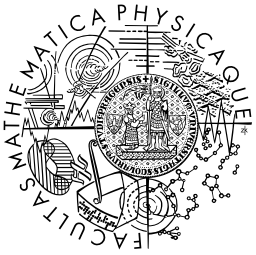


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Karel Chuchel

Úplně nejmenší čtverce a jejich asymptotické vlastnosti

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Michal Pešta, Ph.D.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2020

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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 má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Děkuji svému vedoucímu práce, doc. RNDr. Michalu Peštovi, Ph.D., za všechno pochopení, trpělivost a pomoc při psaní této práce.

Název práce: Úplně nejmenší čtverce a jejich asymptotické vlastnosti

Autor: Karel Chuchel

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Tato práce se zabývá metodou úplně nejmenších čtverců, která slouží pro odhad parametrů v lineárních modelech. V práci je uveden základní popis metody a její asymptotické vlastnosti. Je vysvětleno, jakým způsobem lze v konceptu metody využít neparametrický bootstrap pro hledání odhadu. Vlastnosti bootstrap odhadů jsou pak simulovány na pseudo náhodně vygenerovaných datech. Simulace jsou prováděny pro dvourozměrný parametr v různých nastaveních základního modelu. Jednotlivé bootstrap odhady jsou v rovině řazeny pomocí Mahalanobis a Tukey statistical depth function. Simulace potvrzují, že bootstrap odhad dává dostatečně dobré výsledky, aby se dal využít pro reálné situace.

Klíčová slova: úplně nejmenší čtverce, neparametrický bootstrap, statistical depth function, asymptotické vlastnosti

Title: Total Least Squares and Their Asymptotic Properties

Author: Karel Chuchel

Department: Department of Probability and Mathematical Statistics

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

Abstract: Total least squares is a method, which solves linear models. In this thesis we state the basic and asymptotic properties of the method. Afterwards we show the application of non-parametric bootstrap in total least squares' design. We then simulate the quality of bootstrap estimates with pseudo-random generated data. For simulations we consider two-dimensional parameter and various settings of the base model. We use Tukey and Mahalanobis statistical depth function for sorting bootstrap resamples in the plain. The results of the simulations are favourable – they confirm the validity of total least squares bootstrapping for the real scenarios.

Keywords: total least squares, non-parametric bootstrap, statistical depth function, asymptotic properties

Obsah

Úvod

Úplné nejmenší čtverce (TLS, total least squares) jsou metoda, která řeší tzv. lineární modely. Kdy předpokládáme, že vysvětlované proměnné jsou lineární funkcí vysvětlujících. Je velmi často srovnávána se standardními nejmenšími čtverci (OLS, ordinary least squares). Jejich hlavní rozdíl spočívá v přístupu k proměnným v rovnici. V případě TLS povolujeme náhodnou chybu i u vysvětlujících proměnných.

Z hlediska našeho myšlenkového aparátu jsou proměnné v podstatě záměnné, při prohození vysvětlující a vysvětlované proměnné se naše analýza příliš nemění. Kdežto u OLS má jedna proměnná výsadní postavení – naše myšlení tak v podstatě implicitně obsahuje kauzalitu a prohozením proměnných otáčíme i tuto příčinnou souvislost.

V případě, že se chyby objevují i u regresorů, je OLS odhad nekonzistentní. Metoda TLS tak našla využití v mnoha oborech lidské činnosti. Např. v astronomii může přispět k zpřesnění výpočtu dráhy planet. Široce využívána je v aplikacích souvisejících se zpracováním signálu nebo v ekonomii pro modelování vývoje cen na burze.

Struktura práce je následující. První kapitola se věnuje jednoduchému modelu přímé úměrnosti. Pro tento model si odvodíme OLS a TLS řešení. Přičemž si při odvození vysvětlíme základní logiku obou metod. První kapitola slouží jako úvod, který můžeme využít pro lepší pochopení následujícího textu.

V druhé kapitole si již pro metodu TLS vybudujeme exaktní matematický aparát. Zadefinujeme si problémy, které řešíme. Odvodíme si řešení a ukážeme jeho statistické vlastnosti.

Protože asymptotické rozdělení TLS odhadu se ukáže jako nesnadno vyjádřitelné, seznámíme se ve třetí kapitole s metodou bootstrap. Bootstrap je metoda, která umožňuje získat představu o rozdělení odhadu, aniž bychom si dopředu kladli na toto rozdělení velká omezení.

Ve čtvrté kapitole pak budeme ukazovat na simulovaných datech, jestli je metoda bootstrap pro odhad v praxi využitelná. Data si budeme generovat pro dvourozměrný parametr. Zároveň simulacemi získáme lepší představu i o vlastnostech asymptotického rozdělení z druhé kapitoly.

Motivace

Ve světě kolem nás se setkáváme se spoustou procesů, u kterých neznáme přesné zákonitosti, jimiž se řídí. Jejich pozorováním ovšem můžeme získat o jejich chování určitou představu. Člověk od počátku existence zapojuje celou svoji inteligenci, aby svět kolem sebe správně pochopil. Lidský rozum se však také může ve svém nahlížení na skutečnost – a její interpretaci – velmi snadno mýlit. Lidstvo tak vynalezlo a vynalézá bezpočet způsobů k nalezení pravdy, kterou můžeme považovat v co nejlepším smyslu za objektivní. (Bohužel spousta invence směřuje i do vynalézaní náhledů, které jsou pro určité skupiny „objektivnější“).

Statistici využívají ve svých základech matematický aparát. Co může být více objektivní než čistá matematika? Z několika málo axiomů je vytvořen systém, v němž u každého tvrzení můžeme pomocí určité posloupnosti kroků dokázat, zda je pravdivé či nepravdivé. Ano, systém je silný jako jeho axiomy. Ale pokud je přijmeme, nelze již o pravdivosti korektně dokázaného tvrzení pochybovat.

Statistika však aplikuje matematiku v situacích, kdy nemáme všechny informace. Na základě dostupných dat se snaží odhadnout, jaká je skutečnost. Z toho vyplývá, že naše závěry můžeme činit jen s určitou pravděpodobností. Do objevování skutečnosti tak vstupuje subjektivita. Protože je jen na nás a našich zkušenostech, jak si myslíme, že bychom měli daný proces modelovat, a jakou vhodnou statistickou metodu použít.

Statistik má v dnešní době k dispozici mnoho metod, jak svůj specifický problém řešit. Záleží na procesu, kterým se zabývá, a na cíli, jehož chce dosáhnout. Některé metody jsou vhodnější pro popis systému a k odhadu, jakým způsobem se chová, jiné pro predikci budoucnosti. Tato práce se zaměřujeme právě na jednu z možných metod. Cílem je vysvětlit si základní myšlenku, která se za ní skrývá. Popsat si její elementární vlastnosti a získat lepší představu, kdy je metoda využitelná pro reálné situace.

Kapitola 1

Vzorová úloha

Procesy kolem nás jsou často tak složité, že je pro pochopení problému výhodné si komplexní realitu zjednodušit. K tomu si vytváříme modely. Tvoříme je tak, abychom v nich zachytili vše podstatné. A zároveň potřebujeme, aby byly dostatečně jednoduché – ať z důvodu uchopitelnosti nebo z důvodu omezené výpočetní kapacity. V této práci se budeme zabývat modely, v kterých předpokládáme lineární vztahy mezi zkoumanými veličinami.

Budeme se zabývat především lineárními errors-in-variables modely, které budeme porovnávat s klasickými modely lineární regrese. Základní rozdíl mezi těmito dvěma modely si ukážeme na jednoduché úloze v následujícím oddíle.

Nejdříve si však ještě trochu ujasněme názvosloví. Tato diplomová práce má v názvu metodu úplně nejmenších čtverců a před chvíli jsme zmínili, že se budeme zabývat lineárními errors-in-variables modely. Vztah mezi těmito dvěma pojmy je podobný jako mezi lineárním regresním modelem a metodou nejmenších čtverců. Lineární regresní model je spíše označení vztahu mezi zkoumanými veličinami a metoda nejmenších čtverců pak název metody, pomocí které zkoumaný model řešíme. Analogicky lineární errors-in-variables model řešíme metodou úplně nejmenších čtverců. Dvojice souvisejících pojmů jsou však do jisté míry záměnné. Velmi často se především anglické ekvivalenty *Ordinary least squares* (Nejmenší čtverce) a *Total least squares* (Úplně nejmenší čtverce) používají i pro označení zkoumaného modelu.

1.1 Popis úlohy

Uvažujme jednoduchý příklad, kdy dvě veličiny x a y jsou spolu provázány přímou úměrou a budeme uvažovat možný posun mimo počátek souřadnic, tj.

$$y = ax + b, \tag{1.1}$$

kde a a b jsou neznámé koeficienty (parametry). V klasických lineárních regresních modelech se y označuje jako *vysvětlovaná proměnná*, x jako *vysvětlující proměnná* a pro b se používá anglický pojem *intercept*.

Na tomto jednoduchém příkladě si ukážeme základní logiku metody úplně nejmenších čtverců. Zároveň uvidíme nejpodstatnější rozdíl oproti klasické metodě nejmenších čtverců.

Mějme k dispozici 20 pozorování x a y, které označíme $x_1, \ldots, x_{20}$ a $y_1, \ldots, y_{20}$. Tato pozorování se však odchylují od ideálního vztahu (1.1). Naším úkolem je na jejich základě odhadnout co nejpřesněji hodnoty parametrů a a b. Geometricky si můžeme představit náš problém jako prokládání bodů v rovině přímkou. Snažíme se o to, aby daná přímka byla bodům v jistém smyslu co nejblíže.

Důvody, proč se pozorování odchylují od ideální přímky, můžeme rozdělit v podstatě na dvě skupiny. První skupinou jsou chyby či nedostatky měření. Např. nemáme k dispozici dostatečně přesná měřidla; používáme více měřidel, která se od sebe liší; osoba, která provádí měření, je unavená/nedbalá. Druhou skupinou důvodů může být to, že se jednotlivé prvky ve zkoumané populaci od sebe odlišují sami navzájem. Tedy přesný vztah (1.1) mezi zkoumanými veličinami vlastně vůbec neexistuje! Co hledáme je „průměrný“ vztah (v klasickém lineárním modelu předpokládáme, že vztah platí pro střední hodnotu vysvětlované proměnné).

Pro lepší pochopení použijme konkrétní příklad. Budeme uvažovat, že vztah přímé úměry (1.1) nám ukazuje souvislost mezi vahou a výškou člověka. Čím je člověk vyšší, tím je pravděpodobněji těžší (a naopak). Nedostatky měření můžou být způsobeny vahou, která nám ukazuje pouze celé kilogramy. Nebo metrem, jehož nejmenší jednotkou je centimetr. V případě výšky a váhy je také jasné, že neexistuje přímá úměra platná pro všechny lidi – odhadujeme pouze, kolik váží „průměrný“ jedinec, který měří např. 180 cm. Na vztah mezi výškou a vahou konkrétního člověka mají vliv genetické předpoklady a stravovací návyky.

1.2 Řešení metodou nejmenších čtverců

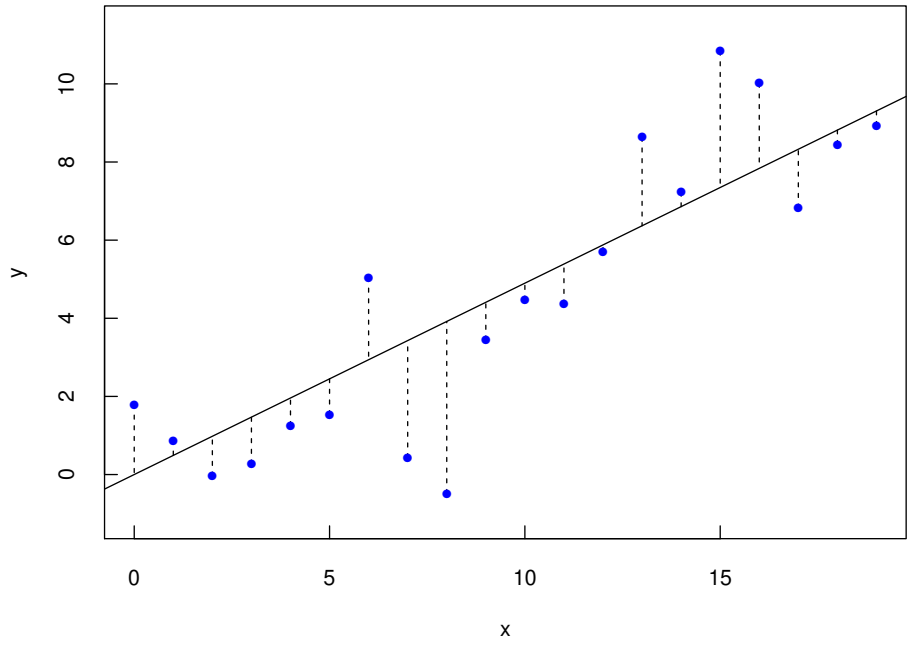
Vraťme se nyní zpět k tomu, jakým způsobem najít hodnotu parametrů a a b. Parametry se snažíme najít tak, aby byla přímka co nejblíže jednotlivým pozorováním (x_i, y_i). Pokud budeme předpokládat, že hodnoty proměnné x jsme schopni získat bez chyb a jediné nepřesnosti vznikají při zjišťování hodnot y, nabízí se měřit vzdálenost přímky a prokládaných dat způsobem ukázaném na obr. 1.1. Protože podstatné jsou odchylky hodnot y_i od přímky. Hodnoty x_i jsme získali přesně.

