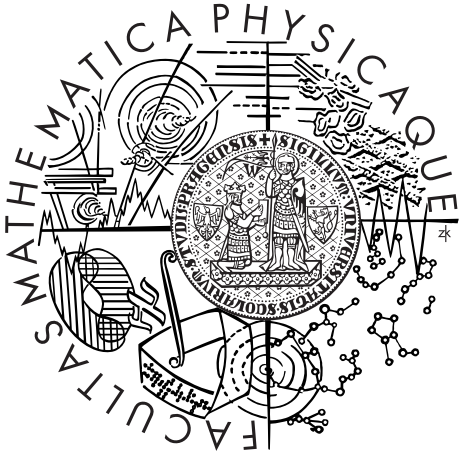
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Michal Gerthofer

Alternativy nejmenších čtverců

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Michal Pešta, Ph.D.
Studijní program: Matematika
Studijní obor: Finanční matematika

Praha 2013

Rád by som sa poďakoval predovšetkým vedúcemu mojej práce RNDr. Michalovi Peštovi, Ph.D. za množstvo času a ochotu, s ktorou sa mi venoval, za všetky jeho rady, pripomienky a poskytnuté materiály.
Ďalej by som sa chcel poďakovať svojim rodičom za podporu počas písania tejto práce a počas celého štúdia.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 23.5.2013 Michal Gerthofer

Název práce: Alternativy nejmenších čtverců

Autor: Michal Gerthofer

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V přiložené práci se věnujeme lineárním regresním modelům založeným na metodě nejmenších čtverců. Ty jsou rozebrány ve dvou skupinách. První se zaměřuje na tři základní postupy rozdělené podle výskytu chyb v proměnných. Tradičním způsobem zabývajícím se chybou jen na straně závislé proměnné je základní metoda nejmenších čtverců (OLS). Opačným případem je metoda datově nejmenších čtverců (DLS), která připouští chyby jen ve vysvětlujících proměnných. Následně se soustředíme na ortogonální regresi (TLS) minimalizující čtverce chyb obou proměnných. Nakonec upřeme pozornost na další skupinu metod s vysokým bodem selhání. Tyto metody se věnují významnosti jednotlivých pozorování (metoda nejmenších vážených čtverců) a eliminaci odlehlých pozorování (metoda useknutých nejmenších čtverců). Hlavním cílem práce je popsat a porovnat tyto modely, jejich předpoklady, charakteristiky a vlastnosti odhadů a demonstrovat je na reálných datech.

Klíčová slova: základní metoda nejmenších čtverců, metoda dátovo nejmenších čtverců, orgtogonální metoda nejmenších čtverců, metoda nejmenších vážených čtverců, metoda useknutých nejmenších čtverců

Title: Least Squares Alternatives

Author: Michal Gerthofer

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In the present thesis we deal with the linear regression models based on least squares. These methods are discussed in two groups. The first one focuses on three primary aproaches devided by occurrence of errors in variables. The traditional approach penalizes only the misfit in the dependent variable part and is called the ordinary least squares (OLS). An opposite case to the OLS is represented by the data least squares (DLS), which allow corrections only in the explanatory variables. Consecutively, we concentrate ourselves on the total least squares approach (TLS) minimizing the squares of errors in the values of both dependent and independent variables. Finally, we give attention to next group of methods whit high breakdown point, which deal with significance of the individual observations (least weighted squares) and elimination of outlying observations (least trimmed squares). The main purpose of this work is to describe and compare these models, their assumptions, characteristics, properties of estimates and show them on real data.

Keywords: ordinary least squares, data least squares, total least squares, least weighted squares, least trimmed squares

Názov práce: Alternatívy najmenších štvorcov

Autor: Michal Gerthofer

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V predloženej práci sa venujeme lineárnym regresným modelom založeným na metóde najmenších štvorcov. Tie sú rozoberané v dvoch skupinách. Prvá sa zameriava na tri základné postupy rozdelené podľa výskytu chýb v premenných. Tradičným prístupom zaoberajúcim sa chybou len na strane závislej premennej je základná metóda najmenších štvorcov. Opačným prípadom je metóda dátovo najmenších štvorcov (DLS), ktorá pripúšťa chyby len vo vysvetľujúcich premenných. Následne sa sústredíme na ortogonálnu regresiu (TLS), minimalizujúcu štvorce chýb oboch premenných. Nakoniec upriamime pozornosť na ďalšiu skupinu metód s vysokým bodom zlyhania, ktoré sa venujú významnosti jednotlivých pozorovaní (metoda najmenších vážených štvorcov) a eliminácii odľahlých pozorovaní (metóda useknutých najmenších štvorcov). Hlavným cieľom práce je popísať a porovnať tieto modely, ich predpoklady, charakteristiky, vlastnosti odhadov a demonštrovať ich na reálnych dátach.

Kľúčové slová: základná metóda najmenších štvorcov, metóda dátovo najmenších štvorcov, ortogonálna metóda nejmenších štvorcov, metóda vážených najmenších štvorcov, metóda useknutých najmenších štvorcov

Obsah

Úvod

Lineárne modely zohrávajú dôležitú úlohu v moderných štatistických a ekonometrických modeloch. Pomocou týchto modelov sme schopní, ak nie v úplnom rozsahu tak aspoň čiastočne, aproximovať veľké množstvo funkcií. Klasická metóda najmenších štvorcov patrí medzi základné postupy lineárnej regresie, od ktorej sa postupne odvodzovali jej alternatívy. Prostredníctvom týchto metód sa snažíme, čo možno najlepšie, zachytiť závislosť pozorovaných veličín respektíve trend.

Cieľom práce je prehľadne priblížiť základné alternatívy metódy najmenších štvorcov, ich predpoklady, konštrukcie, simulácie a vlastnosti odhadov. Ďalej na základe predpokladov odporučiť metódu vhodnú pre daný typ dát.

Metódy budeme skúmať najskôr podľa výskytu chýb, či už na strane závislej, nezávislej premennej alebo na oboch stranách zároveň. Pozornosť ďalej venujeme postupom, pri ktorých budú mať pozorovania rôzny vplyv a ako túto skutočnosť zahrnúť do výpočtu. V poslednej časti práce budeme skúmať dopad odľahlých pozorovaní na náš odhad a riešenie tohto problému.

Prvá kapitola je zameraná na lineárny model lineárnej regresie a odhad neznámeho parametra pomocou základnej metódy najmenších štvorcov, kde sa chyba vyskytuje na strane vysvetľovanej premennej. Uvedieme si postup hľadania tohto odhadu a jeho dôležité vlastnosti.

Druhá kapitola sa venuje singulárnemu rozkladu matice nevyhnutnému pre následný popis metód a vlastností odhadov. Postupne sa dopracujeme k dátovo najmenším štvorcom, kde sa chyba vyskytuje na strane vysvetľujúcej premennej. Ako poslednou z tejto skupiny metód sa oboznámime s ortogonálnou metódou najmenších štvorcov, kde sa chyba vyskytuje na strane vysvetľovanej premennej. Zhrnutie týchto metód je ilustrované na regresných priamkach a analýze reálnych dát.

V závere práce je spracovaná metóda najmenších vážených štvorcov, v ktorej každé pozorovanie má rôzny význam (váhu) a metóda useknutých štvorcov, ktorá sa snaží eliminovať vplyv odľahlých pozorovaní a odvodenie vlatností týchto odhadov.

Pre výber témy bakalárskej práce som sa rozhodol na základe širokého využitia týchto metód, či už v štatistike, ekonometrii alebo poisťovníctve.

Obrázky používané v práci sú vyrobené pomocou softvéru R.

1. Metóda najmenších štvorcov

1.1 Motivácia

Pri pozorovaní určitých javov sa snažíme zachytiť ich správanie. Chceme, čo najlepšie odhadnúť správanie pozorovanej veličiny na základe napozorovaných dát. Napríklad sledujeme, ako sa s rastúcou výškou mení váha. Označme výšku ako premnnú X_i a váhu ako sledovanú premennú Y_i u i-teho pozorovaného človeka. Ako môžeme vidieť na obrázku 1.1, s rastúcou výškou stúpa aj váha. Z toho môžeme usúdiť, že medzi výškou a váhou existuje určitá závislosť.

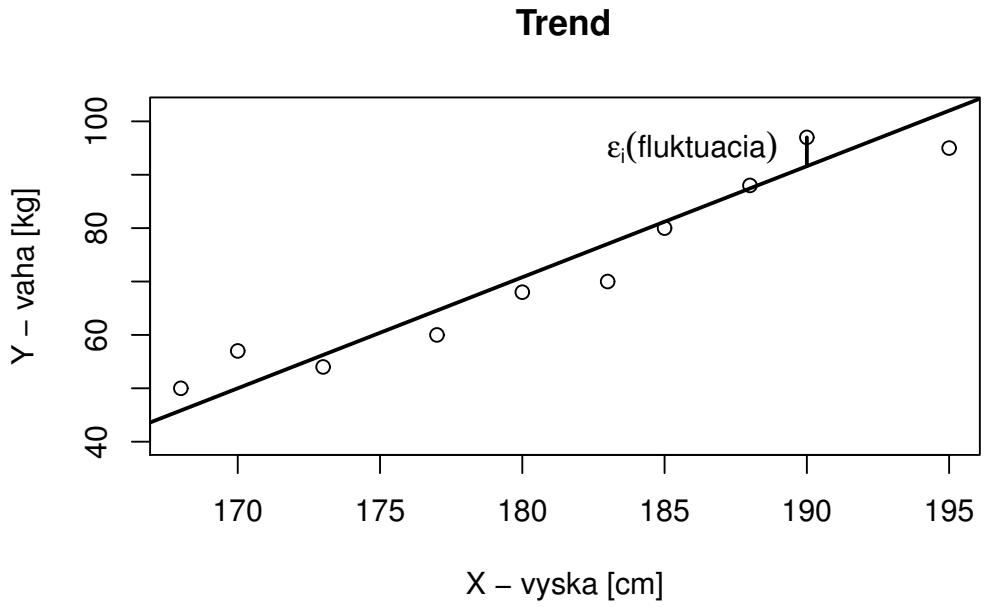
Zachytením takýchto javov sa zaoberá regresia. V našom zjednodušenom prípade ide o klasický model lineárnej regresie

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,$$

pomocou ktorého sa snažíme zachytiť daný trend tzn. závislosť veličiny Y na nezávislej premennej X vzťahom

$$Y = \beta_1 + \beta_2 X,$$

kde ε_i je nevysvetlená fluktuácia s nulovou strednou hodnotou, kladným rozptylom, pre $i \in \{1,2,\ldots,n\}$ a počet pozorovaní n. Našou úlohou je čo najlepšie odhadnúť parametre β_1, $\beta_2 \in \mathbb{R}$ tak, aby daná rovnica, čo možno najlepšie vystihovala všetky namerané dáta, aby sme ďalej pomocou nej mohli odhadnúť váhu človeka len so znalosťou jeho výšky. V práci budeme jednorozmerné premenné označovať veľkým písmom a vektory, matice veľkým tučným písmom.



