regesních parametrů. Do té doby se výlučně používala operace součtu, až v této době se začalo experimentovat i s jiným operátorem. Myšlenka byla následující: podle nových požadavků na robustní odhady se hledal takový odhad, který bude mít vysoký breakdown point, nejlépe 50%.[9] A protože medián má tuto vlastnost při odhadu střední hodnoty, definovala se metoda LMS následovně (Hampel (1997), str. 330)

Definice 9 *Odhad metodou the Least median of Squares se rozumí:*

$$\hat{\beta}^{(LMS,n)} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\ \text{med}\ \{r_i^2(\beta)\},\ i \in \{1, \ldots, n\}. \tag{2.6}$$

Prakticky okamžitě byla však tato definice zobecněna na následující:

$$\hat{\beta}^{(LMS,n,h)} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\ r_{(h)}^2(\beta) \tag{2.7}$$

kde výraz $r_{(h)}^2$ značí h-tou pořádkovou statistiku[10] čtverců reziduí. Jinými slovy seřadíme čtverce reziduí tak, aby platilo:

$$r_{(1)}^2 \leq r_{(2)}^2 \leq \ldots \leq r_{(i)}^2 \leq \ldots \leq r_{(n)}^2 \tag{2.8}$$

Nutno podotknout, že tento odhad nemá (jako mnoho nových robustních odhadů) jednoduchou formuli pro výpočet, jako tomu je u metody OLS. Přesto Rousseeuw a Leroy brzy poskytli i program, který byl schopen zpracovat data touto metodou.

Z dobrých vlastností, které tato metoda poskytovala, uveďme vysoký breakdown point - asymptoticky dosahoval horní hodnoty 50%. Avšak, jak se brzy ukázalo, měl i docela podstatné nedostatky, které vedly k nahrazení této metody jinými. Jednalo

[9] Až později se ukázalo, že tento hon za vysokým ϵ^* nebyl tak docela výhodný.

[10] Lze si povšimnout, že pro $h = \frac{n}{2}$ se tyto dvě definice shodují.

se především o malou eficienci. Zatímco metoda OLS byla $\sqrt{n}$-konzistentní, ukázalo se, že LMS je pouze $\sqrt[3]{n}$-konzistentní. To znamenalo, že zatímco pro stejně vydatný odhad metodou OLS stačilo například 100 pozorování, metoda LMS vyžadovala 1000 pozorování. Detailnější rozbor čtenář nalezne v textu Víšek (1994).

2.3.2 LTS - definice a vlastnosti

Druhou výše zmíněnou metodou je odhad nazvaný the Least trimmed squares, překládaný do češtiny jako nejmenší usekané čtverce. Jak již název napovídá, bude tato metoda ořezávat některá data. Avšak oproti klasickému TLS (Trimmed the Least squares)[11], kdy jsou data osekávána podle nějakého externího pravidla, je v metodě LTS použit vnitřní algoritmus pro ořezávání.

Definice 10 *Odhadem metodou the Least trimmed squares rozumíme*

$$\hat{\beta}^{(LTS,n,h)} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{h} r^2_{(i)}(\beta), \tag{2.9}$$

kde $h \in \langle \frac{n}{2}; n\rangle$ *je vhodně zvolený parametr ořezávání.*

Z výše uvedené definice je patrné, že postup, kterým se získává tento odhad, má mnoho společného s předchozí metodou LMS. Tak jako předchozí odhad, i tento pracuje s pořádkovými statistikami, ze kterých si ale vybírá pouze h nejmenších podle nějakého interního pravidla. Existují studie[12], které ukazují, že odhad metodou LTS je dokonce $\sqrt{n}$-konzistentní, což neplatilo pro výše uvedený odhad LMS. V tomto textu lze také nalézt důkaz tohoto tvrzení a podmínky konzistence. Než budeme pokračovat dále, dodejme, že optimální hodnota parametru ořezávání je $h = [\frac{n}{2}] + [\frac{p+1}{2}]$, kde operace $[a]$ udává celou část čísla a. Při této hodnotě parametru h dosahuje asymptoticky odhad metodou LTS magické hranice $\epsilon^* = 50\%$.

[11] Ano, pořadí slov v názvu metody naznačuje rozdíl mezi těmito dvěma metodami.

[12] Např. Víšek (2000b).

2.3.3 LTS - algoritmus výpočtu

Jak již bylo řečeno, nedílnou součástí těchto nových metod odhadů je i fakt, že jsou formulovány ve formě extremální úlohy. Narozdíl od klasických nejmenších čtverců, které mají "příjemnou" formulku pro výpočet odhadů koeficientů, není u metody nejmenších usekaných čtverců řešení tak jednoduše dostupné. Abychom dostali hledané řešení extremální úlohy (2.9), musíme buďto vyzkoušet všechny možné kombinace, nebo aplikovat něaký vhodný algoritmus výpočtu. V analýze, kterou provedeme na konci této práce, budeme pracovat s daty o velikosti $n = 100$. Pokud bychom chtěli v tomto případě počítat odhady metodou LTS způsobem porovnání všech možných kombinací, museli bychom provést minimálně $\binom{n=100}{h=50} = 10^{29}$ výpočtů. To však není ani v dnešní době vysoce výkonných procesorů časově dostupné.

Proto se musíme uchýlit k nějakému vhodnému algoritmu, který poskytuje přibližné řešení. Uvedeme strukturu algoritmu[13], který byl otestován, zda vhodně aproximuje řešení extremální úlohy (2.9).