Poslední ale podstatnou otázkou zůstává, co přesně chápeme pod pojmem vzdálenost přímky a bodů y_i. Logicky se nabízí brát vzdálenost jako

$$d_i = |y_i - (ax_i + b)|.$$

Tedy měřit ji jako rozdíl ypsilonových souřadnic pozorovaného bodu a jeho průmětu na přímku (průmět ve smyslu obr. 1.1).

Hledání řešení pomocí minimalizace součtu těchto vzdáleností by byl validní postup. V nynější době se však používá především minimalizace součtu čtverce



Obrázek 1.1: Nepřesnosti jen v proměnné y.

vzdáleností (odtud i čtverec v názvu této práce). Následkem je, že při použití mocniny se vzdálené body penalizují více – přímka tak bude těmto bodům blíže než v případě použití absolutní hodnoty. Větší vliv těchto tzv. odlehlých pozorování (*outliers*) se v některých případech hodí, v jiných vhodný není. V čem má čtverec vzdáleností nespornou výhodu jsou jeho matematické vlastnosti – derivace druhé mocniny je spojitá, což nám pomůže při hledání minima.

Řešíme minimalizovat součet čtverců vzdáleností (v našem případě $n = 20$)

$$\min \sum_{i=1}^{n} d_i^2 = \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Minima se nabývá v bodě $(\hat{a}, \hat{b})$

$$\begin{aligned} \hat{b} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{a}\bar{x}, \\ \hat{a} &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned} \tag{1.2}$$

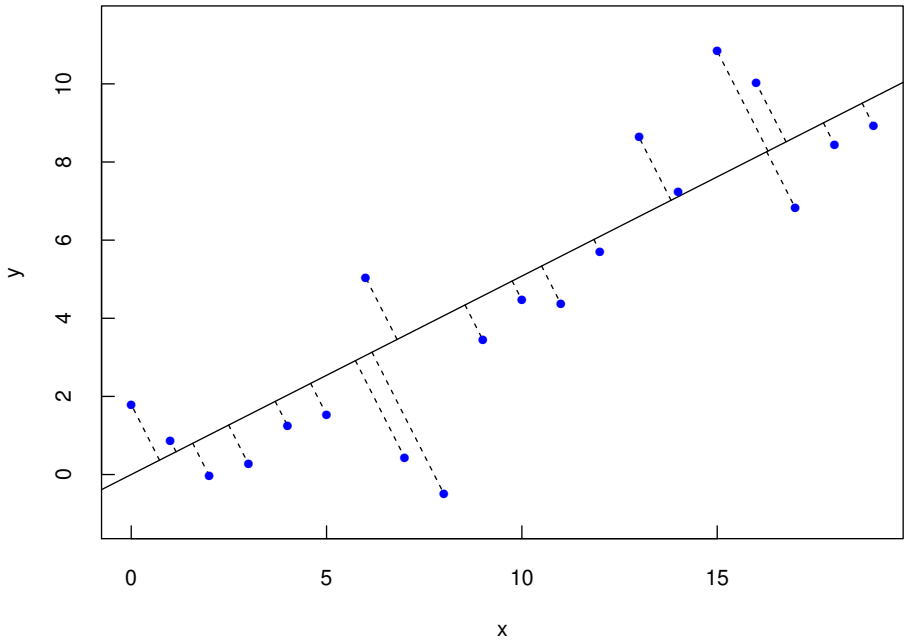
Kde označujeme výběrové průměry

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Řešení $(\hat{a}, \hat{b})$ budeme nazývat *řešení metodou nejmenších čtverců*. Jak jsme k řešení dospěli viz oddíl 1.5.1.

1.3 Řešení metodou úplně nejmenších čtverců

Pokud u vztahu (1.1) předpokládáme chyby měření i u hodnot x, není vhodné upřednostňovat ani jednu proměnnou x či y. Logičtější je dívat se na problém dle obr. 1.2. Nejkratší vzdálenost od pozorování (x_i, y_i) k přímce je od bodu k jeho kolmému průmětu.



Obrázek 1.2: Nepřesnosti v obou proměnných x i y.

A protože i v tomto případě budeme brát čtverec vzdálenosti, hledáme řešení jako

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{1 + a^2}. \tag{1.3}$$

Jak se došlo k výrazu v sumě v (1.3), je zřejmé z obrázku 1.3.

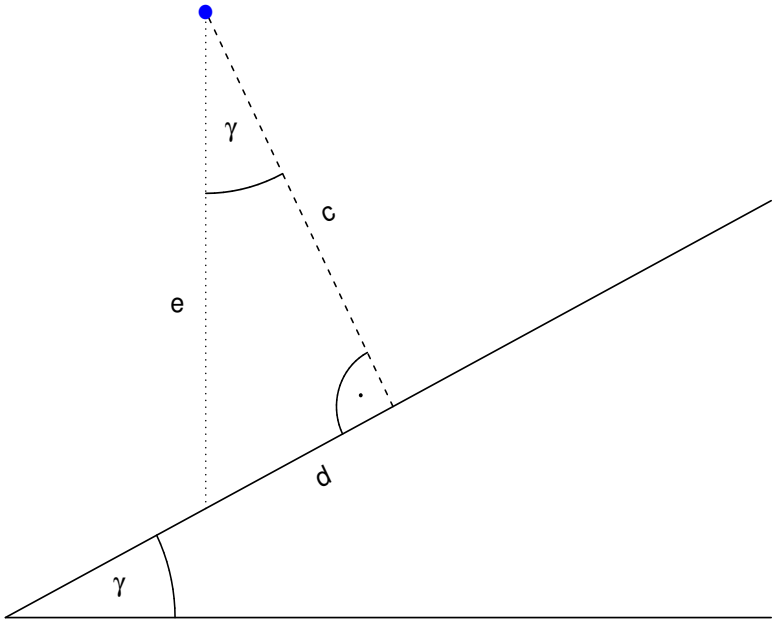
Zajímá nás hodnota c^2, což je strana v pravoúhlém trojúhelníku. Přičemž hodnotu e jsme použili v předchozí analýze. Úhel γ je dán sklonem přímky a. Tedy

$$c^2 = e^2 - d^2, \quad e = (y_i - ax_i - b), \quad \tan\gamma = a = \frac{d}{c}.$$

Dosadíme-li a vyjádříme-li c^2, dospějeme k výrazu v sumě v (1.3).

$$c^2 = (y_i - ax_i - b)^2 - a^2c^2$$

$$c^2 = \frac{(y_i - ax_i - b)^2}{1 + a^2}$$



Obrázek 1.3: Odvození odhadu $\tilde{a}$.

Řešením (1.3) pak je

$$\begin{aligned} \tilde{b} &= \bar{y} - \tilde{a}\bar{x}, \\ \tilde{a} &= \frac{s_y^2 - s_x^2 + \sqrt{(s_x^2 - s_y^2)^2 + 4s_{xy}^2}}{2s_{xy}}. \end{aligned} \tag{1.4}$$

U odhadu $\tilde{a}$ používáme značení pro výběrový rozptyl a výběrovou kovarianci

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \\ s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \\ s_{xy}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \end{aligned}$$

Říkáme, že $\tilde{a}$ a $\tilde{b}$ jsou odhady parametrů a a b získané *metodou úplně nejmenších čtverců*. Odvození viz oddíl 1.5.2.

Poznámka 1.1 (Ortogonální regrese)**.** Protože odchylky jsou kolmé na přímku, kterou prokládáme, mluvíme v situaci z obr. 1.2 také o ortogonální regresi.

1.4 Simulace

Předpokládejme, že správná hodnota parametrů je $a = 0.5$ a $b = 1$. Vztah, který zkoumáme mezi veličinami x a y, je pak konkrétně

$$y = 0.5\,x + 1, \tag{1.5}$$

Jednotlivá pozorování y_i a x_i těchto veličin však neměříme správně. Označme chybu u měření veličiny x jako ν a chybu u veličiny y jako ε. Hvězdičkou označme skutečné hodnoty. Pak

$$x_i = x_i^* + \nu_i$$
$$y_i = y_i^* + \varepsilon_i$$

Dosadíme do vztahu (1.5). Pro jednotlivá pozorování platí

$$y_i = 0.5\,(x_i - \nu_i) + 1 + \varepsilon_i, \quad i = 1, \ldots, n.$$

Hodnoty parametrů a a b však neznáme a odhadujeme je z modelu

$$y_i = ax_i + b + (\varepsilon_i - a\nu_i), \quad i = 1, \ldots, n. \tag{1.6}$$

Podívejme se na vlastnosti odhadů získané výše uvedenými metodami nejmenších čtverců a úplně nejmenších čtverců. Vlastnosti si v této kapitole ukážeme pouze na simulovaných datech. Budeme postupovat následovně:

1. Skutečnou hodnotu x_i^* získáme náhodně jako výběr z rovnoměrného rozdělení na intervalu $(0, 20)$.
2. Hodnotu y_i^* dopočítáme podle vztahu (1.5).
3. Hodnoty x_i^* a y_i^* upravíme o chyby ν_i, ε_i a získáme naměřené hodnoty x_i a y_i.
4. Chyby měření veličin si vygenerujeme náhodně. Budeme předpokládat, že mají nezávislá normální rozdělení se střední hodnotou 0 a rozptylem 2. Symbolicky $\nu \sim \mathsf{N}(0{,}2)$, $\varepsilon \sim \mathsf{N}(0{,}2)$.
5. Zopakujeme body 1-4 pro všechny $i = 1, \ldots, n$.
6. Na základě x_i, y_i, $i = 1, \ldots, n$ pak najdeme OLS odhady $\hat{a}$, $\hat{b}$ a TLS odhady $\tilde{a}$, $\tilde{b}$.
7. Nakonec si podle bodů 1-6 vytvoříme m OLS a TLS odhadů parametrů a, b.

Simulacemi získáme m OLS a m TLS odhadů. Spočítáme si výběrový průměr a výběrovou směrodatnou odchylku těchto odhadů. Na jejich základě budeme usuzovat o vlastnostech odhadu metodou nejmenších čtverců a metodou úplně nejmenších čtverců. Průměr nám napoví o vychýlenosti odhadu, směrodatná odchylka o konzistenci. Budeme uvažovat několik variant n a m

$$n \in (20, 100, 1000, 10000),$$
$$m \in (10, 100, 1000).$$

V tabulce 1.1 najdeme průměrné hodnoty odhadů parametrů a a b v závislosti na hodnotách n a m. Není překvapivé, že odhady metodou nejmenších čtverců jsou vychýlené – jsou porušeny předpoklady, veličina x je v modelu (1.6) korelovaná s chybou. Simulace naopak naznačuje, že TLS odhady vychýlené nejsou (v další části práce si ukážeme, že je tomu skutečně tak).

Tabulka 1.2 nám pak ukazuje výběrové směrodatné odchylky. Je pozitivní, že s přibývajícím počtem pozorování n se hodnoty odchylky zmenšují. Počet simulací m nemá velký vliv na velikost průměrné odchylky – tedy nic nenasvědčujeme tomu, že bychom pro různé náhodné výběry docházeli pravidelně k naprosto odlišným výsledkům.

m\n	20	100	1000	10000
10	0.495	0.467	0.474	0.470
100	0.480	0.475	0.472	0.471
1000	0.472	0.472	0.472	0.472

(a) Průměr OLS odhadů parametru a.

m\n	20	100	1000	10000
10	0.950	1.357	1.297	1.303
100	1.161	1.269	1.273	1.286
1000	1.258	1.282	1.283	1.283

(b) Průměr OLS odhadů parametru b.

m\n	20	100	1000	10000
10	0.532	0.496	0.503	0.498
100	0.509	0.504	0.501	0.500
1000	0.501	0.500	0.500	0.500

(c) Průměr TLS odhadů parametru a.

m\n	20	100	1000	10000
10	0.593	1.060	1.009	1.023
100	0.867	0.980	0.988	1.004
1000	0.964	0.997	1.000	1.000

(d) Průměr TLS odhadů parametru b.

Tabulka 1.1: Průměry odhadů a a b metodou OLS a TLS.

m\n	20	100	1000	10000
10	0.049	0.023	0.009	0.003
100	0.062	0.025	0.010	0.003
1000	0.065	0.026	0.008	0.003

(a) Odchylka OLS odhadů parametru a.

m\n	20	100	1000	10000
10	0.699	0.159	0.114	0.023
100	0.674	0.316	0.108	0.036
1000	0.754	0.303	0.098	0.030

(b) Odchylka OLS odhadů parametru b.

m\n	20	100	1000	10000
10	0.054	0.025	0.010	0.004
100	0.067	0.027	0.010	0.003
1000	0.071	0.028	0.009	0.003

(c) Odchylka TLS odhadů parametru a.

m\n	20	100	1000	10000
10	0.834	0.169	0.122	0.025
100	0.727	0.332	0.115	0.037
1000	0.801	0.321	0.103	0.031

(d) Odchylka TLS odhadů parametru b.