Obrázok 1.1: Trend Y - váhy [kg] v závislosti na premennej X - výške [cm].

1.2 Základné pojmy lineárnej regresie

V regresii sa všeobecne snažíme docieliť pomocou rôznych metód to, aby vyrovnané dáta očistené od náhodných fluktuácií vystihovali trend. Skúmame podmienené rozdelenie veličiny Y_i, ak poznáme vektor $\mathbf{X}_{i,\bullet}{}^\top$ o p zložkách, kde $\mathbf{X}_{i,\bullet}$ je i-ty riadok matice $\mathbf{X} \in \mathbb{R}^{n\times p}$. Zaujíma nás, ktoré komponenty vektoru $\mathbf{X}_{i,\bullet}{}^\top$ a akým spôsobom ovplyvňujú strednú hodnotu $\mathsf{E}\,Y$ a ako už bolo spomínané, chceme predpovedať Y pre ľubovoľnú hodnotu nezávislej premennej.

Dáta pozostávajú z n nezávislých pozorovaní vektorov $(Y_i,\mathbf{X}_{i,\bullet})^\top$, kde matica $\mathbf{X}$ má menej stĺpcov ako je počet pozorovaní teda riadkov tzn. $p < n$. V prípade, že by stĺpce matice $\mathbf{X}$ boli lineárne závislé, niektorý z neznámych regresorov by bol zbytočný, pretože jeho závislosť by sa dala popísať zvyšnými regresormi.

Hodnoty $X_{i,j}$ pre $j \in \{1,\ldots,p\}$, z vektoru $\mathbf{X}_{i,\bullet}{}^\top$ nemusia byť vždy hodnoty napozorovaných náhodných vektorov, ale aj ich transformácie, tzn. $\mathbf{X}_{i,\bullet} = f(\mathbf{Z}_{i,\bullet})$, kde $\mathbf{Z}_{i,\bullet}$ sú hodnoty napozorovaných dát.

Existujú dva spôsoby, ako môžeme daný problém riešiť, pričom v jednom z nich je vektor $\mathbf{X}_{i,\bullet}{}^\top$ chápaný ako náhodná veličina a v druhom ako vektor konštant. Oba spôsoby však vedú k rovnakému výsledku. V tomto texte sa zaoberáme spôsobom, kde $\mathbf{X}_{i,\bullet}{}^\top$ je vektor vopred známych hodnôt.

1.2.1 Lineárny model lineárnej regresie

Definícia 1. *Hovoríme, že dáta $(Y_i,\boldsymbol{X}_{i,\bullet})$ pre $i \in \{1,2,\ldots,n\}$ spĺňajú model lineárnej regresie, keď platí*

$$Y_i = \beta_1 X_{i,1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \tag{1.1}$$

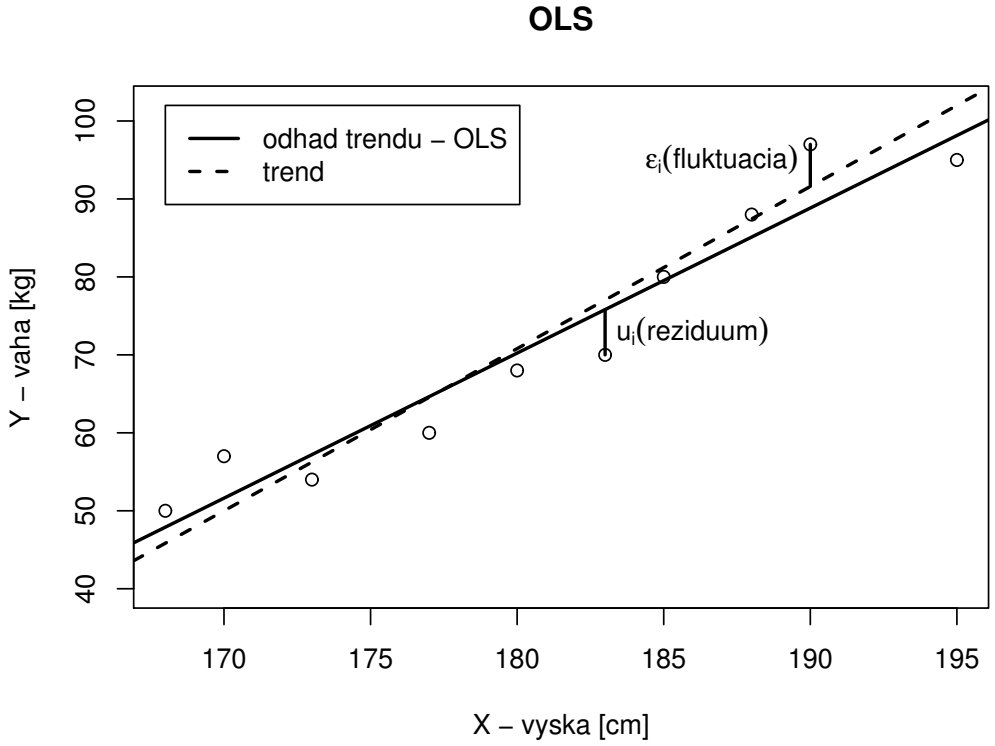
kde $\beta = (\beta_1,\ldots,\beta_p)^\top$ je vektor neznámych parametrov, ktoré inak nazývame regresnými koeficientami a $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)^\top$ je vektor obsahujúci n nezávislých náhodných veličín s nulovou strednou hodnotou a rozptylom $\sigma^2 > 0$.

Ak v modeli lineárnej regresie $p = 2$ a $X_{\bullet,1} = \mathbf{1}_n$, kde $\mathbf{1}_n = (1,\ldots,1)^\top$ je vektor n jednotiek, potom

$$Y_i = \beta_1 + \beta_2 X_2 + \varepsilon_i,$$

ktorú označujeme ako *jednoduchú lienárnu regresiu* (použitie v motivácii, viď obrázok 1.2). V prípade ak $X_{i,1} = 1$ pre $\forall\, i \in \{1,2,\ldots,n\}$, zostane z $\beta_1 X_{i,1}$ len β_1, označovaný ako absolútny člen, intercept.

V lineárnom modeli lineárnej regresie pracujeme s vektormi napozorovaných dát $(Y_i,\mathbf{X}_{i,\bullet})^\top$, kde Y_i je odozva (závislá premenná) a $\mathbf{X}_{i,\bullet} = (X_{i,1},\ldots,X_{i,n})$ je riadkový vektor regresorov (nezávislá premenná) z rovnice (1.1), kde ε_i je náhodná veličina, fluktuácia, ktorá zahŕňa všetko, čo ovplyvňuje Y_i a nie je vysvetlené regresormi $X_{i,1},\ldots,X_{i,n}$, môže zahŕňať aj chybu. Je to rozdiel medzi napozorovanou hodnotou a trendom.



Obrázok 1.2: Odhad trendu pomocou metódy OLS a ukážka rozdielu medzi reziduom u_i a fluktuáciou ε_i.

Pomocou rôznych metód, napríklad metódou najmenších štvorcov sa snažíme, aby vyrovnané dáta v tvare

$$\widehat{Y}_i = \widehat{\beta}_1 X_{i,1} + \widehat{\beta}_2 X_{i,2} + \cdots + \widehat{\beta}_p X_{i,p},$$

čo možno najlepšie vystihovali trend, kde vektor $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^\top \in \mathbb{R}^{n\times 1}$ je odhad neznámeho parametra $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

1.2.2 Alternatívny zápis lineárneho modelu lineárnej regresie

Vektorový zápis rovnice (1.1) je $Y_i = \mathbf{X}_{i,\bullet}\boldsymbol{\beta} + \varepsilon_i$, kde $\mathsf{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma^2$ a ε_i sú nezávislé, z čoho následne platí

$$\mathsf{E}\,Y_i = \mathbf{X}_{i,\bullet}\boldsymbol{\beta} \text{ a } \operatorname{var} Y_i = \sigma^2.$$

Maticovo

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.2}$$

kde $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbf{X} = \left(\mathbf{X}_{1,\bullet}{}^\top, \ldots, \mathbf{X}_{n,\bullet}{}^\top\right)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, za platnosti

$$\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0} \text{ a } \operatorname{var}\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}_n,$$

pričom $\mathbf{X}$ je regresná matica, ktorej $h(\mathbf{X}) = p < n$. Symbolom $\mathbf{I}_n$ rozumieme n-rozmernú štvorcovú maticu, ktorá má na diagonále jednotky inak nuly.

1.2.3 Interpretácia parametrov regresného modelu

Ak $X_{\bullet,1} = \mathbf{1}_n$, regresný koeficien β_1 nám v modeli jednoduchej lineárnej regresie udáva posun vo zvislom smere. Koeficient β_2 zase hovorí o koľko sa zmení stredná hodnota po zväčšení regresoru $X_{i,2}$ o jednotku, čo možno chápať ako sklon priamky. Ak sa $X_{\bullet,1} \neq \mathbf{1}_n$ ale rovná sa pozorovaným hodnotám, regresná nadrovina prechádza počiatkom súradnicovej sústavy.

V lineárnom modeli lineárnej regresie koeficient β_j, $j \in \{1, 2, \ldots, p\}$ vyjadruje zmenu strednej hodnoty $\mathsf{E}\, Y_i$ po zvýšení regresoru $X_{i,j}$ o jednotku, pri konštantných hodnotách ostatných regresorov. Zo štatistického hľadiska je model bez interceptu horšie aplikovateľný na reálne dáta, preto je vhodnejšie používať model s interceptom.

1.3 Odhady regresných koeficientov pomocou základnej metódy najmenších štvorcov

Táto metóda inak označovaná ako OLS (*anglicky Ordinary least squres*) je základnou metódou odhadovania neznámych parametrov v linenárnom modeli lineárnej regresie. Tento postup vychádza z minimalizovania súčtu druhých mocnín vertikálnej vzdialenosti vysvetľovanej premennej Y_i od regresnej nadroviny. Argument minima tohto súčtu je odhad neznámeho parametru. Snažíme sa tak, o čo najlepšie preloženie nadroviny množinou pozorovaných dát, viď obrázok 1.2 a prvá časť obrázku 2.3.

V tejto metóde budeme pre lepšiu ilustráciu na reálnych dátach pracovať s lineárnym modelom lienárnej regresie s interceptom, teda budeme mať regresnú maticu, ktorej prvý stĺpec budú samé jednotky. Postupy, výpočty a vlastnosti odhadu sú však v tejto metóde totožné, či už používame model s interceptom, alebo bez.

1.3.1 Predpoklady metódy OLS

Aby odhad parametru metódou najmenších štvorcov mal určité vlastnosti, musí model spĺňať stanovené predpoklady.