1. Náhodně vybereme $p + 1$ pozorování a najdeme regresní nadrovinu.
2. Vypočítáme rezidua pro všechna pozorování vzhledem k této nadrovině.
3. Vezmeme h nejmenších pozorování a zapíšeme sumu čtverců reziduí.
4. Pokud není tato suma menší než předchozí hodnota, přejdeme na 6
5. Aplikací nejmenších čtverců na h zvolených pozorování najdeme novou regresní nadrovinu a přejdeme na 2.
6. Pokud byl stejný odhad nalezen už poněkolikáté (např. 20x), ukončíme proces, pokud ne, přejdeme na 1.

Netrvalo dlouho a objevila se i u metody LTS slabina, která vyžadovala nápravu. Obecně přijímaný názor, že odhad s vysokým breakdown pointem je obrněný proti špatným datům, se ukázal jako mylný v tom směru, že odhad byl sice odolný vůči odlehlým

[13] Viz. Víšek (2000a).

pozorováním, avšak nebyl dostatečně citlivý na malé změny v datech uvnitř oblaku dat. Tento efekt se objevil při zkoumání dat, ve kterých bylo chybně zaznamenáno jedno měření. Po jeho korekci se model přepočítal a k úžasu statistiků se hodnoty odhadů razantně změnily, i když korekce na jediném pozorování byla minimální.

Důvod pro toto chování objasňuje ve své práci Víšek (2000b). Autor rovněž nabízí východisko z tohoto problému. Představuje další metodu odhadu, kterou uvedeme v následující kapitole.

Kapitola 3

The Least Weighted Squares

Metoda nejmenších vážených čtverců vychází z požadavku, aby robustní regresní metody neignorovaly odlehlá data úplně. Tím se připravují o kus informace, kterou každé takto useklé pozorování nese. Uveďme jeden akademický případ, na kterém popíšeme konkrétně výše zmíněnou vlastnost metody LTS. Pokud máme data, která vynesena do grafu vykazují známky dvou různých populací, může se stát, že "vhodným" výběrem dat pouze z jednoho podsouboru bude výsledný odhad koeficientů regrese popisovat pouze tato data a ignorovat druhou, nevybranou populaci. To může být někdy dobře, někdy špatně.To musí ten, kdo zpracovává data, celkově posoudit (např. s pomocí některého grafického editoru).

Proto zavedeme nový odhad

Definice 11 *Odhadem metodou the Least Weighted Squares rozumíme*

$$\beta^{(LWS,h,w)} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} w_i r^2_{(i)}(\beta), \tag{3.1}$$

kde $\{w_i\}_{i=1}^n$ *je nerostoucí posloupnost čísel, pro která platí* $w_1 = 1, w_i \geq 0$.

Pro důkaz existence řešení odkazujeme čtenáře na text Víšek (2006), který se zabývá i důkazem konzistence metody LWS.

Opět můžeme velmi lehce nahlédnout podobnost s předchozími odhady. Co je v tomto případě nové, je vektor vah, kterým násobíme pořádkové statistiky čtverců reziduí (order statistics of squared residuals), a tím můžeme potlačit vliv dané pořádkové statistiky

na součet v (3.1), přidat nebo ubrat na důležitosti. Při podrobnějším zkoumání lze objevit, že pro specifickou volbu vah se odhad metodou LWS stává odhadem metodou LTS. Stačí pouze položit $w_i = 1$ pro $i \in \{1, \ldots, h\}$ a $w_i = 0$ pro $i \in \{h+1, \ldots, n\}$. Lze tedy říci, že odhad metodou LTS je speciálním případem odhadu metody LWS.

Přidáním vah jako parametru do extremální úlohy vlastně vkládáme do procesu další ladící prvek, který nám dovoluje ovládat proces hledání optimálního řešení úlohy. Jinými slovy máme možnost ovlivnit průběh výběru těch dat, která "více" a která "méně" determinují model, a tím i ovlivnit konečný tvar modelu.

Ještě než se pustíme do simulační analýzy, uveďme pozměněný algoritmus pro výpočet LWS.

3.1 LWS - Algoritmus výpočtu

1. Náhodně vybereme $p + 1$ pozorování a najdeme regresní nadrovinu.

2. Zvolíme vektor vah $w \in \mathbb{R}^n$.

3. Vypočítáme rezidua pro všechna pozorování vzhledem k této nadrovině.

4. Seřadíme rezidua podle velikosti a vynásobíme je příslušnými váhami a zaznamenáme součet $S = \sum_{i=1}^{n} r_{(i)}^2 w_i$.

5. Pokud není tato suma menší než předchozí hodnota, přejdeme na 7.

6. Aplikací vážených nejmenších čtverců (WLS) na všechna pozorování seřazená dle velikosti čtverců reziduí najdeme novou regresní nadrovinu a přejdeme na 3. Jinými slovy, původní data seřadíme dle velikosti čtverců reziduí (což jsme vlastně už udělali v bodě 4.) a spočteme $\hat{\beta}^{(WLS)} = (\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{Y}})$, kde $\tilde{\mathbf{X}}$ a $\tilde{\mathbf{Y}}$ jsou "přerovnaná data" a $\mathbf{W} = \mathrm{diag}\{w_1, w_2, \ldots, w_n\}$ (diag{ } značí diagonální matici s prvky na diagonále $w_1, w_2, \ldots, w_n$).