Tabulka 1.2: Směrodatné odchylky odhadů a a b metodou OLS a TLS.

Poznámka 1.2. Všechny náhodné výběry jsme generovali pseudonáhodně v programu R Core Team (2019).

1.5 Rozvinutí témat vzorové úlohy

V této doplňující podkapitole si podíváme podrobněji na některá témata z předchozího textu. Ukážeme, že jsme našli korektně použitá řešení jak metodou nejmenších čtverců, tak metodou úplně nejmenších čtverců. Následně si ještě odvodíme, jak moc jsou nekonzistentní OLS odhady parametrů v modelu (1.6).

1.5.1 Nalezení řešení metodou nejmenších čtverců

Ukažme si, že řešení (1.2) metodou nejmenších čtverců je opravdu minimum funkce

$$\min_{(a,b)\in\mathbb{R}^2} f_{ols}(a,b) = \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2. \tag{1.7}$$

K tomu využijeme znění vět ze skript docenta Lachouta k předmětu Teorie optimalizace (Lachout, 2018).

Věta 1.3 (Globální minimum konvexní funkce)**.** Nechť $\mathcal{M} \subset \mathcal{G} \subset \mathbb{R}^n$. $\mathcal{M}$ je konvexní množina, $\mathcal{G}$ je otevřená konvexní množina, $\tilde{x}$ je vnitřní bod množiny $\mathcal{M}$. Nechť $f : \mathcal{G} \to \mathbb{R}$ je konvexní funkce, která má v bodě $\tilde{x}$ gradient

$$\nabla f(\tilde{x}) = \left(\frac{\partial f(\tilde{x})}{x_1}, \ldots, \frac{\partial f(\tilde{x})}{x_n} \right).$$

Potom

$$\tilde{x} \text{ je globální minimum f na } \mathcal{M} \Leftrightarrow \nabla f(\tilde{x}) = \mathbf{0}.$$

Důkaz. Viz Lachout (2018), Lemma 3.4. □

Označíme si $\mathcal{M} = \mathcal{G} = \mathbb{R}^n$. Ukážeme-li, že funkce $f_{ols}(a,b)$ je konvexní, pak podle věty 1.3 je extrém funkce $f_{ols}(a,b)$ v takovém bodě $(\hat{a},\hat{b})$, kde jsou parciální derivace rovné nule

$$\frac{\partial \sum_{i=1}^{n} (y_i - ax_i - b)^2}{\partial a} \stackrel{!}{=} 0,$$
$$\frac{\partial \sum_{i=1}^{n} (y_i - ax_i - b)^2}{\partial b} \stackrel{!}{=} 0.$$

Derivujeme podle b

$$\begin{aligned} \frac{\partial f_{ols}(a,b)}{\partial b} &= \frac{\partial \sum_{i=1}^{n} (y_i - ax_i - b)^2}{\partial b} \\ &= \sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 2nb + 2a\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i. \end{aligned}$$

Položíme rovno 0.

$$0 = 2n\hat{b} + 2\hat{a}\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i$$
$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i$$

Nyní budeme derivovat podle a

$$\begin{aligned}\frac{\partial f_{ols}(a,b)}{\partial a} &= \frac{\partial \sum_{i=1}^{n}(y_i - ax_i - b)^2}{\partial a} \\ &= \sum_{i=1}^{n} 2(y_i - ax_i - b)(-x_i) = 2b\sum_{i=1}^{n} x_i + 2a\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} y_i x_i.\end{aligned}$$

Dosadíme vyjádřené $\hat{b}$ a opět položíme rovno 0.

$$0 = 2\hat{b}\sum_{i=1}^{n} x_i + 2\hat{a}\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} y_i x_i$$
$$0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{a}\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i$$
$$0 = \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i + \hat{a}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\right)$$
$$\hat{a} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

Tím jsme dospěli k řešení (1.2). Zbývá dokázat, že $f_{ols}(a,b)$ je konvexní funkce. K tomu použijeme další větu z Lachout (2018), která ukazuje vztah mezi konvexitou a Hessovou maticí.

Definice 1.4 (Hessova matice)**.** Nechť $\mathcal{G} \subset \mathbf{R}^n$. $f : \mathcal{G} \to \mathbb{R}$ a $x \in \mathcal{G}$. Pak *Hessovou maticí* funkce f v bodě x budeme nazývat (pokud má výraz smysl)

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix}.$$

Její determinant pak nazýváme *Hessián*.

Věta 1.5 (Konvexita a Hessova matice)**.** Nechť $\mathcal{D} \subset \mathbb{R}^n, \mathcal{D} \neq \emptyset$ je otevřená konvexní množina a $\nabla^2 f(x)$ je spojitá na $\mathcal{D}$. Pak

$$\text{f je konvexní} \Leftrightarrow \nabla^2 f(x) \text{ je pozitivně semidefinitní matice pro } \forall x \in \mathcal{D}.$$

Důkaz. Viz Lachout (2018), Lemma 2.28 + Věta 2.49. □

Druhé parciální derivace vypadají následovně

$$\begin{aligned}
\frac{\partial^2 f_{ols}(a,b)}{\partial b^2} &= \frac{\partial \left(2nb + 2a\sum_{i=1}^n x_i - 2\sum_{i=1}^n y_i\right)}{\partial b} = 2n, \\
\frac{\partial^2 f_{ols}(a,b)}{\partial b\,\partial a} &= \frac{\partial \left(2nb + 2a\sum_{i=1}^n x_i - 2\sum_{i=1}^n y_i\right)}{\partial a} = 2\sum_{i=1}^n x_i, \\
\frac{\partial^2 f_{ols}(a,b)}{\partial a^2} &= \frac{\partial \left(2b\sum_{i=1}^n x_i + 2a\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n y_i x_i\right)}{\partial a} = 2\sum_{i=1}^n x_i^2.
\end{aligned}$$

A Hessova matice funkce $f_{ols}(a,b)$

$$\nabla^2 f_{ols}(a,b) = \begin{pmatrix} 2\sum_{i=1}^n x_i^2 & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2n \end{pmatrix} \tag{1.8}$$

je spojitá pro $\forall a \in \mathbb{R}$ a $\forall b \in \mathbb{R}$.

Zbývá ukázat, že $\nabla^2 f(a,b)$ je pozitivně semidefinitní, k čemuž využijeme Sylvestrovo kritérium.

Věta 1.6 (Sylvestrovo kritérium)**.** Nechť $\boldsymbol{A}_{n\times n}$ je symetrická matice. Označme její prvky $a_{i,j}$, $i = 1, \ldots, n, j = 1, \ldots, n$. Pak

(i) $\boldsymbol{A}$ je pozitivně semidefinitní právě tehdy, když platí pro každou k-tici přirozených čísel $1 \leq i_1 < \cdots < i_k \leq n$, $k \in \{1, \ldots, n\}$

$$\begin{vmatrix} a_{i_1,i_1} & a_{i_1,i_2} & \ldots & a_{i_1,i_k} \\ a_{i_2,i_1} & a_{i_2,i_2} & \ldots & a_{i_2,i_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_k,i_1} & a_{i_k,i_2} & \ldots & a_{i_k,i_k} \end{vmatrix} \geq 0.$$

(ii) $\boldsymbol{A}$ je pozitivně definitní právě tehdy, když platí pro každé $k \in (1, \ldots, n)$

$$\begin{vmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,k} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \ldots & a_{k,k} \end{vmatrix} > 0.$$

Důkaz. Viz Horn a Johnson (2012), Věta 7.2.5. □

Aplikujeme-li Sylvestrovo kritérium na Hessovu matici (1.8), zjistíme, že je pozitivně semidefinitní. Neboť

$$
\begin{aligned}
&i) \quad 2\sum_{i=1}^{n} x_i^2 \geq 0,\\
&ii) \quad 2n \geq 0,\\
&iii) \quad 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 = 4n\sum_{i=1}^{n} x_i^2 - 4\cdot 2\left(\sum_{i=1}^{n} x_i\right)^2 + 4\left(\sum_{i=1}^{n} x_i\right)^2\\
&\qquad = 4n\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \geq 0.
\end{aligned}
$$

Podle věty 1.3 je tak řešení $\hat{a}, \hat{b}$ metodou nejmenších čtverců globálním minimem (1.7).

Poznámka 1.7. Výše jsme odvodili OLS řešení minimalizací (1.7). Ke stejnému řešení lze dojí i jinými metodami. Viz např. Wooldridge (2008), kde je uvedeno, jak dospět ke stejnému řešení pomocí momentové metody či metodou maximální věrohodnosti.

1.5.2 Nalezení řešení metodou úplně nejmenších čtverců

Opět ukážeme, že řešení (1.4) metodou úplně nejmenších čtverců je opravdu minimum funkce

$$\min_{(a,b)\in\mathbb{R}^2} f_{tls}(a,b) = \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{1+a^2}. \tag{1.9}$$

(V této kapitole si pouze ukážeme, že se jedná o lokální minimum. Že se jedná taktéž o globální minimum uvidíme z obecného řešení v následující kapitole.)

Řešení (1.3) pak opět nalezneme pomocí parciálních derivací. Nejdříve budeme derivovat podle b

$$\frac{\partial f_{tls}(a,b)}{\partial b} = \sum_{i=1}^{n} \frac{2(y_i - ax_i - b)(-1)}{1+a^2} \overset{!}{=} 0.$$

Tedy

$$
\begin{aligned}
&\frac{1}{1+a^2}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb\right) = 0\\
\implies \quad &\tilde{b} = \bar{y} - \tilde{a}\bar{x}.
\end{aligned}
$$

Poznámka 1.8. Můžeme si všimnout, že řešení $\tilde{b}$ je ve stejném vztahu s $\tilde{a}$ jako u odhadu nejmenších čtverců.

Derivace podle a je

$$\frac{\partial f_{tls}(a,b)}{\partial a} = \sum_{i=1}^{n} \frac{2(y_i - ax_i - b)(-x_i)(1+a^2) - (y_i - ax_i - b)^2(2a)}{(1+a^2)^2} \stackrel{!}{=} 0$$

$$\begin{aligned}
\implies 0 = \sum_{i=1}^{n} [&-x_i y_i - a^2 x_i y_i + a x_i^2 + \cancel{a^3 x_i^2} + b x_i + a^2 b x_i + \\
&- a y_i^2 + 2a^2 x_i y_i + 2ab y_i - \cancel{a^3 x_i^2} - 2a^2 b x_i - ab^2] \\
= \sum_{i=1}^{n} &-x_i y_i + a^2 \sum_{i=1}^{n} x_i y_i + a \sum_{i=1}^{n} (x_i^2 - y_i^2) + b \sum_{i=1}^{n} x_i + \\
&- a^2 b \sum_{i=1}^{n} x_i + 2ab \sum_{i=1}^{n} y_i - nab^2.
\end{aligned}$$

Do výrazu dosadíme za b

$$\begin{aligned}
0 = &- \sum_{i=1}^{n} x_i y_i + a^2 \sum_{i=1}^{n} x_i y_i + a \sum_{i=1}^{n} (x_i^2 - y_i^2) + \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \\
&- a \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 - a^2 \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \cancel{a^3 \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2} + \\
&+ 2a \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 \cancel{- 2a^2 \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i} - a \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 + \\
&\cancel{+ 2a^2 \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i} \cancel{- a^3 \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2}.
\end{aligned}$$

Po dalších úpravách

$$\begin{aligned}
0 = a^2 &\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right] - \left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right] + \\
&+ a \left[\sum_{i=1}^{n} (x_i^2 - y_i^2) + \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right] \\
0 = a^2 &\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + a \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 - \sum_{i=1}^{n} (y_i - \bar{y})^2 \right] + \\
&- \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).
\end{aligned}$$

Označíme-li

$$\begin{aligned}
A &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \\
B &= \sum_{i=1}^{n} (x_i - \bar{x})^2 - \sum_{i=1}^{n} (y_i - \bar{y})^2,
\end{aligned} \tag{1.10}$$

pak řešeními kvadratické rovnice jsou

$$\begin{aligned} a_1 &= \frac{-B + \sqrt{B^2 + 4A^2}}{2A}, \\ a_2 &= \frac{-B - \sqrt{B^2 + 4A^2}}{2A}. \end{aligned} \tag{1.11}$$

Máme dva adepty pro řešení $\tilde{a}$. Ukážeme si, že v a_1 je lokální minimum funkce $f_{tls}(a,b)$.

Věta 1.9 (Ostré lokální minimum)**.** Nechť $f : \mathbb{R}^n \to \mathbb{R}$ a $\tilde{x} \in \mathbb{R}^n$. Pokud je funkce f dvakrát diferencovatelná na nějakém okolí bodu $\tilde{x}$, druhé derivace jsou spojité v $\tilde{x}$, $\nabla f(\tilde{x}) = 0$ a Hessova matice je pozitivně definitní, pak má funkce f v bodě $\tilde{x}$ ostré lokální minimum.