Definujme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, maticu $\mathbf{X} = \left(\mathbf{X}_{1,\bullet}{}^\top, \ldots, \mathbf{X}_{1,\bullet}{}^\top\right)^\top$, vektor neznámych parametrov $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ a náhodný vektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, ktoré splňujú rovnicu (1.1).

Predpokladajme, že stredná hodnota vektoru $\boldsymbol{\varepsilon}$ je rovná nule, jeho zložky sú navzájom nezávislé a kovariančná matica je $\sigma^2 \mathbf{I}_n$, pre $\sigma^2 > 0$. Táto vlastnosť sa nazýva homoskedasticita.

Ďalšou podmienkou je lineárna nezávislosť stĺpcov matice $\mathbf{X}$. Ak je splnená vieme, že $h(\mathbf{X}) = p$ a $p < n$, z čoho následne plynie, že matica $\mathbf{X}^\top \mathbf{X}$ je regulárna. V príklade z motivácie je táto podmienka splnená, kedže matica $\mathbf{X}$ má v prvom stĺpci samé jednotky, odpovedajúce interceptu β_1 a v druhom navzájom rôzne hodnoty.

Prepoklad nulovosti strednej hodnoty náhodného vektoru $\boldsymbol{\varepsilon}$ hovorí, že dáta vyvážene oscilujú okolo trendu, čo umožňuje vyrovnať dáta

$$\begin{aligned}\mathsf{E}\, Y_i &= \mathsf{E}\,(\beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i) = \\ &= \mathsf{E}\,(\beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}) + \mathsf{E}\,\varepsilon_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p},\end{aligned}$$

pri rozptyle

$$\operatorname{var}\mathbf{Y} = \sigma^2\mathbf{I}_n = \operatorname{var}\boldsymbol{\varepsilon}.$$

Nech $\widehat{\boldsymbol{\beta}}$ je nejaký odhad parametru $\boldsymbol{\beta}$. Potom označme $\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ odhadnuté stredné hodnoty odozvy t. j.

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_{i,2} + \cdots + \widehat{\beta}_p X_{i,p} = \mathbf{X}_{i,\bullet}\widehat{\boldsymbol{\beta}}.$$

Ako už vieme, vektor $\widehat{\boldsymbol{\beta}}$ sa snažíme určiť tak, aby euklidovský rozdiel skutočných dát $\mathbf{Y}$ a odhadnutých dát $\widehat{\mathbf{Y}}$ bol čo najmenší, tzn.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sqrt{\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_{i,\bullet}\boldsymbol{\beta})^2}.$$

1.3.2 Postup hľadania odhadu parametra – OLS

Hľadáme

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sqrt{\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_{i,\bullet}\boldsymbol{\beta})^2} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_{i,\bullet}\boldsymbol{\beta})^2 = \\ = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Postupujeme ako pri hľadaní minima funkcie, čiže danú funkciu zderivujeme a položíme ju rovnú nule. Koreň tejto funkcie je odhadnutý parameter $\widehat{\boldsymbol{\beta}}$.
Zderivujeme funkciu

$$\frac{\partial}{\partial\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = -\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X}\right]^\top = \\ = -2\left(\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right), \tag{1.3}$$

a následne $\widehat{\boldsymbol{\beta}}$ dostaneme ako riešenie p lineárnych rovnic o p neznámych, ktoré sme získali položením derivácie (1.3) rovnej nule

$$-2\left(\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{0}$$
$$\Downarrow$$
$$\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{Y}$$

Z predpokladov, že matica $\mathbf{X}$ má hodnosť p a teda matica $\mathbf{X}^\top\mathbf{X}$ je regulárna, plynie existencia inverznej matice $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ a následne existencia práve jedného riešenia sústavy rovníc

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}. \tag{1.4}$$

Odhad $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(n)$ je funkciou rozsahu pozorovaní. $\widehat{\boldsymbol{\beta}}$ je taktiež globálnym minimom, pretože funkcia $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ je konvexná v $\boldsymbol{\beta}$.

Veta 1. *Majme model lineárnej regresie v tvare* (1.1)*, v ktorom stĺpce matice* $\boldsymbol{X}$ *sú navzájom nezávislé vektory a platí homoskedasticita vektoru* ε*, potom odhad metódou najmenších štvorcov* $\widehat{\boldsymbol{\beta}}_{OLS}$ *je v tvare* (1.4)*.*

Dôkaz. Viď vyššie *Postup hľadania odhadu parametra – OLS*.

❑

Definícia 2. *Nech* $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$ *sa nazýva projekčná matica.*

Potom môžeme regresnú nadrovinu vyjadriť v tvare

$$\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{OLS} = \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

Dôležitá vlastnosť projekčnej matice $\mathbf{H}$ je inedepotentnosť, tzn. $\mathbf{HH} = \mathbf{H}$.

1.4 Vlastnosti odhadu

Medzi základné vlastnosti odhadov patrí nestrannosť a konzistencia, ale existujú aj ďalšie vlastnosti, ako je napríklad eficientnosť.

1.4.1 Nestrannosť a konzistencia odhadu – OLS

Definícia 3 (Nestrannosť)**.** *Odhad sa nazýva nestranný (nevychýlený), ak jeho stredná hodnota je rovná hodnote odhadovaného parametru:* $\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

Veta 2. $\widehat{\boldsymbol{\beta}}_{OLS}$ *je nestranný odhad parametru* $\boldsymbol{\beta}$*.*

Dôkaz.

$$\begin{aligned}\mathsf{E}\,\widehat{\boldsymbol{\beta}}_{OLS} = \mathsf{E}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\,\mathsf{E}\,\mathbf{Y} &\overset{\mathsf{E}\,\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}}{=} \\ \overset{\mathsf{E}\,\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}}{=} \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \boldsymbol{\beta},\end{aligned} \tag{1.5}$$

z čoho platí $\mathsf{E}\,\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}$

❑

O kvalite odhadu hovoria jeho asymtotické vlastnosti ako napríklad konzistencia.

Definícia 4 (Konzistencia)**.** *Odhad je konzistentný, ak pri rastúcom rozsahu výberu n jeho hodnota konverguje v pravdepodobnosti ku skutočnej hodnote parametru t. j.* $P\left(|\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}| < \varepsilon\right) \to 1$.

Veta 3. $\widehat{\boldsymbol{\beta}}_{OLS}$ *je konzistentným odhadom* $\boldsymbol{\beta}$*, čiže* $\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta}$ *pre* $n \to \infty$*.*

Dôkaz. Viď [18] podkap. 5.1.

❑

1.4.2 Ďalšie vlastnosti

Definícia 5. *Odhad $\widehat{\boldsymbol{\beta}}$ sa nazýva eficientný voči inému odhadu $\widehat{\boldsymbol{\beta}^*}$ toho istého parametra, ak nemá väčší rozptyl.*

$$var\left(\widehat{\boldsymbol{\beta}}\right) \leq var\left(\widehat{\boldsymbol{\beta}^*}\right),$$

kde $var\,\widehat{\boldsymbol{\beta}} = \sigma^2\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$.

Veta 4. *Odhad $\widehat{\boldsymbol{\beta}}_{OLS}$ je v lineárnom modeli lineárnej regresie najlepším nestranným lineárnym odhadom (anglicky Best linear unbiased estimator, tzv. BLUE-odhadom) parametra $\boldsymbol{\beta}$ (Gaussova-Markovova veta).*

Dôkaz. Viď [15] podkap. 3.4. ❑

Veta 4. hovorí, že $\widehat{\boldsymbol{\beta}}_{OLS}$ je súčasne lineárnou funkciou hodnôt vysvetľovanej premennej $Y_1, \ldots, Y_n$, nestranným odhadom parametru $\boldsymbol{\beta}$ a eficientným odhadom voči každému lineárne nestrannému odhadu parametru $\boldsymbol{\beta}$. Podrobnejší popis a ďalšie vlastnosti odhadu viď podkap. 3.3. v [4] a kap. 3.,4. a 5. v [18].

2. Alternatívy metódy najmenších štvorcov

Predtým než uvedieme jednolivé metódy, vysvetlíme pojem singulárneho rozkladu matice inak SVD (*anglicky Singular value decomposition*) a zadefinujeme Frobeniovu normu, ktoré budeme neskôr používať.

2.1 Singulárny rozklad matice – SVD

Definícia 6. *Matica Q je ortonormálna, keď platí* $Q^\top = Q^{-1}$.

Veta 5. *Nech* $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, *potom existujú ortonormálne matice* $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ *a* $\boldsymbol{V} \in \mathbb{R}^{m\times m}$, *pre ktoré platí*

$$\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Sigma} = diag\left\{\sigma_1,\ldots,\sigma_p\right\} \in \mathbb{R}^{n\times m}, \sigma_1 \geq \cdots \geq \sigma_p \geq 0, \tag{2.1}$$

kde $p = \min\{n,m\}$.

Dôkaz. Viď [9] kapitola 8.6. ❑

Matica $\boldsymbol{\Sigma}$ v singulárnom rozklade je jednoznačne určená maticou $\mathbf{A}$. Ďalej si pre maticu $\mathbf{A}$ definujeme *deliaci bod* $r \in \{1,\ldots,p\}$, ako najväčší index, pre ktorý platí $\sigma_r > 0$ teda

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0, \text{ kde } p = \min\{n,m\},$$

kde σ_i pre $i \in \{1,\ldots,p\}$ sú *simgulárne čísla* matice $\mathbf{A}$.

Ak sú matice $\mathbf{V}$ a $\mathbf{U}$ z (2.1) ortonormálne, potom $r = h(\mathbf{A})$ a maticu $\mathbf{A}$ môžeme vyjadriť v *dyadickom rozvoji*

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{U}_{\bullet,i} \mathbf{V}_{\bullet,i}{}^\top,$$

kde $\mathbf{U}_{\bullet,i}$ a $\mathbf{V}_{\bullet,i}$ sú i-te stĺpce príslušných matíc.

Definícia 7. *Označme* $\boldsymbol{U}_{\bullet,i}$ *ako ľavé a* $\boldsymbol{V}_{\bullet,i}$ *ako pravé singulárne vektory matice* $\boldsymbol{A}$. *Potom* $\boldsymbol{U}_{min}$ $(\boldsymbol{V}_{min})$ *nazývame najmenším ľavým (pravým) singulárnym vektorom, ktorý sa spája s najmenšou singulárnou hodnotou* $\sigma_{min} := \sigma_r$.