7. Pokud byl stejný odhad nalezen už poněkolikáté (např. 20x), ukonči proces, pokud ne, přejdi na 1.

3.2 Simulační analýza metody LWS

V poslední části této práce provedeme sérii různých simulací, ve kterých budeme srovnávat výsledky různých metod odhadů s výsledky pomocí metody LWS. Veškeré výpočty a simulace budeme provádět pomocí programu MATLAB. Tento program jsme zvolili na doporučení, neboť má (velmi) dobrý generátor náhodných čísel. Nutno dodat, že požadavek, aby náhodná čísla byla opravdu náhodná, je nutný pro jakoukoliv analýzu.

Nejprve provedeme sérii výpočtů s daty, která budou mít silně kontaminované hodnoty regresorů. Jak jsme si řekli v druhé kapitole, říkáme takovým datům leverage pointy. Také jsme zmínili, že jejich vliv na odhad metodou LS je veliký, a proto jsou pro tuto metodu velkým úskalím. Data pro analýzu vygenerujeme následujícím postupem:

1. Zvolíme základní počet pozorování našich dat ($n = 100$) a počet regresorů včetně konstanty ($p = 5$). Volba těchto hodnot má následující význam: 100 pozorování je na jedné straně obecně dostatečné množství pro regsesní analýzu, na straně druhé se s takovým množstvím dat setkáme v reálných regresních analýzách. Vyšší počet regresorů znamená, že je již obtížnější (bez grafického programu skoro nemožné) odhalit odlehlá pozorování pouhým pohledem.

2. Vytvoříme náhodnou matici regresorů $\mathbf{X}$ typu ($n = 100 \times p = 5$), kde jednotlivé prvky matice jsou normálně rozdělené s nulovou střední hodnotou ($\mu_{X_{ij}} = 0$) a rozptylem $\sigma^2_{X_{ij}} = 2$. Zvolením normálního rozdělení dosáhneme oblakovitého tvaru dat, který bude mít směrem ke středu hustší strukturu.

3. Podobně vygenerujeme vektor disturbancí ϵ, normálně rozdělených ($\mu_\epsilon = 0, \sigma^2_\epsilon = 1$) náhodných veličin.[1]

[1] Je dobré si uvědomit, že v případě, kdy rozptyl generovaných disturbancí bude větší, než rozptyl dat generované matice X, pak bude model z větší části determinovám disturbancemi a nikoliv daty. Proto v této simulaci volíme $\sigma^2_{X_{ij}} > \sigma^2_\epsilon$.

4. Zkonstruujeme model, který bude generovat data závislé proměnné:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i \tag{3.2}$$

   Konkrétní hodnoty koeficientů β by neměly mít vliv výsledky analýzy, proto jsme zvolili $\beta = [1, 2, 3, 4, 5]$ s ohledem na přehlednost v generovaných grafech.

5. Vygenerujeme vektor $\mathbf{Y}$ pomocí modelu z 3 a dat z 1, 2.

6. Vytvoříme vektor vah $\mathbf{w}$ tak, že prvních 53 dat dostane velké váhy, u dalších 10 budou váhy strmě klesat a posledních 30 dat dostane nulovou váhu - bude z regrese vyjmuto. Tyto parametry vycházejí z faktu, že metoda LTS dosahuje pro hodnotu $h = \frac{n+p+1}{2}$ asymptoticky nejvyšší možný breakdown point $\epsilon^* = 50\%$. Vzhledem k podobnosti obou metod proto volíme p

7. Postupně budeme kontaminovat prvních 1 - 50 řádků matice $\mathbf{X}$ tak, že každou kontaminovanou hodnotu vynásobíme 10. Nekontaminujeme 1. sloupec, který obsahuje konstantní jedničku, ostatní sloupce kontaminujeme.