Důkaz. Viz Lachout (2018), Lemma 2.28 + Věta 3.12. □

Podíváme se, jak vypadají druhé parciální derivace funkce $f_{tls}(a,b)$.

$$\begin{aligned} \frac{\partial^2 f_{tls}(a,b)}{\partial b^2} &= \frac{\partial \left(\frac{2nb + 2a\sum_{i=1}^n x_i - 2\sum_{i=1}^n y_i}{1+a^2} \right)}{\partial b} = \frac{2n}{1+a^2} \\ \frac{\partial^2 f_{tls}(a,b)}{\partial b\, \partial a} &= \frac{\partial \left(\frac{2nb + 2a\sum_{i=1}^n x_i - 2\sum_{i=1}^n y_i}{1+a^2} \right)}{\partial a} \\ &= \frac{-2a(2nb - 2\sum_{i=1}^n y_i)}{(1+a^2)^2} + \frac{(1-a^2)2\sum_{i=1}^n x_i}{(1+a^2)^2} \\ &= \frac{-2a^2\sum_{i=1}^n x_i + 4a\sum_{i=1}^n y_i + 2\sum_{i=1}^n x_i - 4nab}{(1+a^2)^2} \end{aligned}$$

Druhá parciální derivace podle ∂a^2 je složitější na zápis

$$\begin{aligned} \frac{\partial f_{tls}(a,b)}{\partial a} &= \frac{2\sum_{i=1}^n [-x_i y_i + a^2 x_i y_i + a x_i^2 + b x_i - a^2 b x_i - a y_i^2 + 2ab y_i - ab^2]}{1+a^2}, \\ \frac{\partial^2 f_{tls}(a,b)}{\partial a^2} &= \frac{2\sum_{i=1}^n (2a x_i y_i + x_i^2 - 2ab x_i - y_i^2 + 2b y_i - b^2)(1+a^2)}{(1+a^2)^2} - \\ &\quad - \frac{2\sum_{i=1}^n (-2a x_i y_i + 2a^3 x_i y_i + 2a^2 x_i^2 + 2ab x_i - 2a^3 b x_i)}{(1+a^2)^2} - \\ &\quad - \frac{2\sum_{i=1}^n (-2a^2 y_i^2 + 4a^2 b y_i - 2a^2 b^2)}{(1+a^2)^2} \\ &= \frac{2\sum_{i=1}^n (4a x_i y_i + x_i^2 - y_i^2 - 4ab x_i + 2b y_i - b^2 - a^2 x_i^2)}{(1+a^2)^2} + \\ &\quad + \frac{2\sum_{i=1}^n (a^2 y_i^2 - 2a^2 b y_i + a^2 b^2)}{(1+a^2)^2}. \end{aligned}$$

Hessova matice funkce $f_{tls}(a,b)$

$$\nabla^2 f_{tls}(a,b) = \begin{pmatrix} \frac{\partial^2 f_{tls}(a,b)}{\partial a^2} & \frac{\partial^2 f_{tls}(a,b)}{\partial b \partial a} \\ \frac{\partial^2 f_{tls}(a,b)}{\partial a \partial b} & \frac{\partial^2 f_{tls}(a,b)}{\partial b^2} \end{pmatrix} \tag{1.12}$$

je tak spojitá pro $\forall a \in \mathbb{R}$ a $\forall b \in \mathbb{R}$. (Jedná se o aritmetické operace spojitých funkcí, přičemž funkce ve jmenovateli nenabývá na $\mathbb{R}^2$ hodnoty 0.)

Dosaďme do Hessovy matice $b = \bar{y} - a\bar{x}$. Pak parciální derivace podle a a b se nám zjednoduší do

$$\begin{aligned} \frac{\partial^2 f_{tls}(a,b)}{\partial b\, \partial a} &= \frac{-2a^2 \sum_{i=1}^n x_i + 4a \sum_{i=1}^n y_i + 2 \sum_{i=1}^n x_i - 4a \sum_{i=1}^n y_i + 4a^2 \sum_{i=1}^n x_i}{(1+a^2)^2} \\ &= \frac{2a^2 \sum_{i=1}^n x_i + 2 \sum_{i=1}^n x_i}{(1+a^2)^2} \\ &= \frac{2 \sum_{i=1}^n x_i}{1+a^2}. \end{aligned}$$

A

$$\begin{aligned} \frac{\partial^2 f_{tls}(a,b)}{\partial a^2} &= \frac{8a \sum_{i=1}^n x_i y_i + 2 \sum_{i=1}^n (x_i^2 - y_i^2) + 2a^2 \sum_{i=1}^n (y_i^2 - x_i^2)}{(1+a^2)^2} + \\ &+ \frac{-8a \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i + 8a^2 \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2 + 4 \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2}{(1+a^2)^2} + \\ &+ \frac{-4a \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i - 2 \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2 + 4a \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{(1+a^2)^2} + \\ &+ \frac{-2a^2 \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2 - 4a^2 \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2 + 4a^3 \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{(1+a^2)^2} + \\ &+ \frac{2a^2 \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2 - 4a^3 \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i + 2a^4 \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2}{(1+a^2)^2}. \end{aligned}$$

Neboli

$$\begin{aligned} \frac{\partial^2 f_{tls}(a,b)}{\partial a^2} &= \frac{8a \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + 2a^2 \sum_{i=1}^n (y_i - \bar{y})^2 - 2a^2 \sum_{i=1}^n (x_i - \bar{x})^2}{(1+a^2)^2} \\ &+ \frac{-2 \sum_{i=1}^n (y_i - \bar{y})^2 + 2 \sum_{i=1}^n x_i^2 + 4a^2 \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2 + 2a^4 \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2}{(1+a^2)^2}. \end{aligned}$$

Ve značení (1.10) můžeme přepsat

$$\frac{\partial^2 f_{tls}(a,b)}{\partial a^2} = \frac{8aA - 2a^2 B + 2B + \frac{2}{n} \left(\sum_{i=1}^n x_i\right)^2 + a^2 \frac{4}{n} \left(\sum_{i=1}^n x_i\right)^2 + a^4 \frac{2}{n} \left(\sum_{i=1}^n x_i\right)^2}{(1+a^2)^2}.$$

Ukážeme, že $\nabla^2 f_{tls}(a,b)$ je v bodě $(a_1,\tilde{b})$ pozitivně definitní. Respektive z důvodu větší aritmetické jednoduchosti si prohodíme souřadnice a budeme ukazovat pozitivní definitnost $\nabla^2 f_{tls}(b,a)$.

Podle Sylvestrova kritéria (Věta 1.6) musíme ukázat, že

(i) $\frac{\partial^2 f_{tls}(b,a)}{\partial b^2} > 0$,

(ii) Hessián $f_{tls}(a,b)$ je větší než nula.

První bod je zřejmý, protože

$$\frac{\partial^2 f_{tls}(b,a)}{\partial b^2} = \frac{2n}{1+a^2} > 0$$

pro všechna $(b,a) \in \mathbb{R}^2$.

Druhý bod znamená ukázat, že

$$H = \left|\nabla^2 f_{tls}(\tilde{b},a_1)\right| = \frac{\partial^2 f_{tls}(\tilde{b},a_1)}{\partial b^2} \cdot \frac{\partial^2 f_{tls}(\tilde{b},a_1)}{\partial a^2} - \left[\frac{\partial^2 f_{tls}(\tilde{b},a_1)}{\partial b \partial a}\right]^2 > 0.$$

Dosadíme

$$\begin{aligned}
H = &\frac{2n}{1+a_1^2}\frac{8a_1A - 2a_1^2B + 2B}{(1+a_1^2)^2} + \\
&+ \frac{2n}{1+a_1^2}\frac{\frac{2}{n}\left(\sum_{i=1}^n x_i\right)^2 + a_1^2\frac{4}{n}\left(\sum_{i=1}^n x_i\right)^2 + a_1^4\frac{2}{n}\left(\sum_{i=1}^n x_i\right)^2}{(1+a_1^2)^2} + \\
&- \frac{4\left(\sum_{i=1}^n x_i\right)^2}{(1+a_1^2)^2} \\
= &\underbrace{\frac{4n}{(1+a_1^2)^3}}_{U} \cdot \underbrace{(4a_1A - a_1^2B + B)}_{V} + \\
&+ \underbrace{\frac{4}{(1+a_1^2)^3}}_{W} \underbrace{\left[a_1^2\left(\sum_{i=1}^n x_i\right)^2 + a_1^4\left(\sum_{i=1}^n x_i\right)^2\right]}_{Z}.
\end{aligned}$$

Zřejmě $U > 0, W > 0$ a $Z \geq 0$. Tedy pokud $V > 0$, pak i $H > 0$.

Dosadíme do V kořen a_1

$$\begin{aligned}
V = (4a_1A - a_1^2B + B) &= 2(\sqrt{B^2+4A^2} - B) + B - \frac{(\sqrt{B^2+4A^2} - B)^2 B}{4A^2} \\
&= 2(\sqrt{B^2+4A^2} - B) + \frac{4BA^2 - B^3 - 4BA^2 + 2B^2\sqrt{B^2+4A^2} - B^3}{4A^2} \\
&= 2(\sqrt{B^2+4A^2} - B) + \frac{B^2(\sqrt{B^2+4A^2} - B)}{2A^2}.
\end{aligned}$$

A protože pro a_1 uvažujeme kladný kořen (1.11) kvadratické rovnice, můžeme psát

$$\sqrt{B^2 + 4A^2} - B > \sqrt{B^2} - B \geq 0 \Rightarrow V > 0.$$

Poznámka 1.10. Proč jsme mohli předpokládat v předchozí rovnici, že

$$A = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \neq 0 \quad ?$$

Nejen že pro $A = 0$ se vůbec nejedná o kvadratickou rovnici, ale A nám i ukazuje korelaci mezi y a x. Pokud jsou veličiny nekorelované, pak náš předpokládaný vztah přímé úměrnosti nedává smysl.

Podívejme se nyní na druhý kořen rovnice a_2. Ukážeme, že snadno najdeme příklady, kdy $\nabla^2 f_{tls}(\tilde{b},a_2)$ nebude pozitivně semidefinitní. Což podle následující věty bude znamenat, že v bodě $(\tilde{b}, a_2)$ není lokální minimum.

Věta 1.11 (Nutná podmínka lokálního minima)**.** Nechť $f : \mathbb{R}^n \to \mathbb{R}$ a $\tilde{x} \in \mathbb{R}^n$. Pokud je funkce f dvakrát diferencovatelná na nějakém okolí bodu $\tilde{x}$, druhé derivace jsou spojité v $\tilde{x}$ a $\tilde{x}$ je lokální minimum funkce f na $\mathbb{R}^n$, pak $\nabla f(\tilde{x}) = 0$ a Hessova matice je pozitivně semidefinitní.

Důkaz. Viz Lachout (2018), Lemma 2.28 + Věta 3.11. □

Uvažujme shodné vektory pozorování $\boldsymbol{x} = (-0{,}5; -0{,}4; \ldots 0{,}4; 0{,}5)^T = \boldsymbol{y}$. Pak Hessova matice není v $(\tilde{b}, a_2)$ pozitivně semidefinitní, neboť Hessián

$$H' = \left|\nabla^2 f_{tls}(\tilde{b}, a_2)\right| < 0.$$

Podle věty 1.11 tak není v $(\tilde{b},a_2)$ lokální minimum.

Závěrem. Ukázali jsme, že lokální minimum funkce (1.3) je v bodech

$$\tilde{b} = \bar{y} - \tilde{a}\bar{x},$$
$$\tilde{a} = \frac{-B + \sqrt{B^2 + 4A^2}}{2A},$$

což je ekvivalentní uvedenému TLS řešení (1.4). V tomto oddíle jsme si ukázali, které řešení kvadratické rovnice (1.11) je preferované (zda a_1 nebo a_2). O globálním minimu více v následující kapitole.

1.5.3 Nekonzistentnost odhadu metodou nejmenších čtverců

V situaci, kdy měříme veličinu x s chybou, je OLS odhad nekonzistentní. V tomto oddíle si ukážeme velikost této nekonzistence za situace z kapitoly 1.4.

Odhadujeme parametry v modelu

$$y_i^* = ax_i^* + b, \quad i = 1, \ldots, n.$$

Místo x_i^*, y_i^* pozorujeme však hodnoty x_i, y_i, které jsou zatížené chybami měření.

$$x_i = x_i^* + \nu_i$$
$$y_i = y_i^* + \varepsilon_i$$

Předpokládali jsme, že chyby měření jsou normálně rozdělené $\nu \sim \mathsf{N}(0{,}2)$, $\varepsilon \sim \mathsf{N}(0{,}2)$. Skutečné hodnoty x_i^* jsou pak náhodné veličiny, které mají rovnoměrné rozdělení $x_i^* \sim \mathsf{R}(0{,}20)$. Všechny náhodné veličiny jsou nezávislé. (Skutečné hodnoty parametrů pak byly $a = 0{,}5$ a $b = 1$.)