Definícia 8. *Frobeniova norma matice* $\boldsymbol{A} \equiv \left(a_{i,j}\right)_{i,j=1}^{n,m}$ *je definovaná nasledovne*

$$\|\boldsymbol{A}\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}^2} = \sqrt{tr\left(\boldsymbol{A}^\top \boldsymbol{A}\right)} = \sqrt{\sum_{i=1}^{p} \sigma_i^2} = \sqrt{\sum_{i=1}^{r} \sigma_i^2},\ p = \min\{n,m\}.$$

2.2 Typy podľa výskytu chýb

Tieto metódy by sme mohli rozdeliť do dvoch skupín. V prvej sa jednotlivé postupy zameriavajú na to, či sa fluktuácie vyskytujú na strane závislej premennej, nezávislej alebo na oboch stranách. V druhej skupine sa metódy viac zameriavajú na to, ktoré pozorovania sú pre odhad podstatnejšie a ako túto skutočnosť zohľadniť pri výpočte. Najbližšie sa budeme venovať prvej skupine sponímaných metód.

Odhadmi parametra $\boldsymbol{\beta}$ pomocou metódy OLS sme sa zaoberali v predchádzajúcej kapitove. Chyba sa tam vyskytovala vo vysvetľovnej premennej

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg \min_{\boldsymbol{\varepsilon}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\varepsilon}\|_2 \text{ za platnosti } \mathbf{Y} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}.$$

Opačným prípadom je, ak sa chyba vyskytuje vo vysvetľujúcich premenných, regresoroch

$$\widehat{\boldsymbol{\beta}}_{DLS} = \arg \min_{\boldsymbol{\Theta}\in\mathbb{R}^{n\times p}, \boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\Theta}\|_F \text{ za platnosti } \mathbf{Y} = (\mathbf{X} + \boldsymbol{\Theta})\,\boldsymbol{\beta}. \tag{2.2}$$

K odhadu parametra $\boldsymbol{\beta}$ v tomto prípade slúži metóda dátovo najmenších štvorcov inak označovaná ako DLS (*anglicky Data least squares*). Poslednou variantou je, ak sa chyba vyskytuje na oboch stranách súčasne

$$\widehat{\boldsymbol{\beta}}_{TLS} = \arg \min_{[\boldsymbol{\varepsilon},\boldsymbol{\Xi}]\in\mathbb{R}^{n\times(p+1)}, \boldsymbol{\beta}\in\mathbb{R}^p} \|\,[\boldsymbol{\varepsilon},\boldsymbol{\Xi}]\,\|_F \text{ za platnosti } \mathbf{Y} + \boldsymbol{\varepsilon} = (\mathbf{X} + \boldsymbol{\Xi})\,\boldsymbol{\beta}, \tag{2.3}$$

čo riešime pomocou ortogonálnej metódy TLS (*anglicky Total least squares*). V každej z týchto metód sa snažíme, čo najlepšie odhadnúť neznámy parameter $\boldsymbol{\beta}$ ako argument minima, v prvom prípade euklidovkej normy vektora a v ďalších dvoch Frobeniovej normy matice.

Kedže v metódach DLS a TLS sa chyby vyskytujú aj na strane nezávislej premennej, pre jednoduchosť výpočtov budeme uvažovať model lineárnej regresie bez interceptu.

2.2.1 Metóda dátovo najmenších štvorcov – DLS

Odhad pomocou metódy DLS vychádza z modifikovanej metódy TLS Abatzogloa a Mendela [1], poskytuje nám odhad parametra $\boldsymbol{\beta}$, ak vieme, že chyby sa vyskytujú len v matici vstupných dát, regresorov. Ako bolo už zmienené odhad parametru je daný argumentom minima funkcie z (2.2).

Na postavenie teórie DLS, je nutné splnenie určitých podmienok. Prvou je plná stĺpcová hodnosť matice, $h(\mathbf{X}) = p$ ako v metóde OLS a druhou je, aby vektor $\mathbf{Y}$ nebol kolmý s najmenším ľavým singulárnym vektorom matice $\mathbf{X}$.

Existencia a jednoznačnosť riešenia DLS

Veta 6. *V prípade splnenia vyššie uvedených predpokladov DLS a* (2.2) *je riešenie tvaru*

$$\widehat{\boldsymbol{\beta}}_{DLS} = \frac{\boldsymbol{Y}^\top \boldsymbol{Y}}{\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{K}_{min}} \boldsymbol{K}_{min},$$

kde $\boldsymbol{K}_{min}$ *je najmenší pravý singulárny vektor matice* $\boldsymbol{P}_{\boldsymbol{Y}}^{\perp}\boldsymbol{X}$*, kde* $\boldsymbol{P}_{\boldsymbol{Y}}^{\perp} = I - \boldsymbol{Y}\left(\boldsymbol{Y}^\top \boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^\top$ *je projekčná matica, ktorá zobrazuje vektorový priestor matice* $\boldsymbol{X}$ *na ortogonálny doplnok vektoru* $\boldsymbol{Y}$*.*

Dôkaz. Viď [10] na str. 408. ❑

Ako je možné vidieť, výsledok nemá zmysel, pokiaľ $\mathbf{Y}^\top \mathbf{X}\mathbf{K}_{min} = 0$.

Komplikácia nastane v prípade, keď sa najmenšia singulárna hodnota bude nachádzať v rozklade viackrát, čo by znamenalo, že existuje nekonečne veľa riešení. Ak chceme aj v tomto prípade nájsť jednoznačné riešenie, musí ležať v priestore určenom najmenšími pravými singulárnymi vektormi $\mathbf{K}_{min_j}$ tzn. $\boldsymbol{\beta}_{DLS} = \mathbf{K}^*_{min}\mathbf{C}$, kde $\mathbf{K}^*_{min}$ je matica, ktorej stĺpce sú vektory $\mathbf{K}_{min_j}$ a $\mathbf{C}$ je vektor konštant. Rovnako musí $\boldsymbol{\beta}_{DLS}$ spĺňať aj rovnicu $\mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta}_{DLS} = \mathbf{Y}^\top \mathbf{Y}$. Jednoznačnosť je určená minimalizačnou normou. Zostáva vyriešiť optimalizačný problém

$$\min_{\mathbf{C}} \boldsymbol{\beta}^\top \boldsymbol{\beta} = \mathbf{C}^\top {\mathbf{K}^*_{min}}^\top \mathbf{K}^*_{min}\mathbf{C}, \text{ za platnosti } \mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{Y}^\top \mathbf{Y},$$

kde použitím Lagrangeových multiplikátorov dostaneme

$$\widehat{\boldsymbol{\beta}}_{DLS} = \frac{\mathbf{Y}^\top \mathbf{Y}}{\left(\mathbf{Y}^\top \mathbf{X}\mathbf{K}^*_{min}\right)\left({\mathbf{K}^*_{min}}^\top \mathbf{X}^\top \mathbf{Y}\right)} \mathbf{K}^*_{min}\left({\mathbf{K}^*_{min}}^\top \mathbf{X}^\top \mathbf{Y}\right).$$

Podobne ako v predchádzajúcej rovnici, ani tu riešenie neexistuje ak

$$\left(\mathbf{Y}^\top \mathbf{X}\mathbf{K}^*_{min}\right)\left({\mathbf{K}^*_{min}}^\top \mathbf{X}^\top \mathbf{Y}\right) = 0.$$

Použitie a vlastnosti odhadu

DLS je vhodný najmä pri určitých typoch signálových procesov, ako pri dekonvolucii, identifikácii a ekvalizácii signálu. Pri tomto využití je DLS omnoho lepším nástrojom ako OLS alebo TLS, viď [10] na str. 409 a 410. Vlastnosti odhadu sú špecifické v závislosti na ďalších vlastnostiach matice regresorov, viac viď [12].

2.2.2 Ortogonálna metóda – TLS

Posledná metóda tejto skupiny, ktorú si popíšeme, sa používa pri chybách vyskytujúcich sa na oboch stranách, ako závislej, tak aj nezávislej premennej. TLS sa v štatistike často označuje ako EIV (*anglicky Errors-in-ariables*) modelovanie. Pracujeme s rovinnými dátami a snažíme sa minimalizovať súčet druhých mocnín vzdialeností dát od odhadovanej nadroviny, tento prístup sa nazýva *ortogonálna regresia*.

Nájdením matice $\left[\widehat{\varepsilon}, \widehat{\Xi}\right]$, ktorá splňuje minimalizačnú podmienku z (2.3), je riešenie TLS určené ako ľubovolné $\boldsymbol{\beta}$ splňujúce rovnicu $\mathbf{Y} + \widehat{\varepsilon} = \left(\mathbf{X} + \widehat{\Xi}\right)\boldsymbol{\beta}$.

Existencia a jednoznačnosť riešenia TLS

Veta 7. *Nech singulárny rozklad matice* $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ *je daný vzťahom* $\boldsymbol{X} = \sum_{i=1}^{m} \sigma_i' \boldsymbol{U}'_{\bullet,i} {\boldsymbol{V}'_{\bullet,i}}^\top$ *a matice* $[\boldsymbol{Y}, \boldsymbol{X}] = \sum_{i=1}^{m+1} \sigma_i \boldsymbol{U}_{\bullet,i} \boldsymbol{V}_{\bullet,i}$. *Ak* $\sigma_m' > \sigma_{m+1}$, *potom*

$$\left[\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{X}}\right] := \left[\boldsymbol{Y} + \widehat{\varepsilon}, \boldsymbol{X} + \widehat{\Xi}\right] = \boldsymbol{U}\widehat{\Sigma}\boldsymbol{V}^\top,\ \widehat{\Sigma} = diag\left\{\sigma_1, \ldots, \sigma_m, 0\right\}, \tag{2.4}$$

s odpovedajúcou TLS korekčnou maticou

$$\left[\widehat{\varepsilon}, \widehat{\Xi}\right] = \sigma_{m+1} \boldsymbol{U}_{\bullet,m+1} \boldsymbol{V}_{\bullet,m+1}^{\top},$$

ktorá rieši rovnice (2.3)*, z ktorých ďalej dostávame odhad*

$$\widehat{\boldsymbol{\beta}}_{TLS} = -\frac{1}{V_{1,m+1}} \left[V_{2,m+1}, \ldots, V_{m+1,m+1}\right]^{\top},$$

ktorý existuje a je jediným riešením $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{X}}\boldsymbol{\beta}$*, pokiaľ* $\boldsymbol{V}_{1,m+1} \neq 0$.

Dôkaz. Viď [14] na str. 90.