8. Pro každý stupeň kontaminace spočítáme odhad metodou LS a metodou LWS.

9. Vyneseme graf závislosti hodnoty odhadů (pro obě metody) na stupni kontaminace dat.

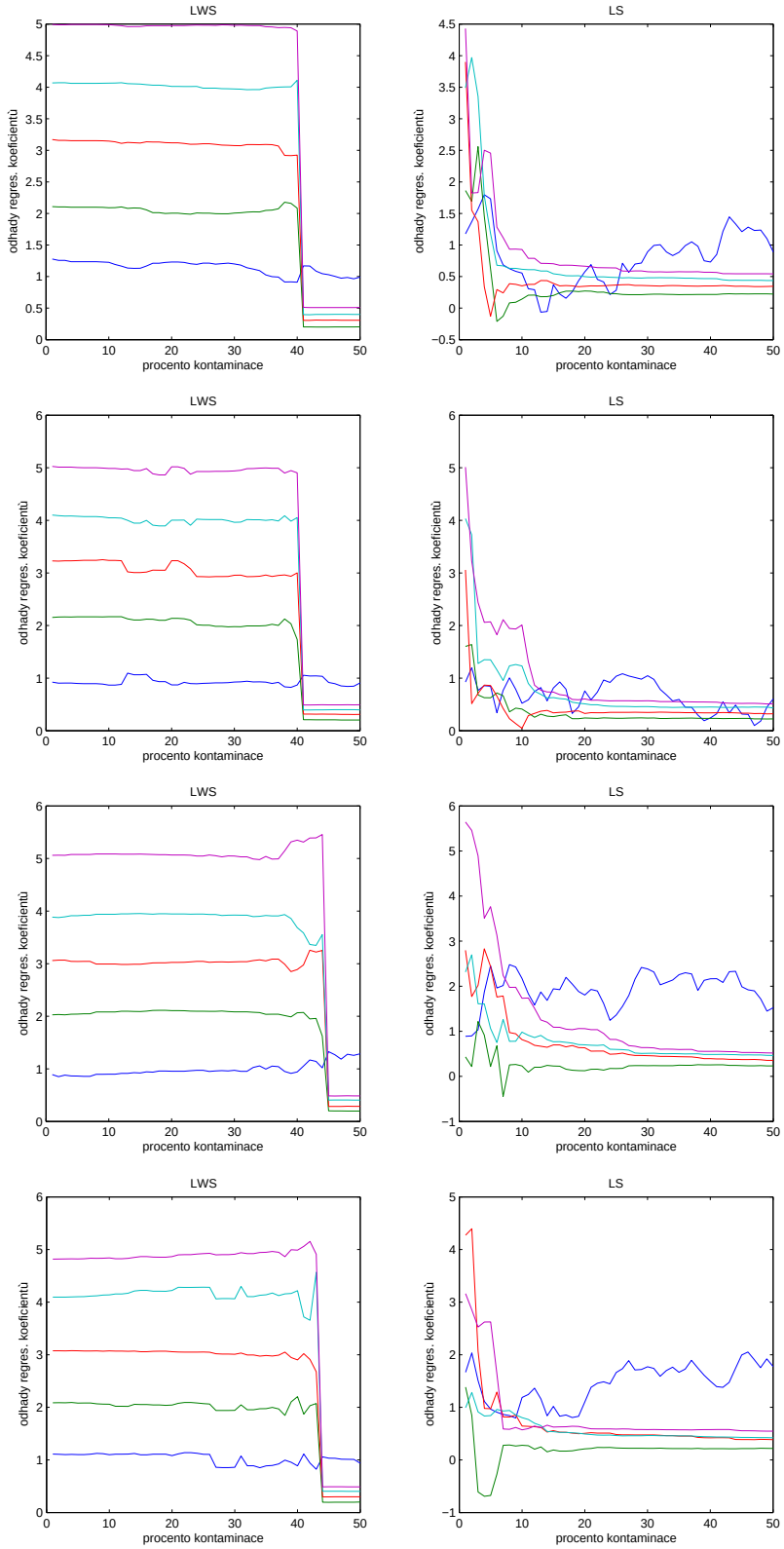
3.3 Interpretace výsledků

Po skončení naprogramované simulace jsme z programu MATLAB získali následující grafický výstup: na obrázku 3.1 jsou zaneseny hodnoty odhadů regresních koeficientů pro jednotlivé metody odhadu (vlevo je metoda the Least Weighted Squares, napravo klasická metoda nejmenších čtverců) při kontaminaci matice $\mathbf{X}$.

Na první pohled je z výsledných grafů patrný markantní rozdíl mezi jednotlivými metodami odhadu. Zatímco na pravých grafech odhad metodou nejmenších čtverců příkře klesá od své skutečné hodnoty již při nízké kontaminaci[2] (a to ve všech čtyřech

[2]Rozsah svislé osy y může být pro každý graf jiná. To platí zejména u pravých grafů, kde je použita metoda nejmenších čtverců.

Obrázek 3.1: Hodnoty odhadů v závislosti na míře kontaminace matice X



případech), metoda LWS, která je zaznamenána na levých grafech, udržuje hodnotu odhadů v relativně úzkém pásmu kolem její skutečné hodnoty a začíná se hroutit až při kontaminaci přesahující 40%.[3] Připomeňme ještě, že skutečné hodnoty regresních koeficientů jsou $\beta = (1, 2, 3, 4, 5)^T$, a tedy na grafech by se objevily jako vodorovné konstantní čáry ve výšce 1 (respektive 2, 3, 4 a 5).