Podívejme se na slabou konzistenci OLS odhadů $\hat{a}$, $\hat{b}$. Tedy nás bude zajímat, jakou mají odhady limitu v pravděpodobnosti. (Viz definice 2.29.)

Upravme nejdříve odhad $\hat{a}$

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(ax_i^* + b + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$= \frac{\frac{1}{n}\sum_{i=1}^n (x_i^* + \nu_i - \frac{1}{n}\sum_{i=1}^n x_i^* - \frac{1}{n}\sum_{i=1}^n \nu_i)(ax_i^* + \varepsilon_i)}{\frac{1}{n}\sum_{i=1}^n (x_i^* + \nu_i - \frac{1}{n}\sum_{i=1}^n x_i^* - \frac{1}{n}\sum_{i=1}^n \nu_i)^2}.$$

Po roznásobení závorek dostaneme

$$\begin{aligned}\hat{a} = {} & \frac{a\frac{1}{n}\sum_{i=1}^n {x_i^*}^2 + a\frac{1}{n}\sum_{i=1}^n x_i^*\nu_i - a\left(\frac{1}{n}\sum_{i=1}^n x_i^*\right)^2}{\frac{1}{n}\sum_{i=1}^n {x_i^*}^2 + \frac{2}{n}\sum_{i=1}^n x_i^*\nu_i - \left(\frac{1}{n}\sum_{i=1}^n x_i^*\right)^2} \cdots \\ & \cdots \frac{-a\left(\frac{1}{n}\sum_{i=1}^n x_i^*\right)\left(\frac{1}{n}\sum_{i=1}^n \nu_i\right) + \frac{1}{n}\sum_{i=1}^n x_i^*\varepsilon_i + \frac{1}{n}\sum_{i=1}^n \nu_i\varepsilon_i}{-2\left(\frac{1}{n}\sum_{i=1}^n x_i^*\right)\left(\frac{1}{n}\sum_{i=1}^n \nu_i\right) + \frac{1}{n}\sum_{i=1}^n {\nu_i}^2 - \left(\frac{1}{n}\sum_{i=1}^n \nu_i\right)^2} \cdots \\ & \cdots \frac{-\left(\frac{1}{n}\sum_{i=1}^n x_i^*\right)\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right) - \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right)\left(\frac{1}{n}\sum_{i=1}^n \nu_i\right)}{-\left(\frac{1}{n}\sum_{i=1}^n \nu_i\right)^2 + \left(\frac{1}{n}\sum_{i=1}^n \nu_i\right)^2}. \end{aligned} \tag{1.13}$$

Následující věta nám umožní si „rozkouskovat“ limitu v pravděpodobnosti odhadu na limity jednotlivých členů ve zlomku. (Je nulová pravděpodobnost, že jmenovatel nabývá hodnoty nula.)

Věta 1.12 (Continuous mapping theorem)**.** Nechť $\mathbf{X}, \mathbf{X}_n, n \in \mathbb{N}$ jsou k-rozměrné reálné náhodné vektory. Funkce $g : \mathbb{R}^k \to \mathbb{R}^m$ je spojitá na otevřené množině $C \subset \mathbb{R}^k$ takové, že $\mathsf{P}(\mathbf{X} \in C) = 1$. Pak

$$\mathbf{X}_n \xrightarrow[n\to\infty]{\mathsf{P}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow[n\to\infty]{\mathsf{P}} g(\mathbf{X})$$

Důkaz. Viz van der Vaart (2000), Věta 2.3. □

Pro konvergenci v pravděpodobnosti členů zlomku pak použijeme slabý zákon velkých čísel.

Věta 1.13 (Slabý zákon velkých čísel)**.** Je-li $\{X_n\}$ posloupnost nezávislých náhodných veličin taková, že

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathsf{var}\, X_i = 0,$$

pak

$$\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}\, X_i \right) \xrightarrow{\mathsf{P}} 0.$$

Důkaz. Pomocí Čebyšovovy nerovnosti, viz Dupač a Hušková (2009). □

Důsledek 1.14. Platí-li předpoklady předchozí věty a veličiny X_i jsou stejně rozdělené, pak

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\mathsf{P}} \mathsf{E}\, X_1.$$

Důkaz. Důkaz je zřejmý. □

Podívejme se nyní na jednotlivé členy zlomku (1.13).

(i) $\underline{\left(\frac{1}{n} \sum_{i=1}^{n} x_i^*\right)^2}$

Protože je druhá mocnina spojitá funkce, můžeme se podle věty 1.12 zaměřit pouze na $\frac{1}{n} \sum_{i=1}^{n} x_i^*$.

Víme, že náhodné veličiny $x_i^*, i \in \mathbb{N}$ jsou rovnoměrně rozdělené na intervalu (0,20). Tedy

$$\mathsf{var}\, x_1^* = \frac{20^2}{12} \Rightarrow \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathsf{var}\, x_i^* = 0.$$

Vidíme, že podmínka věty 1.13 je splněna. A proto

$$\frac{1}{n} \sum_{i=1}^{n} x_i^* \xrightarrow{\mathsf{P}} \mathsf{E}\, x_1^* = 10.$$

Závěrem

$$\left(\frac{1}{n} \sum_{i=1}^{n} x_i^* \right)^2 \xrightarrow{\mathsf{P}} (\mathsf{E}\, x_1^*)^2 = 100.$$

(ii) $\underline{\frac{1}{n}\sum_{i=1}^{n} x_i^* \nu_i}$

Náhodné veličiny x_i^* a $\nu_i, i \in \mathbb{N}$ jsou nezávislé. Rozptyl jejich součinu $x_i^* \nu_i$ zjistíme podle následující věty.

Věta 1.15. Nechť X a Y jsou nezávislé náhodné veličiny. Potom

$$\mathsf{var}\, XY = (\mathsf{var}\, X)(\mathsf{var}\, Y) + (\mathsf{var}\, X)(\mathsf{E}\, Y)^2 + (\mathsf{var}\, Y)(\mathsf{E}\, X)^2$$

Důkaz. Protože X a Y jsou nezávislé, tak i X^2 a Y^2 jsou nezávislé. Pak

$$\begin{aligned}
\mathsf{var}\, XY &= \mathsf{E}\, X^2Y^2 - (\mathsf{E}\, XY)^2 \\
&= \mathsf{E}\, X^2\, \mathsf{E}\, Y^2 - (\mathsf{E}\, X\, \mathsf{E}\, Y)^2 \\
&= \left(\mathsf{var}\, X + (\mathsf{E}\, X)^2\right)\left(\mathsf{var}\, Y + (\mathsf{E}\, Y)^2\right) - (\mathsf{E}\, X)^2(\mathsf{E}\, Y)^2 \\
&= (\mathsf{var}\, X)(\mathsf{var}\, Y) + (\mathsf{var}\, X)(\mathsf{E}\, Y)^2 + (\mathsf{var}\, Y)(\mathsf{E}\, X)^2.
\end{aligned}$$

□

Tedy rozptyl si vyjádříme následovně.

$$\begin{aligned}
\mathsf{var}\, x_1^*\nu_1 &= \mathsf{var}\, x_1^*\, \mathsf{var}\, \nu_1 + \mathsf{var}\, x_1^*\, \mathsf{E}\, \nu_1^2 + \mathsf{E}\, x_1^{*2}\, \mathsf{var}\, \nu_1 \\
&= \frac{20^2}{12} \cdot 2 + \frac{20^2}{12} \cdot 2 + \left(\frac{20^2}{12} + 100\right) \cdot 2 \qquad (1.14)
\end{aligned}$$

Pro nalezení limity v pravděpodobnosti zkoumané sumy pak znovu aplikujeme větu 1.13 (respektive její důsledek 1.14).

Podle (1.14) je podmínka věty splněna a tedy za použití nezávislosti veličin

$$\frac{1}{n}\sum_{i=1}^{n} x_i^*\nu_i \xrightarrow{\mathsf{P}} \mathsf{E}\, x_i^*\nu_i = \mathsf{E}\, x_i^*\, \mathsf{E}\, \nu_i = 0.$$

(iii) Podobně bychom postupovali pro všechny členy zlomku (1.13). Většina z nich má limitu v pravděpodobnosti rovnou nule.

Z bodů (i)-(iii) pak dostaneme

$$\hat{a} \xrightarrow{\mathsf{P}} \frac{a\, \mathsf{E}\, x_1^{*2} - a(\mathsf{E}\, x_1^*)^2}{\mathsf{E}\, x_1^{*2} - (\mathsf{E}\, x_1^*)^2 + \mathsf{E}\, \nu_1^2} = a \cdot \frac{\mathsf{var}\, x_1^*}{\mathsf{var}\, x_1^* + \mathsf{var}\, \nu_1} = 0{,}5 \cdot \frac{\frac{100}{3}}{\frac{100}{3} + 2} = 0{,}472.$$

OLS odhad tak správnou hodnotu parametru $a = 0{,}5$ podhodnocuje. V předchozí rovnici si můžeme všimnout, čím je dána výše podhodnocení. Závisí na velikosti $\mathsf{var}\, \nu$ – což odpovídá tomu, jak velkých nepřesností se dopouštíme při měření veličiny x^*. Vidíme, že rovnice odpovídá naší intuitivní představě. Čím je chyba měření veličiny x^* větší, tím je OLS odhad nepřesnější.

Podívejme se nyní podrobněji na odhad $\hat{b}$

$$\begin{aligned}\hat{b} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i^* + b + \varepsilon_i) - \hat{a}\frac{1}{n}\sum_{i=1}^{n}(x_i^* + \nu_i) \\ &= b + a\frac{1}{n}\sum_{i=1}^{n} x_i^* + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i^* - \hat{a}\frac{1}{n}\sum_{i=1}^{n}\nu_i^*.\end{aligned}$$

Za použití stejného postupu jako u odhadu $\hat{a}$ najdeme limitu v pravděpodobnosti.

$$\hat{b} \xrightarrow{\mathsf{P}} b + a\,\mathsf{E}\,x_1^* + 0 - a\cdot\mathsf{E}\,x_1^*\cdot\frac{\mathsf{var}\,x_1^*}{\mathsf{var}\,x_1^* + \mathsf{var}\,\nu_1} - 0 = b + a\cdot\mathsf{E}\,x_1^*\cdot\frac{\mathsf{var}\,\nu_1}{\mathsf{var}\,x_1^* + \mathsf{var}\,\nu_1}$$

$$\hat{b} \xrightarrow{\mathsf{P}} 1{,}283$$

Vidíme, že i OLS odhad $\hat{b}$ je nekonzistentní.

Teoreticky vypočítané limity v pravděpodobnosti odhadů $\hat{a}$ a $\hat{b}$ odpovídají hodnotám, které jsme nasimulovali v kapitole 1.4 (tabulka 1.1).

Kapitola 2

Základní pojmy a vlastnosti

2.1 Úvod do problematiky

Tato práce se zabývá modely, u kterých předpokládáme lineární vztahy mezi zkoumanými proměnnými. Tedy pokud vysvětlujeme proměnnou y^* pomocí proměnných $x_1^*, \ldots, x_k^*$ předpokládáme, že

$$y^* = \boldsymbol{x}^{*\mathsf{T}}\boldsymbol{\beta},$$

kde $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)^\mathsf{T}$ označujeme jako vektor *vysvětlujících* proměnných a y^* jako *vysvětlovanou* proměnnou. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\mathsf{T}$ je vektor čísel, označován jako *parametry* modelu.

Respektive předpokládáme lineární vztah pro libovolné transformace jednotlivých zkoumaných veličin. Tedy pro libovolné funkce

$$\begin{aligned} &\phi(y^*) : \mathbb{R} \to \mathbb{R}, \\ &\varphi_i(x_i^*) : \mathbb{R} \to \mathbb{R}, \quad i = 1, \ldots, k, \end{aligned}$$

přepíšeme vztah jako

$$\phi(y^*) = \varphi(\boldsymbol{x}^*)^\mathsf{T}\boldsymbol{\beta}, \tag{2.1}$$

kde

$$\varphi(\boldsymbol{x}^*) = (\varphi_1(x_1^*), \ldots, \varphi_k(x_k^*))^\mathsf{T}.$$

V reálné situaci, kdy je rozumné předpokládat, že se zkoumaný proces dá aproximovat lineárními vztahy, však neznáme skutečné hodnoty proměnných y^*, $\boldsymbol{x}^*$. O procesu získáme představu pouze z námi naměřených dat. Ta se však od správných hodnot mohou velmi lišit. Máme tak k dispozici pouze vektory pozorování $\boldsymbol{y}$, $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$. Pro n pozorování tedy $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T}$, $\boldsymbol{x_i} = (x_{1,i}, \ldots, x_{n,i})^\mathsf{T}$.

Naším problémem je pak nalezení parametrů $\boldsymbol{\beta}$ jako řešení soustavy rovnic

$$\boldsymbol{y} \approx \boldsymbol{X}\boldsymbol{\beta}, \tag{2.2}$$

kde

$$\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_k}) = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,k} \end{pmatrix}.$$

Vlnovka ve výrazu $\boldsymbol{y} \approx \boldsymbol{X\beta}$ je použita proto, protože na soustavu hledíme jako na soubor naměřených dat, která považujeme z různých důvodů za nepřesná. O žádném pozorování nemůžeme říct a dokonce ani nepředpokládáme, že pro něj platí ideální lineární vztah v rovnici.