❑

Definícia 9. *Nech* $\boldsymbol{M} \in \mathbb{R}^{n \times n}$*. Číslo* $\lambda \in \mathbb{R}$ *sa nazýva vlastným číslom matice* $\boldsymbol{M}$*, ak existuje nenulový vektor* $\boldsymbol{W}$ *taký, že* $\boldsymbol{M}\boldsymbol{W} = \lambda \boldsymbol{W}$*. Vektor* $\boldsymbol{W}$ *sa nazýva vlastným vektorom matice* $\boldsymbol{M}$*.*

Veta 8. *Ak sú navyše* $\boldsymbol{V}_{\bullet,i}$ *z* (2.4) *vlastnými vektormi matice* $[\boldsymbol{Y}, \boldsymbol{X}]^{\top} [\boldsymbol{Y}, \boldsymbol{X}]$*, potom* $\widehat{\boldsymbol{\beta}}$ *splňuje rovnicu*

$$[\boldsymbol{Y}, \boldsymbol{X}]^{\top} [\boldsymbol{Y}, \boldsymbol{X}] \begin{bmatrix} -1 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}^{\top}\boldsymbol{Y} & \boldsymbol{Y}^{\top}\boldsymbol{X} \\ \boldsymbol{X}^{\top}\boldsymbol{Y} & \boldsymbol{X}^{\top}\boldsymbol{X} \end{bmatrix} \begin{bmatrix} -1 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \sigma_{m+1}^2 \begin{bmatrix} -1 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix}. \tag{2.5}$$

Z tej následne plynie

$$\widehat{\boldsymbol{\beta}}_{TLS} = \left(\boldsymbol{X}^{\top}\boldsymbol{X} - \sigma_{m+1}^2 \boldsymbol{I}_m\right)^{-1} \boldsymbol{X}^{\top}\boldsymbol{Y}.$$

Dôkaz. Odvodenie rovnosti (2.5) viď [17] na str. 37.

❑

Zo štatistického hľadiska je situácia $\sigma_m^{'} = \sigma_{m+1}$ veľmi málo pravdepodobná, ale predpokladajme, že nastala. V tom prípade riešenie existuje, ale je lineárnou kombináciou *najmenších singulárnych vektorov*, čiže riešení je nekonečne mnoho. V prípade, že $\sigma_m^{'} \neq \sigma_{m+1}$, ale $V_{1,m+1} = 0$, pre algoritmus z (2.3) neexistuje riešenie.

EIV modelovanie a vlastnosti odhadu

Od existencie a formy riešenia sa teraz postupne dopracujeme k jeho štatistickým vlastnostiam. Ako bolo spomínané, v štatistike sa označuje tento postup ako EIV modelovanie. Označme napozorované dáta $\mathbf{X}$, $\mathbf{Y}$ meraných veličín $\mathbf{X}_0$, $\mathbf{Y}_0$ s chybami ε a Θ splňujúce

$$\mathbf{Y} = \mathbf{Y}_0 + \varepsilon,$$
$$\mathbf{X} = \mathbf{X}_0 + \theta,$$

kde riadky matice $[\varepsilon, \Theta]$ sú rovnako rozdelené náhodné vektory s kovariančnou maticou $\sigma_*^2 \mathbf{I}_n$, kde $\sigma_*^2 > 0$ a nulovou strednou hodnotou.

Veta 9 (Slabá konzistencia)**.** *Predpokladajme, že distribúcia riadkov matice* $[\varepsilon,\Theta]$ *má konečný štvrtý moment. Ak*

$$\frac{1}{\sqrt{n}}\lambda_{min}\left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right) \longrightarrow \infty,\ n \longrightarrow \infty,$$
$$\frac{\lambda_{min}^2\left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)}{\lambda_{max}\left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)} \longrightarrow \infty,\ n \longrightarrow \infty,$$

kde $\lambda_{min}\,(\lambda_{max})$ *označujeme minimálnu (maximálnu) vlastnú hodnotu danej matice, potom*

$$\widehat{\boldsymbol{\beta}}_{TLS} \xrightarrow{P} \boldsymbol{\beta},\ pre\ n \longrightarrow \infty.$$

Dôkaz. Podľa [7] na str. 9. ❑

Veta 10 (Silná konzistencia)**.** *Ak* $lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}_0^\top \boldsymbol{X}_0$ *existuje a je väčšia ako nula, potom*

$$lim_{n\to\infty}\widehat{\boldsymbol{\beta}}_{TLS} \overset{\text{a.s.}}{=} \boldsymbol{\beta}.$$

Dôkaz. Viď [8] na str. 35. ❑

Ďalšie vlastnosti TLS odhadu viď [8] kap. 4.

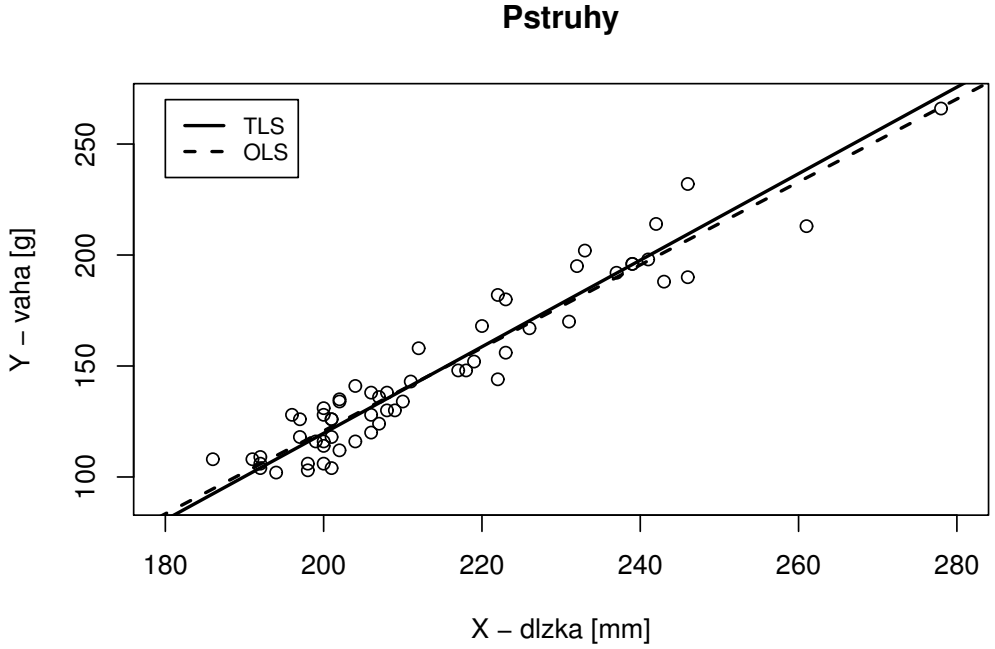
2.2.3 Analýza dát pstruha obyčajného

V tejto časti nadviažeme na príklad z motivácie, kde sme sa zaoberali závislosťou váhy a výšky. Uvažovali sme zjednodušený prístup, pri ktorom sme predpokladali fluktuáciu len na strane vysvetľovanej premennej teda váhy. V skutočnosti je treba uvažovať chybu aj na strane vysvetľujúcej premennej, v našom prípade výšky, ktorá je taktiež ovplyvňovaná rôznymi prírodnými faktormi.

Podobný problém ako v motivácii riešia aj ekológovia z Výskumného ústavu vodohospodárskeho T. G. Masaryka, ktorých zaujíma jednoduchý lineárny vzťah medzi dĺžkou a váhou pstruha obyčajného (latinsky Salmo trutta morpha fario) z malých horských potokov Národného parku Šumava. Dáta obsahujú 59 párov meraní dĺžok a váh dospelých pstruhov. Pstruhy boli chytané v rozličných častiach viacerých potokov. Po meraniach boli všetky pstruhy pustené späť do potokov.

Je treba si uvedomiť určité vlastnosti pozorovaných veličín. Ako sme sa zmienili, chyby sa nachádzajú na oboch stranách premenných, dĺžky aj váhy, ktoré boli spôsobené rôznymi prírodnými faktormi. Teraz na základe spomínaných predpokladov môžeme pristúpiť k hľadaniu riešenia metódou TLS respektíve EIV modelovaniu.

Keďže ťažisko pozorovaní neleží v počiatku súradnicovej sústavy, na naše dáta aplikujeme model s interceptom. Na obrázku 2.1 môžeme porovnať rozdiel medzi použitím OLS a TLS.



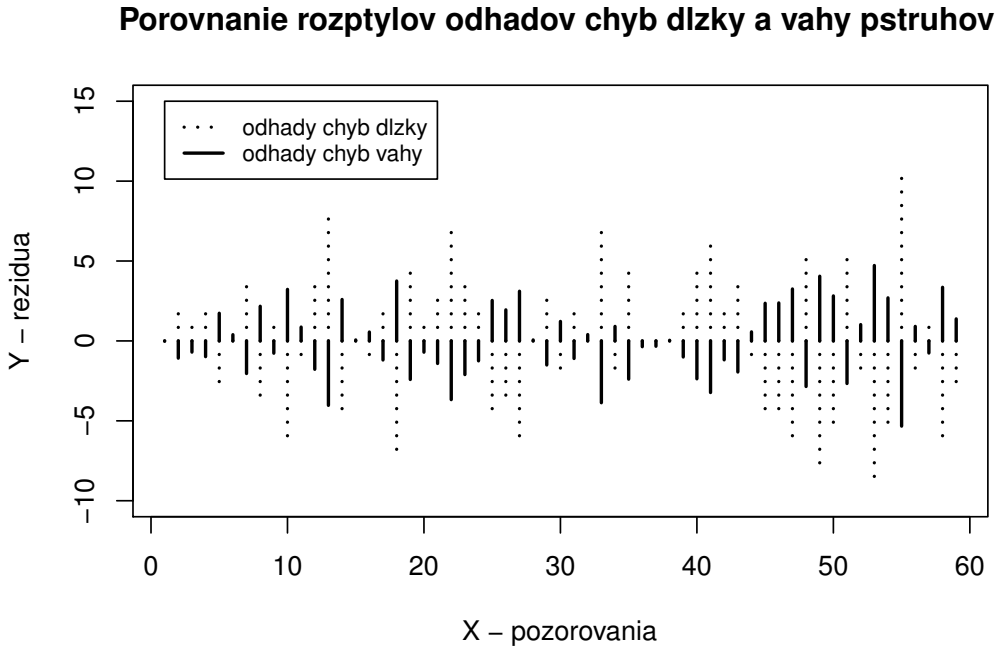
Obrázok 2.1: Ukážka ortogonálnej metódy a základnej metódy najmenších štvorcov na reálnych dátach (dĺžka – váha pstruha obyčajného).

Po výpočte regresnej priamky metódou TLS overíme predpoklady modelu EIV tzn. nulovosť stredných hodnôt, rovnosť rozptylov chýb ε, Θ pomocou ich odhadov $\widehat{\varepsilon}$ a $\widehat{\Theta}$. Tieto odhady získame ako rozdiely pozorovaní a ich kolmých priemetov na regresnú priamku, kde prvá súradnica bude reprezentovať odhad chyby dĺžky, $\widehat{\Theta}$ a druhá váhy, $\widehat{\varepsilon}$. Na obrázku 2.2 sú uvedené tieto odhady pre jednotlivé pozorovania, z ktorých možno vidieť, že rozptyly odhadov chýb dĺžky a váhy pstruhov sú zhruba rovnaké. Bodové odhady strednej hodnoty ε a Θ z dát $\widehat{\varepsilon}$ a $\widehat{\Theta}$ vyšli $\widehat{\mathsf{E}\,\varepsilon} = 0.0000170749$, $\widehat{\mathsf{E}\,\Theta} = -0.0000332824$. Takmer nulové hodnoty týchto odhadov nenaznačujú porušenie predpokladu EIV modelu o nulovej strednej hodnote chýb.