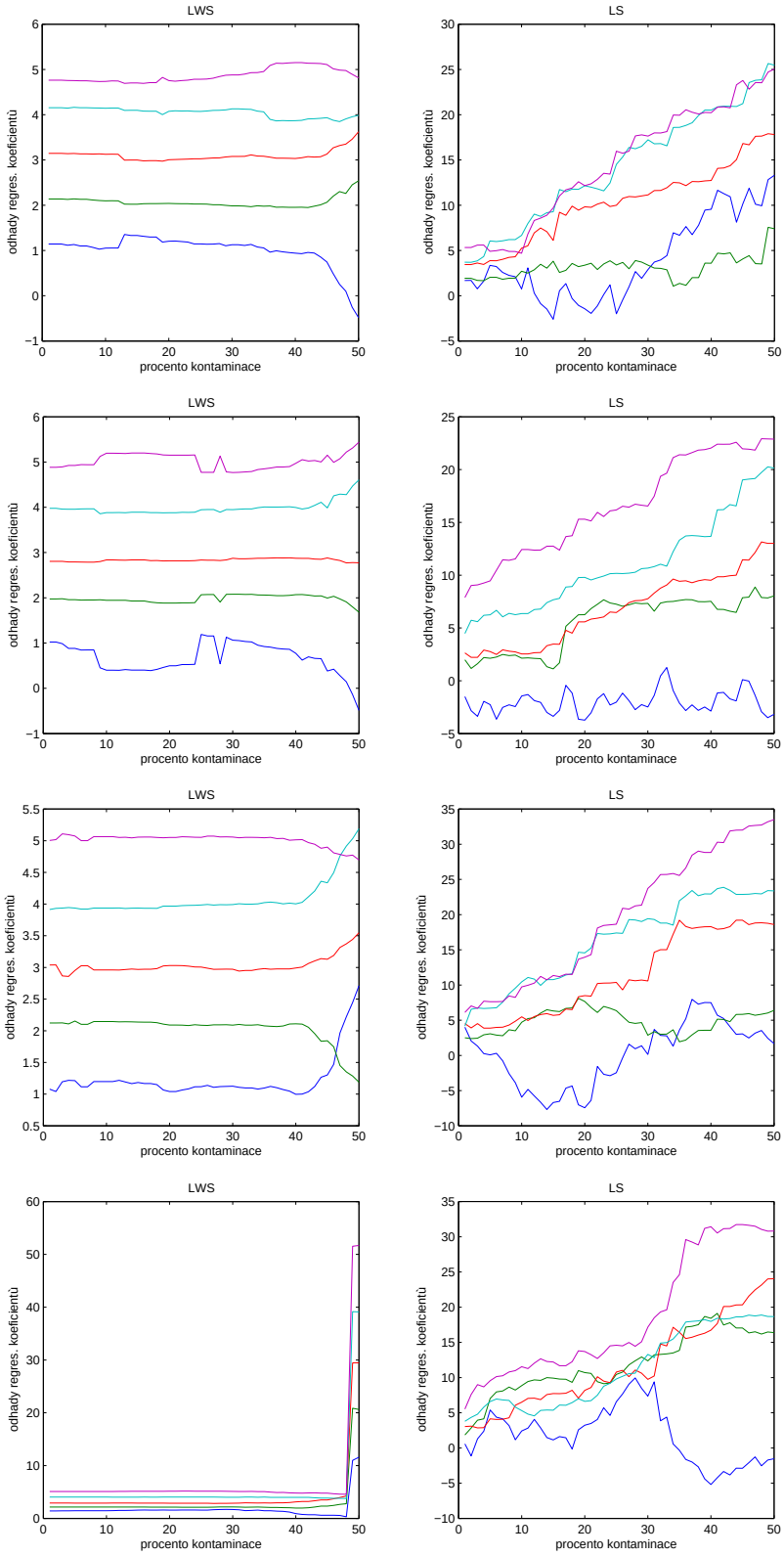
Tento výsledek dokazuje náchylnost klasické metody nejmenších čtverců na odlehlá pozorování. Lze dokázat, že jedno dostatečně vzdálené pozorování může způsobit naprosté zhroucení odhadů, neboli klasické metody mají asymptoticky nulový breakdown point. Co se týče metody LWS, tak zde toto nemusí nutně platit. Námi zvolenými váhami jsme ze všech dat do výpočtu zahrnuli jen 63 nejlépe popisujících daný model. Ostatním byla přiřazena nulová váha, tedy byly z regrese vyjmuty.

Je však pravdou, že pokud bychom zvolili vektor vah se všemi členy nenulovými, pak by metoda LWS měla (s použitím těchto vah) nulový breakdown point, podobně jako výše zmíněná metoda nejmenších čtverců. Jinými sovy by teoreticky stačilo jedno dostatečně odlehlé pozorování, aby zhroutilo odhad. Např. i s váhou $w_n = 0,0001$ by mělo pozorování $X_n = (1, 10000, 10000, 10000, 10000)$ srovnatelný vliv jako obyčejné pozorování s jednotkovou váhou.

Podobně jako na předchozím obrázku, jsou na obrázku 3.2 zaznamenány výsledky při použití dat s kontaminací vektoru závislé proměnné $\mathbf{Y}$. I v tomoto případě odhady metodou nejmenších čtverců (umístěných v pravém sloupci grafů) s rostoucí kontaminací rostou a vzdalují se od svých skutečných hodnot. Rozdíl mezi "explozí" v tomto případě a "implozí" v případě předchozí simulace se objasní, uvědomíme-li si, jakým způsobem byla data kontaminována. Pokud kontaminujeme hodnoty matice $\mathbf{X}$ a nezvýšíme přímo úměrně i hodnoty vysvětlované proměnné $\mathbf{Y}$, pak se samozřejmě odhady regresních koeficientů zmenšují s rostoucí mírou kontaminace. A přesně naopak tomu je, pokud zvětšíme pouze hodnoty vektoru $\mathbf{Y}$ a ponecháme hodnoty matice $\mathbf{X}$ nezměněny. Nakonec poznamenejme, že pokud bychom vynásobením deseti kontaminovali 100% dat

[3] Podotkněme, že 40% kontaminace je blízko hranice, kdy se kontaminovaná data stávají většinou, a kdy je obtížné rozeznat, která data určují model, a která jsou kontaminací.