Naší představou je, že o systému generujícím data získáváme větší představu sbíráním dalších informací, tedy snažíme se o co největší počet pozorování n. Soustava (2.2) je pak přeurčená ($n > k$). A zpravidla tak nemá řešení. Dokonce čím více pozorování, tím menší pravděpodobnost, že by řešení soustavy existovalo. Tento rozpor nám jiným způsobem ilustruje, proč soustavu nechápeme jako rovnici.

Pokud nemůžeme najít přesné řešení soustavy, tak hledáme takové řešení, které by bylo nejlepší. Definice nejlepšího je pak do velké míry subjektivní. Avšak rozumná definice vždy závisí na správném pochopení struktury modelu, který řešíme. Dále si ještě upřesníme model, kterým se v práci budeme zabývat. A pokusíme se vysvětlit, v jakých ohledech je řešení metodou úplně nejmenších čtverců nejlepší.

Poznámka 2.1. Protože se hledání řešení soustavy (2.2) věnovalo i mnoho numerických matematiků, je v literatuře obvyklé i značení

$$\boldsymbol{Ax} \approx \boldsymbol{b}.$$

Tedy značení používané pro soustavy lineárních rovnic. Vzhledem k zamýšleným čtenářům této práce však dále v práci budeme používat značení, které je častěji používané statistiky při lineární regresi.

Druhy chyb

Výše jsme mluvili o proměnné y_i^*, jež se odlišuje od hodnoty pozorování y_i. Chybu pozorování si můžeme rozdělit na dvě části v_i a t_i.

Pod v_i můžeme rozumět souhrnnou chybu, která spočívá v nepřesnostech měření. Tedy čím lepší měřící přístroj budeme mít k dispozici, tím bude chyba v_i menší. V dokonalém případě ji pak můžeme celou eliminovat.

Na t_i se pak můžeme dívat jako na odchylku specifickou pro jedince i. Tedy pro daného jedince i je to hodnota. jakou se odlišuje od ideálního „průměrného“ jedince. Pro ideál platí přesný lineární vztah (2.2), ale jedinec/pozorování i má nějaké predispozice (např. genetické), kterými se od ideálu odlišuje. Chybu t_i tak nemůžeme nikdy eliminovat, protože je jedinci i přímo vlastní.

Je vůbec nutné chybu vždy rozdělovat na typy uvedené výše? Uvažujme o chybách jako o náhodných veličinách a předpokládejme, že jsou nezávislé a mají normální rozdělení, $v_i \sim \mathsf{N}(\mu_1, \sigma_1^2)$, $t_i \sim \mathsf{N}(\mu_2, \sigma_2^2)$. Pak i jejich součet má normální rozdělení

$$\varepsilon_i = v_i + t_i \sim \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Stejně jako jsme chápali v_i jako souhrn nepřesností našich měření a t_i jako souhrn odlišností jedince i od ideálu, tak ε_i můžeme jednoduše chápat jako souhrn obou typů chyb dohromady. Přičemž je v tomto případě zachována normalita. Za této situace by se tedy mohlo zdát, že rozdělovat chybu ε_i na dvě části není pro naší analýzu příliš podstatné.

Rozdělení je však podstatné v tom, jakým způsobem chápeme zkoumaný problém. Co se snažíme odhadovat. Máme veličiny provázány lineárními vztahy a protože nejsme schopni veličiny přesně změřit, je naším úkolem odhadnout konstanty těchto vztahů? Nebo hledáme parametry pro vztah, který platí pro průměrného jedince, jenž se v populaci nemusí vůbec vyskytovat?

Pokusme se zkonkretizovat rozdíl na jednoduchém modelu se dvěma proměnnými.

1. První model má tvar
$$Y^* = \alpha + \beta X^*,$$
kde pozorujeme hodnoty
$$Y = Y^* + v,$$
$$X = X^* + u.$$
Pro vztah mezi veličinami platí přímá úměrnost. Nejistota do vztahu vstupuje, protože reálné hodnoty nejsme schopni přesně získat. Naměříme je s nějakou chybou.

2. Pro druhý model si zadefinujeme další proměnnou Z^*, která bude zohledňovat chybu t, která je specifická danému jedinci. Tedy jedinec Z^* se od ideálu Y^* odlišuje o hodnotu t, která zohledňuje všechny jeho specifické vlastnosti/predispozice
$$Z^* = Y^* + t.$$
Další nepřesnosti do modelu vstupují při měření veličin X^* a Z^*, pro které pozorujeme hodnoty
$$Z = Z^* + v,$$
$$X = X^* + u.$$
Tedy jinými slovy, Z^* označujeme jedince, který se odlišuje od ideálu Y^*. Změřit ideál Y^* však není možné, a proto se naše pozorování Z vztahuje ke skutečnému jedinci Z^*.

   V modelu pak ztotožňujeme ideálního jedince s „průměrem“ skutečných jedinců
$$Y^* \approx \mathsf{E}(Z^* \,|\, X^*) = \alpha + \beta X^*.$$

První typ se v literatuře označuje jako *errors-in-variables* modely. Druhý typ připomíná model klasické lineární regrese. Tomu i odpovídá označení chyb v a t. v můžeme označovat jako chybu v proměnné (*error in variable*) a t jako chybu v rovnici (*error in equation*) – Malinvaud (1980)[kap. 10]. Nadále v práci budeme pracovat pouze s prvním typem modelu. Tedy odhadujeme model, kde jsou proměnné v lineárním vztahu a „jenom“ je nedokážeme přesně změřit.

Existují i další možnosti, jakým způsobem můžeme dát do vztahu dvě zkoumané veličiny. Je užitečným myšlenkovým cvičením si uvědomit rozdíly mezi jednotlivými modely – bližší diskuze viz Moran (1971).

Poznámka 2.2. V prvním modelu jsou všechny proměnné v rovnici záměnné. Tedy není podstatné, jakou proměnnou máme na levé straně rovnice – jedná se o jednoduché aritmetické operace v lineární rovnici. V druhém modelu však máme na levé straně podmíněnou střední hodnotu a aritmetické operace už nedávají smysl. Specifická odlišnost jedince od průměru se vztahuje přímo k veličině Z^*.

Functional a *structural relationship*

Další otázkou je, jak se díváme na skutečné hodnoty vysvětlujících proměnných X_i^*. Zda je považujeme za pevná čísla, nebo náhodné veličiny. Pro nedostatek českého ekvivalentu budeme modely označovat anglickými názvy, které se pro ně vžily. Modely, kde vysvětlující proměnné jsou konstanty se označují jako *linear functional relationship* a kde stochastické jako *linear structural relationship*.

Poznámka 2.3. Mnemotechnickou pomůckou pro zapamatování, který model označujeme jak, může být první písmeno názvu modelu. S jako stochastický a f jako fixní.

Pro bližší popis rozdílu se inspirujeme příkladem z Madansky (1959). Snažíme se odhadnout hustotu ρ železa na základě vztahu

$$hmotnost = \rho \cdot objem.$$

Ze všech vzorků železa, které máme k dispozici, můžeme vybrat zcela náhodně ty, které použijeme pro naši analýzu. V takovém případě je *objem* v rovnici náhodná veličina. Nebo si předem řekneme, o jakém objemu chceme vzorky železa vybírat. Pak do proměnné *objem* vstupují pevně zvolená čísla. V poplatnosti s modely, kterými se v této práci zabýváme, je vhodné zdůraznit, že se jedná o pevně zvolená čísla, která však nejsme schopni přesně změřit. Skutečná hodnota *objemu* je nenáhodná, ale pozorování již náhodné je. (Rozdělení pozorování pak závisí na tom, jakou a jak přesnou techniku měření zvolíme.)

Někdy ale může být obtížné rozhodnout, o jaký vztah (*structural* vs *functional*) se jedná. Respektive náš pohled musí být odpovídajícímu tomu, jaký problém máme v úmyslu řešit.

Nechť máme na naší zahradě 10 vzorků železa. Považujeme železo, které máme

k dispozici, za náhodný výběr z celé populace všech želez? Nebo považujeme naše vzorky za celkovou populaci – populaci želez na naší zahradě? V prvním případě je výběr náhodný, v druhém případě je výběr nenáhodný (vybrali jsme jednoho každého zástupce populace). V prvním případě děláme závěry o hustotě železa obecně, v druhém odhadujeme hustotu železa vyskytujícího se na naší zahradě.

Shrnutí úvodu

Zatím jsme si pouze zadefinovali, co chápeme pod pojmem lineární model. Pak jsme narazili na několik otázek, jejichž zodpovězení je důležité pro přesné pochopení toho, jaký model řešíme. První otázkou bylo, zda na levé straně rovnice máme přímo veličinu Y, nebo její podmíněnou střední hodnotu. Druhou pak, zda hodnoty vysvětlujících proměnných jsme získali náhodně, nebo je můžeme považovat za fixní.

V naší práci se budeme zabývat pouze modely, kde předpokládáme u zkoumaných veličin jen chyby vzniklé při jejich měření. U vysvětlujících proměnných pak budeme předpokládat, že jejich pozorování jsou pevné hodnoty.

Závěrem úvodu neuškodí poznamenat, že po správném pochopení modelu, musí i metoda použitá k jeho řešení reflektovat naše cíle. Snažíme se o pochopení struktury modelu? (Tj. odhadujeme parametry modelu.) Nebo se snažíme potvrdit či vyvrátit naše představy? (Tj. testujeme hypotézy o parametrech.) Anebo jsou naším cílem předpovědi do budoucnosti?

2.2 Základní úlohy OLS a TLS

V kapitole 1 jsme ukazovali na jednoduché úloze rozdíl mezi metodou nejmenších čtverců a úplně nejmenších čtverců. Rozdíl spočíval v tom, jak měříme vzdálenost pozorování a přímky, kterou jsme pozorování prokládali. Zpřesněme si nyní tento vágní pojem vzdálenosti z motivační úlohy. $\mathrm{M}(n,k)$ budeme označovat množinu všech matic o n řádcích a k sloupcích s reálnými prvky.

Definice 2.4 (Normy)**.**

(i) *Eukleidovskou normu* vektoru $\boldsymbol{x} = (x_1, \ldots, x_n)$ definujeme jako

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

(ii) Nechť $\boldsymbol{X} \in \mathrm{M}(n,k)$. Tedy

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,k} \end{pmatrix}.$$

Frobeniovou normou matice $\boldsymbol{X}$ rozumíme

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k} x_{i,j}^2}.$$

(iii) Nechť $\boldsymbol{X} \in \mathrm{M}(n,k)$. *2-normu* matice $\boldsymbol{X}$ definujeme pomocí Eukleidovské normy vektoru jako

$$\|\boldsymbol{X}\|_2 = \sup_{\boldsymbol{y}\in\mathbb{R}^k,\, \boldsymbol{y}\neq 0} \frac{\|\boldsymbol{X}\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2}.$$

Ekvivalentně

$$\|\boldsymbol{X}\|_2 = \max_{\boldsymbol{y}\in\mathbb{R}^k,\, \|\boldsymbol{y}\|_2=1} \|\boldsymbol{X}\boldsymbol{y}\|_2.$$

(Důkaz, že definice jsou ekvivalentní viz Golub a Van Loan (2013), kapitola 2.3.1.)

Poznámka 2.5. Výše jsme si zadefinovali jednu vektorovou a dvě maticové normy. Eukleidovská norma nám bude sloužit pro měření vzdálenosti mezi dvěma vektory. Frobeniovu a 2-normu pak využijeme pro určování, jak jsou od sebe vzdálené dvě matice. Podrobnější diskuzi ohledně možných norem pro řešení úloh nejmenších čtverců lze naleznout ve Watson (1998).

Poznámka 2.6. V rámci lepší čitelnosti většinou nebudeme v práci označovat rozměry vektorů a matic. Počet řádků a sloupců budeme uvádět pouze v těch případech, kdy nám jejich označení pomůže k lepšímu pochopení problému. Případně, pokud by vynechání rozměrů mohlo čtenáře zmást. Matici o n řádcích a k sloupcích tak budeme označovat jako $\boldsymbol{X}$ nebo $\boldsymbol{X}_{n\times k}$.