Z Obrázku 2.2 nebadať porušenie homoskedasticity (konštantnosti rozptylu chýb), už len preto, že nie je viditeľný žiadny "vzor"(napríklad nepozorujeme znižujúce sa hodnoty chýb s narastajúcim počtom pozorovaní) a žiadna chyba sa nejaví ako odľahlá.

Odhad metódou TLS by mal v tomto prípade poskytovať lepší a presnejší odhad, keďže sme použili vhodnejší model využívajúci viac informácií z dát. Ďalším veľmi doležitým pozorovaním z obrázka 2.1 vpravo hore, je pôsobenie odľahlých pozorovaní na regresné priamky. Môžeme si všimnúť, že metóda TLS je voči nim odolnejšia než OLS. Táto problematika je ďalej rozoberaná v kap. 2 [6].

Dáta dĺžky a váhy pstruha obyčajnéhu pochádzjú z Výskumního ústvu vodohospodářskeho T. G. Masaryka T. G. Masaryka, v.v.i. http://www.vuv.cz/.



Obrázok 2.2: Grafické porovnanie rozptylov dĺžky a váhy pstruhov pomocou odhadov chýb $\widehat{\varepsilon}$, $\widehat{\mathbf{\Theta}}$.

2.2.4 Regresné priamky OLS, DLS a TLS

V tejto časti sa zaoberáme prekladaním dát regresnou priamkou s interceptom v tvare

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X. \tag{2.6}$$

Veta 11. *Koeficienty regresnej priamky OLS sú*

$$\widehat{\beta}_1 = \bar{Y} - \widehat{\beta}_2 \bar{X},\ \widehat{\beta}_2 = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}, \tag{2.7}$$

kde $\bar{X} = \sum_{i=1}^n X_i/n$ *a* $\bar{Y} = \sum_{i=1}^n Y_i/n$.

Dôkaz. Vyjadrenia dostaneme z rovníc (1.4) kde $\mathbf{X}$ má prvý sĺpec jednotiek a v druhom stĺpci sú napozorované hodnoty nezávislej premennej.

❑

Veta 12. *DLS koeficienty priamky sú*

$$\widehat{\beta}_1 = -\frac{\bar{X} - \widehat{\beta}_2 \bar{Y}}{\widehat{\beta}_2},\ \widehat{\beta}_2 = \frac{\sum_{i=1}^n Y_i^2 - n\bar{Y}^2}{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}. \tag{2.8}$$

Dôkaz. Kedže tento postup minimalizuje chyby na strane jednorozmernej vysvetľujúcej premennej, použijeme metódu OLS na upravených dátach, kde sme zložky X_i zamenili

so zložkami Y_i. Takto vypočítané koeficienty však platia pre inverznú priamku, keďže sme ich získali výpočtom z prehodených súradníc. Posledným krokom je aplikovanie inverzného zobrazenia na spomínanú rovnicu priamky, čím sme dospeli k hľadanej regresnej priamke a koeficientom.

❑

Veta 13. *Pre priamku TLS platí*

$$\widehat{\beta}_1 = \bar{Y} - \widehat{\beta}_2 \bar{X},\ \widehat{\beta}_2 = \frac{s_Y^2 - s_X^2 + \sqrt{\left(s_Y^2 - s_X^2\right)^2 + 4s_{XY}^2}}{2s_{XY}}, \tag{2.9}$$

za predpokladu $s_{XY} \neq 0$, *pričom* $s_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}$, $s_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$, $s_Y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2$.

Dôkaz. Viď [3] na str. 222.

❑

Definícia 10. *Majme body* $(X_1, Y_1), \ldots, (X_n, Y_n)$. *Ťažisko týchto bodov je dané vzťahom*

$$\boldsymbol{T} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \tag{2.10}$$

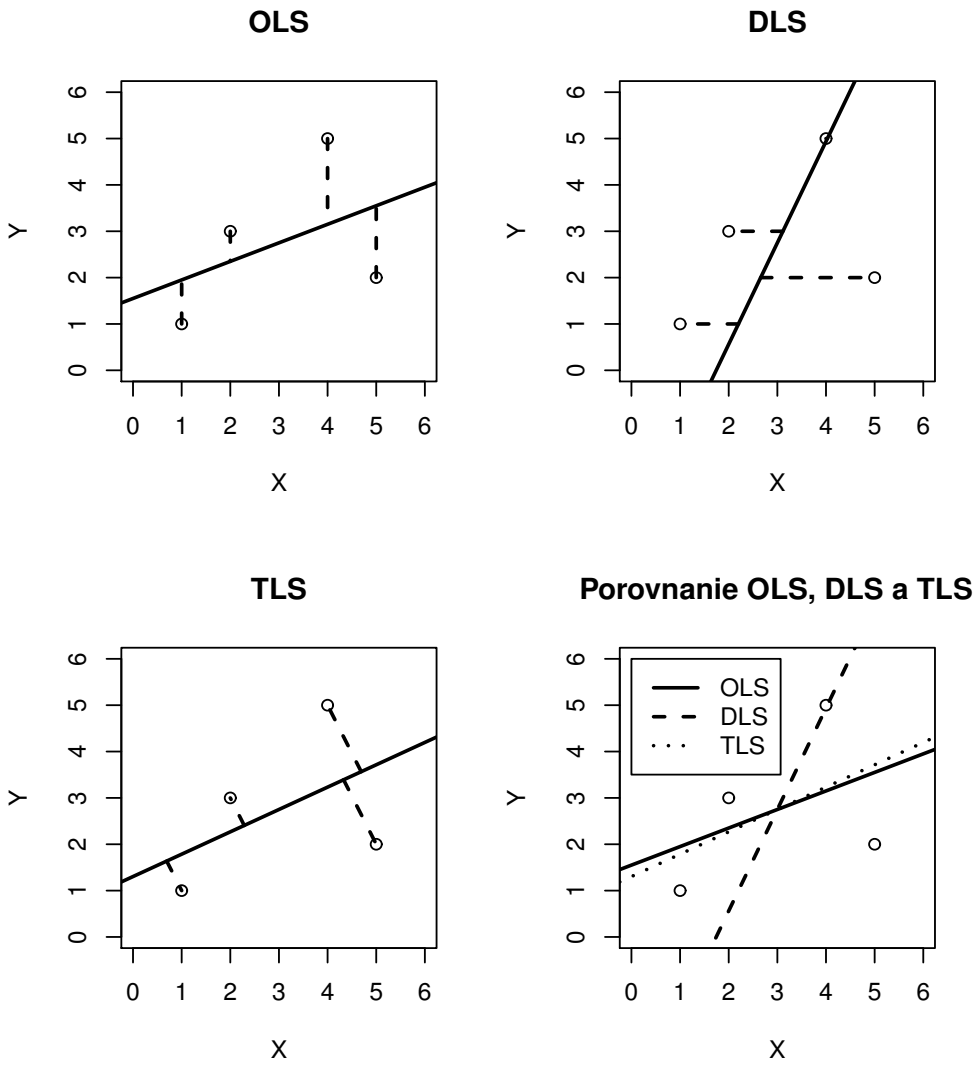
Lemma 14. *Každá z vyššie uvedených regresných priamok prechádza ťažiskom, z čoho plynie, že priamky sa pretínajú práve v ňom (viď obrázok 2.3).*

Dôkaz. Jednoduchým dosadením súradníc ťažiska (2.10) do rovnice priamky (2.6), kde za $\widehat{\beta}_1$ a $\widehat{\beta}_2$ dosadíme keoficienty (2.7), (2.8) a (2.9) sa môžeme presvedčiť, že ťažisko skutočne leží na každej zo spomínaných regresných priamok.

❑

Využitie ťažiska pri výpočte odhadu s interceptom

Skutočnosť, ktorú sme si v predchádzajúcej časti ukázali na priamkach, môžeme rovnakým spôsobom rozšíriť aj pre viacrozmerné regresory, čo nám umožňuje postup, v ktorom si najskôr spočítame ťažisko. Následne posunieme všetky pozorovania opačným smerom ako je vektor bodu ťažiska, čím docielime, že ťažisko takto vzniknutých bodov bude ležať v počiatku súradnicovej sústavy. Potom aplikujeme vybranú metódu bez interceptu a takto získanú regresnú priamku posunieme naspäť o už spomínaný vektor. Výsledkom bude regresná nadrovina s absolútnym členom.



Obrázok 2.3: Porovnanie regresných priamok metód OLS, DLS, TLS, ktoré sa navzájom pretínajú v ťažisku pozorovaní.

2.3 Odhady s vysokým bodom zlyhania

Doteraz uvedené metódy používali v plnej miere všetky napozorované hodnoty a boli extrémne citlivé na odľahlé pozorovania, preto sa v nasledujúcom texte budeme venovať metódam, ktoré tento problém riešia. Patrí medzi ne metóda vážených štvorcov LWS (*anglicky Least weighted squares*), v ktorej je každému bodu priradená príslušná váha a metóda useknutých štvorcov LTS (*anglicky Least trimmed squares*). Graficky v metóde LTS ide o nájdenie najtesnejšieho pásu pokrývajúceho najväčšiu časť pozorovaní.

K posudzovaniu robustnosti odhadu, odolnosti odhadu voči odľahlým pozorovaniam (ak existujú), bola navrhnutá celá rada meradiel. Jedným z nich je napríklad bod zlyhania *(anglicky breakdown point)*. Vyjadruje percentuálne zastúpenie chybných hodnôt, pre ktoré je ešte odhad "správny"(teda neovplyvnený podstatným výskytom odľahlých hodnôt).

2.3.1 Metóda najmenších vážených štvorcov – LWS

Postup je založený na klasickej metóde OLS, v ktorej sa sčitujú štvorce reziduí násobené príslušnými váhami. Tento postup síce poskytuje lepší odhad pri vychýlených pozorovaniach, ale za cenu vysokého bodu zlyhania.

Varianty prístupov hľadania riešenia

Existuje viac prístupov, jeden z nich je založený na určovaní hodnôt $w_1,\ldots,w_n$, ktoré sú priradené implicitne v priebehu výpočtu. Váhy potom závisia na hodnotách reziduí príslušného hrubého odhadu $\widehat{\boldsymbol{\beta}}$ skutočného parametra $\boldsymbol{\beta}$, kde reziduum $u_i \equiv u_i\left(\widehat{\boldsymbol{\beta}}\right)$ je rozdiel medzi vyrovnanou a napozorovanou hodnotou v danom bode, $u_i = Y_i - \mathbf{X}_{i,\bullet}\widehat{\boldsymbol{\beta}}$.