Obrázek 3.2: Hodnoty odhadů v závislosti na míře kontaminace vektoru Y



matice $\mathbf{X}$, pak by se odhady koeficientů pohybovaly okolo desetiny skutečných hodnot. Naopak při 100% kontaminaci vektoru $\mathbf{Y}$ se hodnoty odhadů budou pohybovat okolo desetinásobku skutečné hodnoty.[4]

3.4 Závěr

Otázkou však zůstává, jakou metodu zvolit při hledání regresních koeficientů. Jestli se spolehnout na klasické požazavky vlastností dat, které jsme nastínili v první kapitole tohoto textu, nebo raději rovnou předpokládat jakékoliv porušení těchto předpokladů. Připomeňme, že restrikce na normálně rozdělené disturbance je velmi drastická. Odpověď může být překvapující, na druhou stranu však velmi jednoduchá. Vždy, když to bude možné, aplikujeme více metod odhadů - klasických i robustních. Pokud se shodují, máme více méně jistotu, že nalezený model opravdu odpovídá modelu, dle kterého byla data naměřena. Pokud se však výsledky klasických a robustních metod významně liší, je to pro nás výstražné znamení, že data nejsou úplně v pořádku. Pak nezbývá, než se zastavit a velmi důkladně tato data prozkoumat, zdali například neobsahují více subpopulací, nebo zkusit odhalit nějakou vnitřní strukturu mezi daty, která by měla na svědomí odlišnost výsledků. Každopádně výše popsaná metoda the Least Weighted Squares, ani žádná další nepřináší automatický lék na poškozená data. Vždy bude třeba, aby statistik či ekonometr, který data zpracovává, ovládal veškeré nástroje do té míry, aby mohl najít příčiny nesrovnalostí jednotlivých metod odhadu.

To, co jsme však konečnou simulací dokázali, je fakt, že pokud máme dostatečně velký soubor dat, tak pokud zvolíme váhy více robustní (pokud si můžeme dovolit většímu počtu dat přiřadít nízké, nebo nulové váhy), pak bude odhad metodou LWS robustní, co se týče míry kontaminace dat. Neboli, ať je kontaminace dat malá či velká, odhad bude blízko skutečné hodnoty.

Zbýva jen zodpovědět otázku, kterou si možná čtenář klade. Proč tedy stále nejpoužívanější metodou zůstává klasická metoda nejmenších čtverců? Možná to je proto,

[4] Toto je dobře vidět v levém spodním grafu na obrázku 3.2, kde při kontaminaci 45% a vyšší se hodnoty odhadů posunuly na hladinu desetinásobku skutečných hodnot koeficientů.

že je stále nejjednodušší co se týče postupu výpočtu odhadů, avšak je tu ještě jeden velmi důležitý aspekt. Jak jsme již zmínili výše v textu, pro metodu nejmenších čtverců bylo navrhnuto velké množství doprovodných testů a procedur, které rozšiřovaly její použití i do oblastí, kde byly některé klasické předpoklady porušeny. Mnohé z toho, co je k dispozici pro metodu nejmenších čtverců, stále není dostupné pro robustní metody. Až budou navrženy tyto doprovodné programy i pro robustní odhady, rozšíří se implementace těchto odhadů do statistických programů, pak věříme, že se robustní metody stanou více používanějšími, než je tomu doposud.