Věta 2.7 (Jednoduché vlastnosti norem)**.**

(i) Nechť $\boldsymbol{X} \in \mathrm{M}(n,k)$

$$\|\boldsymbol{X}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})} = \sqrt{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})}.$$

Kde $\mathrm{tr}(\cdot)$ označujeme stopu čtvercové matice, tedy pro $\boldsymbol{Y} \in \mathrm{M}(m,m)$

$$\mathrm{tr}(\boldsymbol{Y}) = \sum_{i=1}^{m} y_{i,i}.$$

(ii) Nechť $\boldsymbol{E} \in \mathrm{M}(n,k)$, $\boldsymbol{F} \in \mathrm{M}(k,n)$. Pak

$$\mathrm{tr}(\boldsymbol{E}\boldsymbol{F}) = \mathrm{tr}(\boldsymbol{F}\boldsymbol{E}).$$

(iii) Nechť $\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{C} \in \mathrm{M}(n,n)$. Pak

$$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}).$$

(iv) Nechť $\boldsymbol{A} \in \mathrm{M}(n,k)$. Nechť $\boldsymbol{R}_1 \in \mathrm{M}(n,n)$ a $\boldsymbol{R}_2 \in \mathrm{M}(k,k)$ jsou ortonormální matice. Potom

$$\|\boldsymbol{A}\|_F = \|\boldsymbol{R}_1\boldsymbol{A}\|_F = \|\boldsymbol{R}_1^\mathsf{T}\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{R}_2\|_F = \|\boldsymbol{A}\boldsymbol{R}_2^\mathsf{T}\|_F,$$

Pozn. ortonormální maticí rozumíme takovou matici, pro kterou platí, že transponovaná matice se rovná inverzní. Tedy

$$\boldsymbol{R}\boldsymbol{R}^\mathsf{T} = \boldsymbol{R}^\mathsf{T}\boldsymbol{R} = \boldsymbol{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

(v) Nechť $\boldsymbol{R} \in \mathrm{M}(n,n)$ je ortonormální matice a $\boldsymbol{x} \in \mathbb{R}^n$. Potom

$$\|\boldsymbol{R}\boldsymbol{x}\|_2 = \|\boldsymbol{R}^\mathsf{T}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2.$$

(vi) Nechť $\boldsymbol{A} \in \mathrm{M}(n,k)$. Nechť $\boldsymbol{R}_1 \in \mathrm{M}(n,n)$ a $\boldsymbol{R}_2 \in \mathrm{M}(k,k)$ jsou ortonormální matice. Potom

$$\|\boldsymbol{A}\|_2 = \|\boldsymbol{R}_1\boldsymbol{A}\|_2 = \|\boldsymbol{R}_1^\mathsf{T}\boldsymbol{A}\|_2 = \|\boldsymbol{A}\boldsymbol{R}_2\|_2 = \|\boldsymbol{A}\boldsymbol{R}_2^\mathsf{T}\|_2,$$

Důkaz.

(i) Plyne přímo z definice, vynásobíme-li mezi sebou matice $\boldsymbol{X}$ a $\boldsymbol{X}^\mathsf{T}$, respektive $\boldsymbol{X}^\mathsf{T}$ a $\boldsymbol{X}$.

(ii) Označme $\boldsymbol{C} = \boldsymbol{E}\boldsymbol{F}$ a $\boldsymbol{D} = \boldsymbol{F}\boldsymbol{E}$. Prvky matic tak jsou $c_{i,j} = \sum_{l=1}^{k} e_{i,l}\, f_{l,j}$, $i,j = 1,\dots,n$ a $d_{i,j} = \sum_{l=1}^{n} f_{i,l}\, e_{l,j}, i,j = 1,\dots,k$. Potom

$$\operatorname{tr}\boldsymbol{C} = \sum_{i=1}^{n} c_{i,i} = \sum_{i=1}^{n}\sum_{l=1}^{k} e_{i,l}\, f_{l,i} = \sum_{l=1}^{k}\sum_{i=1}^{n} f_{l,i}\, e_{i,l} = \sum_{l=1}^{k} d_{l,l} = \operatorname{tr}\boldsymbol{D}.$$

(iii) Plyne z předchozího bodu (ii), označíme-li $\boldsymbol{E} = \boldsymbol{A}\boldsymbol{B}$ a $\boldsymbol{F} = \boldsymbol{C}$.

(iv) Rovnosti lze ověřit jednoduchou aplikací bodů (i) a (iii). Dokažme si tvrzení např. pro $\|\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{R}_2\|_F$

$$\|\boldsymbol{A}\boldsymbol{R}_2\|_F^2 \overset{\text{(i)}}{=} \operatorname{tr}\left(\boldsymbol{R}_2^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{R}_2\right) \overset{\text{(iii)}}{=} \operatorname{tr}\left(\boldsymbol{R}_2\boldsymbol{R}_2^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right) = \operatorname{tr}\left(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right) = \|\boldsymbol{A}\|_F^2.$$

(v) Dokažme nejdříve, že $\|\boldsymbol{R}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$.

$$\|\boldsymbol{R}\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\mathsf{T}\boldsymbol{R}^\mathsf{T}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2.$$

Druhou rovnost lze dokázat analogicky.

(vi) Pro ortonormální matici na levé straně

$$\begin{aligned}\|\boldsymbol{R}_1\boldsymbol{A}\|_2^2 &= \max_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{R}_1\boldsymbol{A}\boldsymbol{y}\|_2^2 \\ &= \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{R}_1^\mathsf{T}\boldsymbol{R}_1\boldsymbol{A}\boldsymbol{y} \\ &= \max_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{A}\boldsymbol{y}\|_2^2 \\ &= \|\boldsymbol{A}\|_2^2.\end{aligned}$$

Pro ortonormální matici na pravé straně využijeme bod (v).

$$\|\boldsymbol{A}\boldsymbol{R}_2\|_2^2 = \max_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{A}\boldsymbol{R}_2\boldsymbol{y}\|_2^2 \overset{\text{(v)}}{=} \max_{\|\boldsymbol{R}_2\boldsymbol{y}\|_2=1} \|\boldsymbol{A}\boldsymbol{R}_2\boldsymbol{y}\|_2^2 = \max_{\|\boldsymbol{z}\|_2=1} \|\boldsymbol{A}\boldsymbol{z}\|_2^2 = \|\boldsymbol{A}\|_2^2.$$

□

Nyní si zadefinujme základní úlohy, kterými se v naší práci budeme zabývat. Výklad sleduje Van Huffel a Vandewalle (1991).

Definice 2.8 (Základní úloha nejmenších čtverců)**.** Nechť máme přeurčenou soustavu n rovnic (2.2)

$$\boldsymbol{y}_{n\times 1} \approx \boldsymbol{X}_{n\times k}\,\boldsymbol{\beta}_{k\times 1}.$$

Úlohou nejmenších čtverců je najít takové $\hat{\boldsymbol{y}}$, které:

1. minimalizuje Eukleidovskou normu

$$\min_{\hat{\boldsymbol{y}}\in\mathbb{R}^n} \|\boldsymbol{y}-\hat{\boldsymbol{y}}\|_2,$$

2. leží v prostoru generovaném sloupci matice $\boldsymbol{X}$

$$\hat{\boldsymbol{y}} \in R(\boldsymbol{X}).$$

Řešením metodou nejmenších čtverců je pak libovolné $\hat{\boldsymbol{\beta}}$, které splňuje

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{2.3}$$

Máme tedy k dispozici vektor pozorování $\boldsymbol{y}$ a matici pozorování $\boldsymbol{X}$. Snažíme se odhadnout parametry modelu $\boldsymbol{\beta}$ tak aby odhadnuté hodnoty $\hat{\boldsymbol{y}}$ byly co nejblíže pozorování $\boldsymbol{y}$.

Definice 2.9 (Základní úloha úplně nejmenších čtverců)**.** Nechť máme opět přeurčenou soustavu n rovnic (2.2)

$$\boldsymbol{y}_{n\times 1} \approx \boldsymbol{X}_{n\times k}\,\boldsymbol{\beta}_{k\times 1}.$$

Úlohou úplně nejmenších čtverců je najít taková $\tilde{\boldsymbol{y}}$ a $\tilde{\boldsymbol{X}}$, která

1. minimalizují Frobeniovu normu

$$\min_{(\tilde{\boldsymbol{X}},\,\tilde{\boldsymbol{y}})\in\mathbb{R}^{n\times(k+1)}} \|(\boldsymbol{X},\,\boldsymbol{y}) - (\tilde{\boldsymbol{X}},\,\tilde{\boldsymbol{y}})\|_F,$$

2. $\tilde{\boldsymbol{y}}$ leží v prostoru generovaném sloupci matice $\tilde{\boldsymbol{X}}$

$$\tilde{\boldsymbol{y}} \in R(\tilde{\boldsymbol{X}}).$$

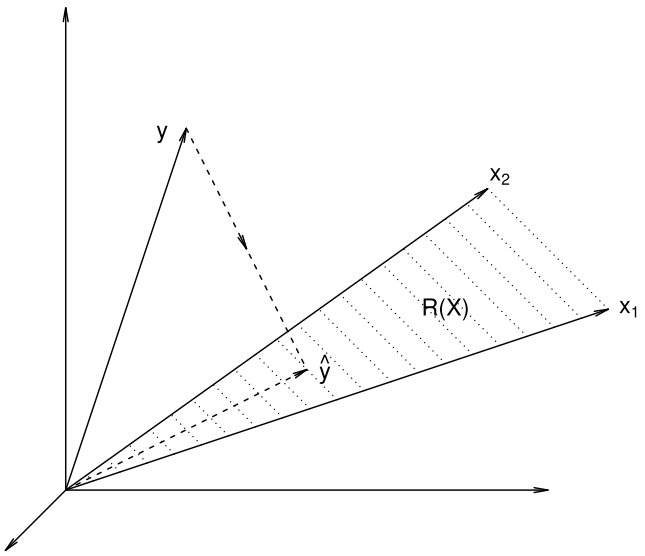
Řešením metodou úplně nejmenších čtverců pak je libovolné $\tilde{\boldsymbol{\beta}}$, které splňuje

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}. \tag{2.4}$$

Zaměřme se nyní na odlišnosti v definicích základních úloh výše. V analogii ze vzorovou úlohou (1.1), je rozdíl v tom, jakým způsobem měříme vzdálenost pozorování od předpokládané skutečné hodnoty.

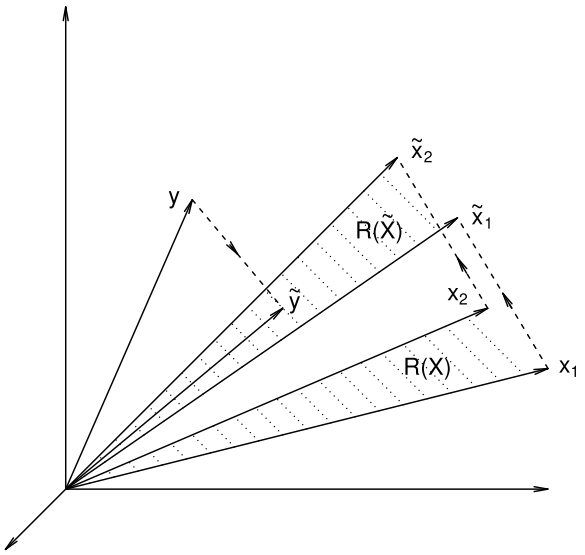
V úloze *nejmenších čtverců* měříme vzdálenost v Eukleidovské normě, v úloze *úplně nejmenších čtverců* ve Frobeniově normě. V prvním případě opět předpokládáme chybu pouze ve veličině y, zatímco v druhém i ve veličinách x_i. Což se odráží v tom, jakým způsobem upravujeme pozorované hodnoty.

Pro OLS manipulujeme hodnoty $\boldsymbol{y}$ tak, aby upravené hodnoty $\hat{\boldsymbol{y}}$ ležely v prostoru generovaném sloupci matice $\boldsymbol{X}$. Jinými slovy promítáme $\boldsymbol{y}$ do prostoru $R(\boldsymbol{X})$. Schématicky je to zobrazeno na obr. 2.1. Máme 2 trojrozměrné sloupce matice $\boldsymbol{X}$, které nám vytvářejí plochu $R(\boldsymbol{X})$. Do této plochy pak promítáme vektor $\boldsymbol{y}$.



Obrázek 2.1: Řešení metodou nejmenších čtverců.

Pro TLS manipulujeme jak pozorování $\boldsymbol{y}$ tak $\boldsymbol{X}$. Přičemž chceme, aby upravené hodnoty $\tilde{\boldsymbol{y}}$ ležely v prostoru generovaném sloupci upravené matice $\tilde{\boldsymbol{X}}$. Promítáme $\boldsymbol{y}$ do prostoru $R(\tilde{\boldsymbol{X}})$. Viz obr. 2.2. Sloupce $\tilde{x}_1$ a $\tilde{x}_2$ upravené matice $\tilde{\boldsymbol{X}}$ nám vytvářejí plochu $R(\tilde{\boldsymbol{X}})$, do které promítáme $\boldsymbol{y}$.



Obrázek 2.2: Řešení metodou úplně nejmenších čtverců.

2.3 Hledání řešení

Hledat řešení základních úloh budeme pomocí singulárního rozkladu.

Věta 2.10 (Singulární rozklad)**.** Nechť $\boldsymbol{C} \in \mathrm{M}(n,k)$. Pak existují ortonormální matice $\boldsymbol{U} \in \mathrm{M}(n,n)$ a $\boldsymbol{V} \in \mathrm{M}(k,k)$ takové, že:

$$\boldsymbol{U}^{\mathsf{T}}\boldsymbol{C}\,\boldsymbol{V} = \boldsymbol{\Sigma},$$

kde $\boldsymbol{\Sigma}$ je diagonální matice s kladnými hodnotami na diagonále. Prvky na diagonále matice $\boldsymbol{\Sigma}$ označujeme jako *singulární hodnoty matice* $\boldsymbol{C}$. Pro jednoznačnost matice $\boldsymbol{\Sigma}$ předpokládáme, že posloupnost singulárních hodnot je nerostoucí.