Usporiadame druhé mocniny reziduí pre dané $\boldsymbol{\beta}$

$$\left(u^2(\boldsymbol{\beta})\right)_{(1)} \leq \cdots \leq \left(u^2(\boldsymbol{\beta})\right)_{(n)}, \tag{2.11}$$

ktorým odpovedá nerastúca postupnosť váh $w_1,\ldots,w_n$. Pre takto zvolené veličiny definujeme odhad metódou LWS ako

$$\widehat{\boldsymbol{\beta}}_{LWS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} w_i \left(u^2(\boldsymbol{\beta})\right)_{(i)}.$$

Váhy môžeme určiť buď lineárne klesajúcim spôsobom $w_i = 1-(i-1)/n$, alebo dvojkrokovou procedúrou pre výpočet adaptívnych váh, ktorá váhy určuje automaticky. Podrobnejší popis tejto procedúry je uvedený v [5] na str. 269.

Postup hľadania odhadu parametra – LWS

Iný postup, ktorý ukážeme, spočíva prevedením problému LWS na OLS za predpokladu, že váhy jednotlivých pozorovaní sú vopred známe.

Veta 15. *Majme lineárny model* (1.1)*, kde ε má nulovú strednú hodnotu, kovariančnú maticu $\sigma^2\mathbf{V}$, kde $\sigma^2 > 0$, $\mathbf{V}$ je pozitívne definitná matica a $\mathbf{X}$ má plnú stĺpcovú hodnosť.*

Potom odhad pomocou váženej metódy najmenších štvorcov je tvaru

$$\widehat{\boldsymbol{\beta}}_{LWS} = \left(\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{2.12}$$

Dôkaz. Tento problém prevedieme na klasickú metódu OLS, kde na základe predpokladov, že $\mathbf{V} > 0$, existuje $\mathbf{V}^{\frac{1}{2}}$, ktorá je symetrická a regulárna. Potom

$$h\left(\mathbf{V}^{-\frac{1}{2}}\mathbf{X}\right) = h(\mathbf{X}) = h\left(\mathbf{X}^\top \mathbf{V}^{-\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} \mathbf{X}\right) = h\left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right),$$

z čoho plynie regularita matice $\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X}$ a existencia inverznej matice.

Definujme premenné $\mathbf{Z} := \mathbf{V}^{\frac{1}{2}}\mathbf{Y}, \mathbf{F} := \mathbf{V}^{\frac{1}{2}}\mathbf{X}, \boldsymbol{\eta} := \mathbf{V}^{\frac{1}{2}}\boldsymbol{\varepsilon}$, ktoré ak substituujeme do (1.2) $\mathbf{Z} \to \mathbf{Y}$, $\mathbf{F} \to \mathbf{X}$ *a* $\boldsymbol{\eta} \to \boldsymbol{\varepsilon}$, dostaneme

$$\mathbf{Z} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\eta} \Rightarrow \mathbf{V}^{\frac{1}{2}}\mathbf{Y} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{\frac{1}{2}}\boldsymbol{\varepsilon}.$$

Zostáva ukázať, že

$$\mathsf{E}\,\boldsymbol{\eta} = \mathsf{E}\,\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon} = \mathbf{V}^{-\frac{1}{2}}\,\mathsf{E}\,\boldsymbol{\varepsilon} = 0 \text{ a } \operatorname{cov}\left(\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}\right) = \sigma^2 \mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}} = \sigma^2 \mathbf{I}_n,$$

čím sme dokázali, že model splňuje predpoklady OLS. Odhad teda môžeme vyjadriť v tvare

$$\widehat{\boldsymbol{\beta}}_{LWS} = \left(\mathbf{F}^\top \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{Z} = \left(\mathbf{X}^\top \mathbf{V}^{-\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{V}^{-\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} \mathbf{Y} =$$
$$= \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}.$$

❑

Matica $\mathbf{V} = diag\{v_1, \ldots, v_n\}$, potom definujeme maticu

$$\mathbf{W} = \mathbf{V}^{-1} = diag\left\{\frac{1}{v_1}, \ldots, \frac{1}{v_n}\right\} = diag\{w_1, \ldots, w_n\},$$

kde $w_1, \ldots, w_n$ sú váhy, ktorých hodnoty sú nepriamo úmerné veľkosti rozptylu. Odhad neznámeho parametra metódou LWS možno vyjadriť v tvare

$$\widehat{\boldsymbol{\beta}}_{LWS} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y}.$$

Vlastnosti odhadu

Veta 16 (Nestrannosť)**.** *Odhad metódou najmenších vážených štvorcov je nestranným odhadom neznámeho parametra*

$$E\,\widehat{\boldsymbol{\beta}}_{LWS} = E\,\boldsymbol{\beta}.$$

Dôkaz. Tento prístup je v podstate metóda OLS. Postup dôkazu viď (1.5) len s inými premennými, kde vektor $\boldsymbol{\eta}$ má tiež nulovú strednú hodnotu a kovariančnú maticu $\sigma^2\mathbf{I}_n$.

❑

Ak by boli ešte zložky vektora $\boldsymbol{\eta}$ nezávislé, odhad by bol konzistentný $\widehat{\boldsymbol{\beta}}_{LWS} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}$ a zachoval by všetky už popisované vlastnosti odhadu OLS.

Rozdielnosť prístupov je daná spomínaným spôsobom určovania váh. Po ich určení je postup identický, len vlastnosti odhadov sú rôzne v závislosti na predpokladoch váhovej matice $\mathbf{W}$, respektíve $\mathbf{V}$. Vlastnosti dvojkrokovej metódy LWS viď [5] na str. 272.

Táto metóda sa vo veľkej miere používa napríklad vo vyhladzovaní obrazu, šumu alebo pri lokalizácii objektov ako oči, tvár, kde je potrebná veľká odolnosť voči odľahlým pozorovaniam, viď [13] na str. 113.

2.3.2 Metóda najmenších useknutých štvorcov – LTS

Týmto problémom sa ako prvý začal zaoberať Rousseeuw v roku 1984. Ako naznačuje názov, snažíme sa určiť najodľahlejšie pozorovania, ktoré do postupu vôbec nezahrnieme a na zvyšné pozorovania použijeme metódu OLS. Celý postup je založený na efektívnej identifikácii týchto pozorovaní.

Postup hľadania odhadu parametra – LTS

Majme zoradené štvorce reziduí ako v (2.11) a označme *bod zlyhania* h ako počet pozorovaní, ktoré sa viažu s prvými h najmenšími reziduami, ktoré budeme používať pri výpočte odhadu neznámeho parametra $\boldsymbol{\beta}$. Odhad metódou LTS definujme vzťahom

$$\widehat{\boldsymbol{\beta}}_{LTS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{h} \left(u^2(\boldsymbol{\beta})\right)_{(i)},$$

kde $\left(u^2(\boldsymbol{\beta})\right)_{(i)}$ je i-tá najmenšia druhá mocnina rezidua pri danom $\boldsymbol{\beta}$.

Bod h závisí na štruktúre a počte pozorovaní n, $h \equiv h(n)$. Čím viac odľahlejších pozorovaní máme v súbore o veľkosti n, tým menší by mal byť podiel n v $h(n)$. Pre maticu $\mathbf{X}$ s hodnosťou p sa všeobecne $h = (n+p+1)/2$, čo je pre veľké n rovné približne $n/2$. V praxi sa volí $h \approx 0{,}75n$. Ako v predchádzajúcej metóde, aj teraz existuje niekoľko prístupov výpočtu odhadu neznámeho paramera, z ktorých uvedieme spôsob vytvorený Rousseeuwom a Van Driessenom.

Algoritmus FAST-LTS

Ide o viackrokový postup tzv. FAST-LTS, ktorý je založený na identifikácii odľahlých pozorovaní pomocou ich štandardizovaných reziduí $u_i/\widehat{\sigma}_{opt}$, kde $\widehat{\sigma}_{opt}$ je hrubý TLS odhad smerodatnej odchýlky, ktorý popíšeme neskôr. Tento prístup ďalej vedie k odhadu metódou vážených najmenších štvorcov.

Algoritmus je založený na iterácii daného kroku, ktorá skončí splnením určitých podmienok. Na začiatku inicializujeme vstupné hodnoty. Majme p náhodne vybraných h-prvkových podmnožín množiny $\{1,\ldots,n\}$, kde p je hodnosť matice $\mathbf{X}$ a h je *bod zlyhania*. Pre každú z týchto podmnožín spočítame odhad pomocou metódy OLS

a následne príslušný súčet štvorcov reziduí, pričom do výpočtov zahrnieme len tie pozorovania, ktorých index sa nachádza v danej podmnožine. Podmnožinu, ktorej už spomínaný súčet štvorcov reziduí bude najmenší, označíme ako počiatočnú množinu H_1 a k nej už spočítaný príslušný odhad označíme ako $\widehat{\boldsymbol{\beta}}_1$.

Ďalším krokom je konštrukcia novej množiny indexov pozorovaní H_2, ktorá bude pozostávať z h indexov h najmenších absolutných hodnôt reziduí $\left|u_i\left(\widehat{\boldsymbol{\beta}}_1\right)\right|$, kde $i \in \{1,\ldots,n\}$. Aplikovaním OLS na pozorovania s indexami z množiny H_2 dostávame opäť odhad parametra $\widehat{\boldsymbol{\beta}}_2$.

Predchádzajúci krok opakujeme až do okamihu, kým

$$\sum_{i=1}^{h}\left(u^2\left(\widehat{\boldsymbol{\beta}}_j\right)\right)_{(i)} = \sum_{i=1}^{h}\left(u^2\left(\widehat{\boldsymbol{\beta}}_{j-1}\right)\right)_{(i)}.$$

V praxi pri numerickom výpočte je obtiažne túto rovnosť dosiahnuť, preto pri programovaní v softvéri vopred stanovíme "presnosť", napríklad $\delta = 10^{-9}$ a iteruje sa kým

$$\left|\sum_{i=1}^{h}\left(u^2\left(\widehat{\boldsymbol{\beta}}_j\right)\right)_{(i)} - \sum_{i=1}^{h}\left(u^2\left(\widehat{\boldsymbol{\beta}}_{j-1}\right)\right)_{(i)}\right| < \delta.$$

Po ukončení iterácií máme optimálnu množinu H_{opt} a odhad parametra $\widehat{\boldsymbol{\beta}}_{opt}$. Použitím tohto odhadu môžeme vyjadriť, už zmienený odhad smerodatnej odchýlky

$$\widehat{\sigma}_{opt} = c_h\sqrt{\frac{1}{h}\sum_{i=1}^{h}\left(u^2\left(\widehat{\boldsymbol{\beta}}_{opt}\right)\right)_{(i)}}.$$

Hodnota c_h zaručuje nestrannosť a konzistenciu $\widehat{\sigma}_{opt}$ pri rovnako rozdelených, vzájomne nezávislých chybách z normálneho rozdelenia s nulovou strednou hodnotou a konštantným rozptylom. Je daná vzťahom

$$c_h = \sqrt{\frac{n}{h(\alpha)}\int_{-q}^{q} u^2\, d\Phi} \text{ a } q = \Phi^{-1}\left(\frac{\alpha}{2}+\frac{1}{2}\right),$$

kde $\alpha = h(\alpha)/n$ a Φ je distribučná funkcia normovaného normálneho rozdelenia.