Příloha A

Matlab - zdrojový kód

Následuje struktura programu napsaného v programu MATLAB, kterým jsme generovali a analyzovali data.

```
% Vygeneruje data pomocí funkce CreateData(),v kterých postupně
% kontaminuje prvních 1 - 50 procent dat (buď X nebo Y). Pro tyto
% data najde odhady regresních koeficientů metodou LS a LWS.
% Nakonec vykreslí graf znázorňující hodnotu odhadů v závislosti
% na míře kontaminace.
% k=1  ... kontaminace X
% k=2  ... kontaminace Y

function[X,Y,w,beta_contaminated_matrix_lws,beta_contaminated_matrix_ls]=
SimulationRepeated(k)
%vytvoření dat
[w,X,Y,beta_real,n,p,e] = CreateData();

for f=1:50;

    %ukazatel procent proběhlých operací
    mesage= [num2str(f*2) '% výpočtů hotovo.'];
    disp(mesage)

    %kontaminace dat
    if k==1;
        [X_contamined] = ContamineX(f,X);
        [Y_contamined] = Y;
    elseif k==2;
        [X_contamined] = X;
        [Y_contamined] = ContamineY(f,Y);
    else
```

```
        disp('Chyba v zadání - parametr k má nepřípustnou hodnotu.');
        break
    end

        betals_mean=zeros(p,1);
        betalws_mean=zeros(p,1);
        betalss=zeros(p,1);
        betalwss=zeros(p,1);

        %aplikace LWS a LS na data
        [betals,betalws]=Simulation(X_contamined,Y_contamined,w);

    for j=1:p
        beta_contaminated_matrix_lws(f,j)=betalws(j);
        beta_contaminated_matrix_ls(f,j)=betals(j);
    end

end

%generování grafu
subplot(1,2,1); plot(beta_contaminated_matrix_lws);title('LWS');
xlabel('procento kontaminace');ylabel('odhady regres. koeficientů');
subplot(1,2,2); plot(beta_contaminated_matrix_ls);title('LS');
xlabel('procento kontaminace');ylabel('odhady regres. koeficientů');

------------------------------------------------------------------------
% Vygeneruje vstupní data do simulace.
%
% n         ... počet pozorování(řádků matice X)
% p         ... počet regresorů (sloupců matice X)
% w         ... vektor vah
% X         ... matice regresorů
% beta_real ... vektor koeficientů modelu
% e         ... vektor disturbancí
% Y         ... vektor závislé proměnné generovaný modelem:
%                     Y = X * beta_real + e

function [w,X,Y,beta_real,n,p,e] = CreateData()

%základní vstupní hodnoty
beta_real=[1;2;3;4;5];
n=100;
p=5;

%parametry váhové funkce
```

```
h=53;
g=63;
k=200;

for i=1:n;
    for j=1:p;
        e(i,1)=0;
        X(i,j)=0;
        Y(i,1)=0;
    end
end

%generování dat
w=Weights(h,g,n,k);
X=RandMatrX(n,p);
e=Genere(n);
Y=GenerY(beta_real,X,e,n,p);

-------------------------------------------------------------------------
% Vygeneruje vektor vah, který je určen parametry h,g,n,k.
% h ... počet pozorování, které získají mírně klesající váhu
% g ... počet pozorování, které získají prudce klesající váhu
% n ... počet pozorování (řádků matice X)
% k ... koeficient určující sklon funkce vah.
% Pro větší k je sklon prvních (h) členů posloupnosti w menší a sklon
% členů (g-h) je stmější. (n-g) členům přiřadí nulovou váhu.

function [w]=Weights(h,g,n,k)
L=(1-h/k)/(g-h+1);
for i=1:n;   w(i,1)=0; end
f=g-h;
for i=1:f;  w(i+h,1)=1-h/k-i*L; end
for i=1:h;   w(i,1)=1-i/k; end

-------------------------------------------------------------------------
% Funkce generující matici X typu (n x p)
% náhodných hodnot z rozdělení N(0,2)
% n   ... počet řádků matice X
% p   ... počet sloupců matice X

function [X]=RandMatrX(n,p)

X=random('norm',0,2,[n,p]);

-------------------------------------------------------------------------
```

```
% Generátor vektoru disturbancí z rozdělení N(0,1)
% n      - počet řádků matice X

function [e]=Genere(n)

e=random('normal',0,1,[n,1]);

-------------------------------------------------------------------------
% Generátor závislé proměnné Y pomocí modelu
%       Y(i)=b(1)*X(i,1)+b(2)*X(i,2)+...+b(p)*X(i,p)+e(i)
%
% beta_real   ... vektor hodnot koeficientů
% n           ... počet pozorování
% p           ... počet sloupců matice X
% e           ... vektor disturbancí

function[Y]=GenerY(beta_real,X,e,n,p)

Y=X*beta_real+e;

-------------------------------------------------------------------------
% Kontaminace matice regresorů X. První sloupec matice X (vektor
% jedniček) nekontaminujeme. Typ kontaminace je posun desetinné čárky
% o jedno místo do prava (vynásobení hodnoty deseti).
%   q   ... počet kontaminovaných regresorů   1<q<p+1
%   z   ... počet pozorování, která kontaminujeme  0<z<n
%   f   ... procento kontaminace dat (f = z / n * 100)

function [X_contamined] = ContamineX(f,X)

q=size(X,2);
z=round(size(X,1)*f/100);
t=20;
X_contamined=X;

for i=1:z;
    for j=2:q;
        X_contamined(i,j)=X(i,j)*10;
    end
end

-------------------------------------------------------------------------
% Kontaminace části dat vektoru Y. Typ kontaminace je posunutí desetinné
% čárky o jedno místo do prava (neboli vynásobení hodnoty deseti).
% f   ... procentu kontaminace
```

```
% z   ... počet kontaminovaných pozorování

function [Y_contamined] = ContamineY(f,Y)

z=round(size(Y,1)*f/100);
Y_contamined=Y;

for i=1:z;
    Y_contamined(i,1)=Y(i,1)*10;
end

-------------------------------------------------------------------------
% Provedení funkcí LS a LWSiter na kontaminovaná data.
% w  ... vektor vah

function[betals,betalws]=Simulation(X_contamined,Y_contamined,w)

betals=LS(X_contamined,Y_contamined);
betalws=LWSiter(X_contamined,Y_contamined,w);

-------------------------------------------------------------------------
% Odhad regresních koeficientů metodou nejmenších čtverců.
%
%   X  ... matice regresorů
%   Y  ... vektor dat závislé proměnné
%   b  ... vektor odhadů parametrů
%   mf ... součet čtverců reziduí

function [b,mf]=LS(X,Y)

b=inv(X'*X)*X'*Y;
mf=sum((Y-X*b).^2);

-------------------------------------------------------------------------
% the Least Weighted Squares -
% iterační algoritmus s předvolenými parametry
%   X - matice regresorů
%   Y - pozorování
%   w - vektor váh

function [b,mf,r,rr,rrr]=LWSiter(X,Y,w)

[b,mf,r,rr,rrr]=LWSiterGeneral(X,Y,w,10000,100,20,0);

-------------------------------------------------------------------------
```

```
% Least Weighted Squares - iterační algoritmus s volitelnými parametry
%   X - matice regresorů
%   Y - pozorování
%   w - vektor váh
%   MaxOpak - maximalni pocet opakovani (napr. 10000)
%   MaxIter - max. pocet iteraci pri jednom opakovani (napr. 100)
%   MaxMatch - po kolika stejnych cilovych modelech skoncit (napr. 20)
%   Info - vypisovat informace
%

function
[b,mf,por,rr,rrr]=LWSiterGeneral(X,Y,w,MaxOpak,MaxIter,MaxMatch,Info)

  % pocet pozorovani a dimenze
  n=size(X,1);
  p=size(X,2);
  Info=0;
  % Hlavni cyklus
  b=zeros(p,1);
 %   por=(n,1);
  mf=Inf;
  Match=0;
  for i=1:MaxOpak
    % projed jednu iteraci
    [bp,mfp,por,rr,rrr]=OneIter(X,Y,w,MaxIter,Info);
    % stejny model jako doposud nejlepsi
    if (mf==mfp) & (b==bp);
      Match=Match+1;
      if Info; disp('match'); end
    end
    if Match==MaxMatch; break; end
    % lepsi model nez doposud nejlepsi
    if mfp<mf
      b=bp;
      mf=mfp;
      Match=0;
      if Info; disp(b); disp(mf); end
    end
  end

function [b,mf,por,rr,rrr]=OneIter(X,Y,w,MaxIter,Info)

  % pocet pozorovani a dimenze
  n=size(X,1);
  p=size(X,2);
```

```
  % vyber startovnich p pozorovani
  SI=find(randperm(n)<=p);
  SX=X(SI,:);
  SY=Y(SI,:);
  if cond(SX)>1e10
    if Info; disp('Bad starting basis'); end
    mf=inf;
    b=zeros(p,1);
    return
  end
  % spocti nadrovinu
  b=inv(SX)*SY;
  % predpocitej vahovou matici
  W=diag(w);
  W=W(find(w),find(w));
  % spust cyklus
  rI=zeros(n,1);

  while 1
    % uloz pozorovani vybrana v predchozim kroku
    rIo=rI;
    % spocti rezidua a serad je
    r=(Y-X*b).^2;
    [rs,rI]=sort(r);
    %objone=sum(rs.*w);
    % WLS na serazena pozorovani
    SX=X(rI,:);
    SY=Y(rI,:);
    SX=SX(find(w),:);
    SY=SY(find(w),:);
    b=inv(SX'*W*SX)*SX'*W*SY;
    % konevrguje ?
    if min(rIo==rI)==1; break; end
    MaxIter=MaxIter-1;
    if MaxIter==0; break; end
  end
  if Info & (MaxIter==0); disp('To many iterations'); end
  % vystup
  [r,por]=sort((Y-X*b).^2);
   rr=abs(Y-X*b);
   rr=sort(rr);
   ww=sqrt(w);
   rrr=(rr.*ww);
%disp(rrr);
  mf=sum(r.*w);
```

Literatura

[1] Anděl J. (2003): *Statistické metody*, Matfyzpress, Praha.

[2] Hampel, F. R. (1986): *The Approach based on Influence Functions*, Wiley, New York.

[3] John, O., et al. (2003): *Matematika (pokračování)*, Matfyzpress, Praha.

[4] Rousseuw, P. J. (1987): *Robust reggression and outlier detection*, Wiley, New York.

[5] Víšek, J. Á. (1994): *A cautionary note on the method of the Least Median of Square reconsidered.* Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague, 1994, s. 254-259.

[6] Víšek, J. Á. (1997): *Ekonometrie*, Karolinum, Praha.

[7] Víšek, J. Á. (2000a): *On the diversity of estimates*, Computational Statistics and Data Analysis 34, 67-89.

[8] Víšek, J. Á. (2000b): *Regression with high breakdown point*, In *Robust '2000: Proceedings of the 11th Conference on Robust Statistics*, ISBN 80-7015-792-5, s. 324-356.

[9] Víšek, J. Á. (2002): *Modern processing data*, In: *Data Analysis 2002/II, Modern Statistical Methods - Modelling, Regression, Classification and Data Mining*, organised by TRYLOBYTE, Bohdaneč ed. K. Kupka, ISBN 80-239-0204-0, s. 131-167.

[10] Víšek, J. Á. (2006): *Consistency of the Least weighted squares*, [elektronická pošta].

UNIVERSITAS CAROLINA PRAGENSIS
založena 1348

Univerzita Karlova v Praze
Fakulta sociálních věd
Institut ekonomických studií

Opletalova 26
110 00 Praha 1
TEL: 222 112 330,305
TEL/FAX: 222 112 304
E-mail: ies@mbox.fsv.cuni.cz
http://ies.fsv.cuni.cz

Akademický rok 2006/2007

TEZE BAKALÁŘSKÉ PRÁCE

Student:	Jaroslav Hlávka
Obor:	Ekonomie
Konzultant:	Doc. RNDr. Jan Ámos Víšek, CSc.

Garant studijního programu Vám dle zákona č. 111/1998 Sb. o vysokých školách a Studijního a zkušebního řádu UK v Praze určuje následující bakalářskou práci

Předpokládaný název BP:

Robustní ekonometrické metody – simulační analýza the Least Weighted Squares

Charakteristika tématu, současný stav poznání, případné zvláštní metody zpracování tématu:

Ve své bakalářské práci se hodlám zabývat teorií robustních odhadů, zvláštní pozornost bude věnována metodě the Least Weighted Squares.

Pomocí programu MATLAB aplikujeme metodu LWS a klasickou metodu LS na vygenerovaná data a porovnáme výsledky.

Struktura BP:

Úvod
Klasické metody - teorie
Robustní metody - teorie
Aplikace robustní metody na konkrétní data (simulace)
Analýza výsledků
Závěr

Seznam základních pramenů a odborné literatury:

Hampel, F. R. (1986): Robust Statistics. The Approach based on Influence Functions. Wiley, New York
Rousseeuw, P. J. (1987): Robust reggression and outlier detection. Wiley, New York
Víšek, J. Á. (2000): Regression with high breakdown point, In Robust '2000: Proceedings of the 11th Conference on Robust Statistics, ISBN 80-7015-792-5, s. 324-356.

Datum zadání:	říjen 2006
Termín odevzdání:	červen 2007

Podpisy konzultanta a studenta:

V Praze dne 31.10. 2006