Prostredníctvom týchto odhadov môžeme určiť váhy jednotlivých pozorovaní, ktoré nadobúdajú hodnoty

$$w_i = \begin{cases} 0, & \text{ak } \left|u_i\left(\widehat{\boldsymbol{\beta}}_h\right)/\widehat{\sigma}_h\right| \leq \sqrt{\chi^2_{1,0.975}}, \\ 1, & \text{inak.} \end{cases}$$

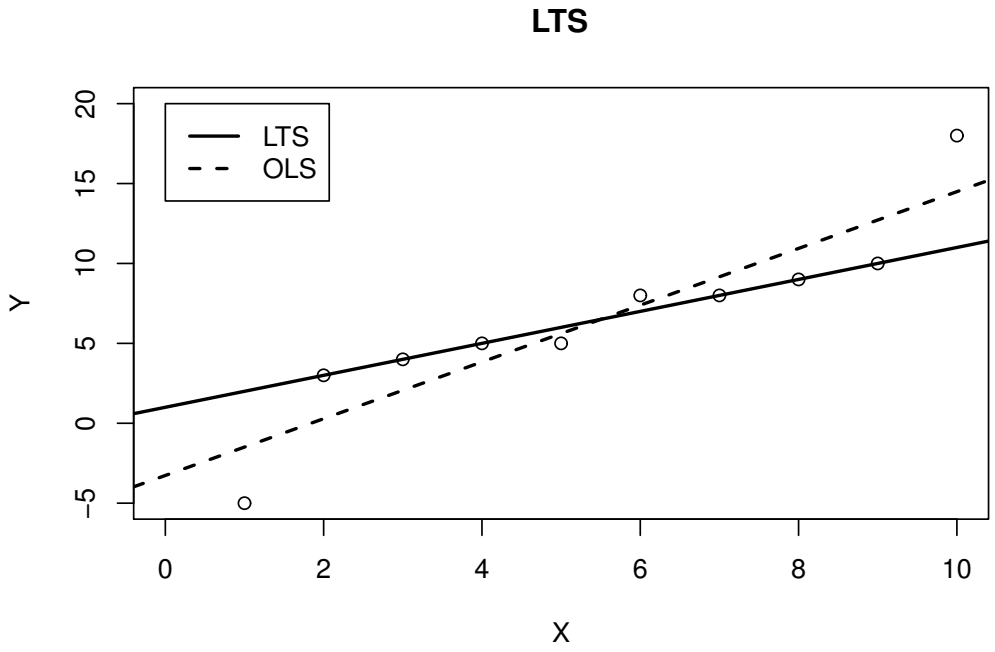
Potom hľadaný odhad parametra je v tvare

$$\widehat{\boldsymbol{\beta}}_{LTS} = \left(\sum_{i=1}^{n} w_i \mathbf{X}_{i,\bullet}^{\top}\mathbf{X}_{i,\bullet}\right)^{-1}\left(\sum_{i=1}^{n} w_i \mathbf{X}_{i,\bullet}^{\top} Y_i\right).$$

Bližší popis môžeme nájsť v [11] kap. 3 alebo iné postupy [16] podkap. 1.2 a podkap. 5.3. Pre lepšie použitie na reálnych dátach používame variantu s interceptom, viď obrázok 2.4.

Vlastnosti odhadu

Asymptotické vlastnosti LTS odhadov nie sú také hladké ako pri OLS, ale možno preukázať ich asymptotickú normalitu, tzn. $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{\mathrm{D}} N(0,\mathbf{V})$ za predpokladu, že chyby pochádzajú taktiež z normálneho rozdelenia a $\mathbf{V}$ je pozitívne definitná matica. Ďalšie vlastnosti a ich dôkazy sú uvedené v [16] na str. 179.



Obrázok 2.4: Odolnosť metódy LTS v porovnaní s citlivosťou metódy OLS voči odľahlým pozorovaniam.

Záver

Alternatívy metódy najmenších štvorcov sú veľmi využívané v rôznych odvetviach, napríklad pri spracovaní obrazu (LWS), signálových procesoch (DLS), štatistických, ekonometrických modeloch (OLS, TLS, LTS) a v mnoho iných, či už teoretických alebo aplikovaných smeroch.

Tieto metódy sme rozdelili do dvoch skupín. V prvej sa tieto postupy zameriavali na výskyt fluktuácií. Ukázali sme si základné vlastnosti odhadov OLS, DLS a TLS za predom stanovených prepokladov. Pozorovali sme simulácie týchto postupov na dátach a porovnávali vhodnosť jednotlivých modelov pre dané sady dát.

Venovali sme sa aj analýze reálnych dát pozostávajúcich z pozorovaní dĺžky a váhy pstruha obyčajného. Na základe našej analýzy sme ako najvhodnejší postup označili TLS, pretože fluktuácia sa vyskytovala na oboch stranách premenných a naše odhady strednej hodnoty a rozptylu neporušovali predpoklady vybranej metódy.

V závere tejto časti sme demonštrovali všetky tri postupy na regresných priamkach. Výsledkom použitia týchto metód na rovnakej sade dát bolo, že regresné priamky sa nám pretli v jednom bode, ktorý bol zároveň aj ťažiskom našich pozorovaní. Z čoho sme následne odvodili postup, ako možno pomocou metódy bez interceptu získať regresnú nadrovinu s absolútnym členom.

Druhá skupina metód bola zameraná na vplyv jednotlivých pozorovaní na odhad trendu. Ako prvý sme si rozobrali postup, kde má každé pozorovanie rôznu váhu. Existujú dva postupy priradenia váh, v prvom sú váhy pozorovaní známe a v druhom sú priradené počas výpočtu. Po tomto kroku sú však postupy identické a pri splnení daných predpokladov majú tieto odhady rovnaké vlastnosti ako odhady parametru metódou OLS.

Posledná časť našej práce bola venovaná odľahlým pozorovaniam. Snažili sme sa efektívne eliminovať ich vplyv na odhad, čo nás v konečnom dôsledku priviedlo k už spomínanej metóde vážených štvorcov, kde však váha je rovná nule pre odľahlé pozorovanie a jedna inak. Tieto hodnoty sú určené na základe hrubých odhadov regresných koeficientov a rozptylu.

Bakalárska práca je svojím obsahom prínosná nielen pre študentov Matematicko-fyzikálnej fakulty UK, ale aj pre odbornú verejnosť. Za najväčší prínos práce považujem podrobné vysvetlenie postupov, odvodenie odhadov neznámych parametrov, grafické znázornenia a porovnania uvedených odhadov aj na reálnych dátach.

V práci sú obsiahnuté základné a v praxi najviac používané postupy. Vzhľadom k širokému uplatneniu v súčasnosti existuje veľké množstvo ďalších alternatív metódy najmenších štvorcov, ktoré sa neustále vylepšujú, a preto ich výskum nemožno považovať za ukončený.

Zoznam použitej literatúry

[1] ABATZOGLOU, T. J. a MENDEL, J. M. (1987): Constrained total least squares. *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP'87*. IEEE, 1485–1488.

[2] ANDĚL, J. (2007): *Základy matematické statistiky*. 2. vyd. Matfyzpress, Praha. ISBN 978-80-7378-001-2.

[3] ANDĚL, J. (2007): *Statistické metody*. 4. vyd. Matfyzpress, Praha. ISBN 80-7378-003-8.

[4] CIPRA, T. (2008): *Finanční ekonometrie*. 1. vyd. Ekopress, Praha. ISBN 978-80-86929-43-9.

[5] ČÍŽEK, P. (2008): Efficient robust estimation of time-series regression models. *Applications of Mathematics*, **53**(3), 267–279.

[6] FIERRO R.D. a BUNCH J.R. (1994): Collinearity and total least squares. *SIAM Journal on Matrix Analysis and Applications*, **15**, 1167–1181.

[7] GALLO, P.P. (1982): Consistency of regression estimates when some variables are subject to error. *Communications in Statistics: Theory and Methods*, **11**, 973–983.

[8] GLESER, L. J. (1981): Estimation in a multivariate „errors in variables" regression model: large sample results. *The Annals of Statistics*, 24–44.

[9] GOLUB, G. H a VAN LOAN, C. F. (1996): *Matrix computation*. 3. vyd. Johns Hopkins University Press, Baltimore. ISBN 978-1-4214-0859-0.

[10] GROAT, R. D. D. a DOWLING, E. M. (1993): Data least squares problem and channel equalization. *IEEE Transactions on signal processing*, **41**, 407–411.

[11] HUBERT, M., ROUSSEEUW, P. J. a VAN AELST, S. (2008): High-breakdown robust multivariate methods. *Statistical Science*, 92–119.

[12] CHANG, X.-W., GOLUB, G. H. a PAIGE, C. C. (2008): Towards a backward perturbation analysis for data least squares problems. *SIAM Journal on Matrix Analysis and Applications*, **30**(4), 1281–1301.

[13] KALINA, J. (2007): Locating the mouth using weighted templates. *Journal of applied mathematics, statistics and informatics*, **3**(1), 111–125.

[14] PEŠTA, M. (2008): Total Least squares approach in regression methods. *WDS'08 Proceedings of Contributed Papers: Part I – Mathematics and Computer Sciences*, MatfyzPress, Praha, 88–93.

[15] RAO, C. R., TOUTENBURG, H., FIEGER, A., HEUMANN, C., NITTNER, T. a SCHEID, S. (1999): *Linear models: least squares and alternatives*. 2. vyd. Springer-Verlag, Berlin, Heidelberg, New York. ISBN 0-387-988488-3.

[16] ROUSSEEUW, P. J. a LEROY, A. M. (1987): *Robust regression and outlier detection*. Wiley, New York. ISBN 0-471-85233-3.

[17] VAN HUFFEL, S. a VANDEWALLE, J. (1991): *The total least squares problem: computational aspects and analysis*. SIAM, Philadelphia. ISBN 0-89871-275-0.

[18] WOOLDRIDGE J.M. (2009): *Introductory econometrics: a modern approach*. South-Western Pub, Mason. ISBN 0-324-66054-5.

Zoznam obrázkov

Prílohy

Príloha č. 1

K práci je priložené CD obsahujúce program math.nb, ktorý je naprogramovaný v systéme Wolfram Mathematica 9.0 for Students, skript programu R s názvom rko.R a súbor data.csv obsahujúci napozorované dáta. Dokumenty obsahujú príkazy, simulácie ako aj tvorbu grafov jednotlivých modelov použitých v našej práci. Na CD sa nachádza tiež táto práca v PDF formáte.