Tedy pro $n \leq k$:

$$\boldsymbol{\Sigma}_{n\times k} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & \underbrace{0 & \dots & 0}_{(k-n)} \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n & 0 & \dots & 0 \end{pmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0.$$

A pro $n > k$:

$$\boldsymbol{\Sigma}_{n\times k} = \begin{matrix} \\ \\ \\ \\ (n-k) \end{matrix} \left\{ \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_k \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \right., \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq 0.$$

Nadále budeme označovat diagonální matici $\boldsymbol{\Sigma}_{n\times k}$

$$\boldsymbol{\Sigma}_{n\times k} = \operatorname{diag}(\sigma_1, \ldots, \sigma_m), \quad m = \min(n, k).$$

Důkaz. Viz Golub a Van Loan (2013), Věta 2.4.1. □

Pokud spočítáme singulární rozklad matice $\boldsymbol{C}$, pak její Frobeniovu i 2-normu získáme snadno pomocí singulárních hodnot.

Důsledek 2.11. Nechť $\boldsymbol{C} \in \mathrm{M}(n, k)$ a $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$ je její singulární rozklad. Pak

(i)
$$\|\boldsymbol{C}\|_2 = \sigma_1,$$

(ii)
$$\|\boldsymbol{C}\|_F = \sqrt{\sigma_1^2 + \ldots + \sigma_p^2}, \quad p = \min\{n, k\}.$$

Důkaz.

(i) Protože $\boldsymbol{U}$ a $\boldsymbol{V}$ jsou ortonormální matice, tak podle věty 2.7 víme, že

$$\|\boldsymbol{C}\|_2 = \|\boldsymbol{U}^\mathsf{T}\boldsymbol{C}\boldsymbol{V}\|_2 = \|\boldsymbol{\Sigma}\|_2.$$

Protože σ_1 je největší singulární hodnota, pak přímo z definice 2-normy platí

$$\|\boldsymbol{\Sigma}\|_2 = \max_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{\Sigma}\boldsymbol{y}\|_2 = \sigma_1.$$

(ii) Postup důkazu je podobný jako v bodě (i). Věta 2.7 nám opět říká, že $\|\boldsymbol{C}\|_F = \|\boldsymbol{\Sigma}\|_F$. Frobeniovu normu pak spočteme jako stopu diagonální matice

$$\|\boldsymbol{\Sigma}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma})} = \sqrt{\sigma_1^2 + \ldots + \sigma_p^2}, \quad p = \min\{n, k\}.$$

□

Uveďme si ještě další dva zřejmé důsledky věty o singulárním rozkladu.

Důsledek 2.12. Hodnost matice $\boldsymbol{C}$, $\mathrm{h}(\boldsymbol{C})$, je rovna počtu nenulových singulárních hodnot.

Důkaz. Protože matice $\boldsymbol{U}$ a $\boldsymbol{V}$ jsou ortonormální (tedy jsou plné hodnosti), tak

$$\mathrm{h}(\boldsymbol{C}) = \mathrm{h}(\boldsymbol{U}^\mathsf{T}\boldsymbol{C}\boldsymbol{V}) = \mathrm{h}(\boldsymbol{\Sigma}).$$

Hodnost diagonální matice je rovna počtu nenulových prvků na diagonále. □

Důsledek 2.13 (Dyadická dekompozice.)**.** Nechť $\boldsymbol{C} \in \mathrm{M}(n,k)$, $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$ je její singulární rozklad a $\mathrm{h}(\boldsymbol{C}) = r$. Pak

$$\boldsymbol{C} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u_i} \boldsymbol{v_i}^\mathsf{T}.$$

Důkaz. Plyne přímo ze singulárního rozkladu

$$\boldsymbol{C} = (\boldsymbol{U}\boldsymbol{\Sigma})\boldsymbol{V}^\mathsf{T} = (\sigma_1\boldsymbol{u_1},\sigma_2\boldsymbol{u_2},\ldots,\sigma_r\boldsymbol{u_r},\boldsymbol{0},\ldots,\boldsymbol{0}) \begin{pmatrix} \boldsymbol{v_1}^\mathsf{T} \\ \vdots \\ \boldsymbol{v_k}^\mathsf{T} \end{pmatrix} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u_i} \boldsymbol{v_i}^\mathsf{T}.$$

□

Dyadická dekompozice nám umožňuje rozpadnout matici o hodnosti r na součet r matic o hodnosti 1.

Následující věta ukazuje, jak nejlépe aproximovat matici jinou maticí s nižší hodností. Nejlépe ve smyslu, že si jsou matice co nejblíže. Přičemž vzdálenost měříme v normách definovaných výše. Singulární rozklad nám pak umožňuje velmi jednoduše charakterizovat matici, která nám slouží pro aproximaci.

Věta 2.14 (Eckart-Young-Mirsky)**.** Nechť $\boldsymbol{C} \in \mathrm{M}(n,k)$, $\mathrm{h}(\boldsymbol{C}) = r$ a $\sum_{i=1}^{r} \sigma_i \boldsymbol{u_i} \boldsymbol{v_i}^\mathsf{T}$ je singulární rozklad $\boldsymbol{C}$. Pokud $m < r$ a $\boldsymbol{C}_m = \sum_{i=1}^{m} \sigma_i \boldsymbol{u_i} \boldsymbol{v_i}^\mathsf{T}$, pak

(i)
$$\min_{\mathrm{h}(\boldsymbol{D})=m} \|\boldsymbol{C} - \boldsymbol{D}\|_2 = \|\boldsymbol{C} - \boldsymbol{C}_m\|_2 = \sigma_{m+1},$$

(ii)
$$\min_{\mathrm{h}(\boldsymbol{D})=m} \|\boldsymbol{C} - \boldsymbol{D}\|_F = \|\boldsymbol{C} - \boldsymbol{C}_m\|_F = \sqrt{\sum_{i\,=\,m+1}^{p} \sigma_i}, \quad p = \min\{n,k\}.$$

Důkaz.

(i) Viz Golub a Van Loan (2013), Věta 2.4.8.

(ii) Viz Eckart a Young (1936).

□

Nyní máme vše připraveno k tomu, abychom pomocí singulárního rozkladu nalezli řešení základních úloh 2.8 a 2.9. Už ve vzorové úloze v kapitole 1 jsme se snažili pozorování prokládat přímkou tak, aby byla přímka těmto pozorováním co nejblíže. Věta 2.14 nám umožňuje v obecné situaci nalézt řešení, které je pozorováním nejblíže ve smyslu norem, které jsme si zadefinovali v 2.4.

Věta 2.15 (Řešení OLS úlohy singulárním rozkladem)**.** Řešme základní úlohu nejmenších čtverců dle definice 2.8. Nechť singulární rozklad matice pozorování $\boldsymbol{X}$ je $\boldsymbol{X}_{n\times k} = \boldsymbol{U}_{n\times n}\,\boldsymbol{\Sigma}_{n\times k}\,\boldsymbol{V}^{\mathsf{T}}_{k\times k}$. Buď $\mathrm{h}(\boldsymbol{X}) = r$. Pak řešením úlohy nejmenších čtverců je

$$\hat{\boldsymbol{\beta}}_{k\times 1} = \boldsymbol{V}_{k\times k}\,\boldsymbol{\Sigma}^{-1}_{k\times n}\,\boldsymbol{U}^{\mathsf{T}}_{n\times n}\,\boldsymbol{y}_{n\times 1}, \tag{2.5}$$

kde

$$\boldsymbol{\Sigma}^{-1}_{k\times n} = \mathrm{diag}\left(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_r},0,\ldots,0\right) = \begin{matrix}\\ (k-r)\end{matrix}\overset{\overbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad}^{n-r}}{\begin{pmatrix} \frac{1}{\sigma_1} & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \frac{1}{\sigma_2} & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \frac{1}{\sigma_r} & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix}}.$$

Ekvivalentně

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{r} \frac{\boldsymbol{u_i}^{\mathsf{T}}\boldsymbol{y}}{\sigma_i}\boldsymbol{v_i}. \tag{2.6}$$

Navíc $\hat{\boldsymbol{\beta}}$ má nejmenší Eukleidovskou normu mezi všemi řešeními úlohy 2.8 a reziduální součet čtverců je roven

$$RSS = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 = \sum_{i=r+1}^{n} (\boldsymbol{u_i}^{\mathsf{T}}\boldsymbol{y})^2.$$

Důkaz. Nechť $\delta \in \mathbb{R}^k$ je libovolné. Potom

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\delta}\|_2^2 \overset{(*)}{=} \|\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} - (\boldsymbol{U}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{V})(\boldsymbol{V}^{\mathsf{T}}\boldsymbol{\delta})\|_2^2 = \|\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{\Sigma}\boldsymbol{\alpha}\|_2^2 =$$
$$= \sum_{i=1}^{r}(\boldsymbol{u_i}^{\mathsf{T}}\boldsymbol{y} - \sigma_i\alpha_i)^2 + \sum_{i=r+1}^{n}(\boldsymbol{u_i}^{\mathsf{T}}\boldsymbol{y})^2,$$

kde $\boldsymbol{\alpha} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{\delta}$. U rovnosti $(*)$ využíváme větu 2.7 (v) a toho, že matice $\boldsymbol{V}$ je ortonormální.

Abychom získali minimum $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\delta}\|_2^2$ musí platit, že $\alpha_i = \boldsymbol{u_i}^{\mathsf{T}}\boldsymbol{y}/\sigma_i$ pro $\forall\, i \in \{1,\ldots,r\}$. Pro minimalizaci Eukleidovské normy $\boldsymbol{\alpha}$ pak položíme $\alpha_i = 0$ pro $\forall\, i \in \{r+1,\ldots,k\}$. Tedy závěrem (2.6) je řešením úlohy nejmenších čtverců s nejmenší Eukleidovskou normou (opět podle věty 2.7 (v) platí, že $\|\boldsymbol{\alpha}\|_2 = \|\boldsymbol{\delta}\|_2$). Charakterizace reziduálního součtu čtverců je zřejmá.

Ekvivalence řešení (2.5) a řešení (2.6) vyjádřeného dyadickou dekompozicí pak plyne přímo z maticového násobení. □

Poznámka 2.16 (Souvislost SVD řešení s Moore-Penroseho pseudoinverzní maticí)**.** Pomocí SVD rozkladu lze také snadno spočítat Moore-Penroseho pseudoinverzi matice $\boldsymbol{X}$. Snadno se ověří, že matice $\boldsymbol{X}^+ = \boldsymbol{V}\,\boldsymbol{\Sigma}\,\boldsymbol{U}^\mathsf{T}$ splňuje všechny požadavky, které na Moore-Penroseho matici klademe. Řešení (2.5) lze tak přepsat jako

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^+\boldsymbol{y}.$$

Podrobnější informace o vlastnostech pseudoinverzních matic a přesnou definici Moore-Penrose matice viz Anděl (2011), dodatek A.

Uveďme si ještě jednu věta, která nám pomůže při hledání řešení TLS problému pomocí singulárního rozkladu.

Věta 2.17. Nechť $\boldsymbol{C} \in \mathrm{M}(n,k)$. A nechť jejími singulárními hodnotami jsou $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_{\min\{n,k\}}$. Vytvořme matici $\boldsymbol{D}$ vynecháním jednoho sloupce z matice $\boldsymbol{C}$. Označme singulární hodnoty matice $\boldsymbol{D}$ jako $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_{\min\{n,k-1\}}$. Pak

(i) Když $n \geq k$: $\quad \gamma_1 \geq \delta_1 \geq \gamma_2 \geq \delta_2 \geq \ldots \geq \delta_{k-1} \geq \gamma_k$,

(ii) Když $n < k$: $\quad \gamma_1 \geq \delta_1 \geq \gamma_2 \geq \delta_2 \geq \ldots \geq \delta_n \geq \gamma_n$.

Důkaz. Viz Thompson (1972). □

Věta 2.18 (Řešení TLS úlohy singulárním rozkladem)**.** Řešme základní úlohu úplně nejmenších čtverců dle definice 2.9. Mějme singulární rozklad matic pozorování $\boldsymbol{X} = \boldsymbol{U'}\,\boldsymbol{\Sigma'}\,\boldsymbol{V'}^\mathsf{T}$ a $(\boldsymbol{X},\boldsymbol{y}) = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^\mathsf{T}$. Jejich singulární hodnoty označíme jako $\sigma'_1 \geq \ldots \geq \sigma'_k$, resp. $\sigma_1 \geq \ldots \geq \sigma_n \geq \sigma_{k+1}$. Nechť $\sigma'_k > \sigma_{k+1}$. Pak řešením úplně nejmenších čtverců je

$$\tilde{\boldsymbol{\beta}} = -\frac{1}{v_{k+1,k+1}} \begin{pmatrix} v_{1,k+1} \\ v_{2,k+1} \\ \vdots \\ v_{k,k+1} \end{pmatrix}. \tag{2.7}$$

Řešení existuje a je jediným řešením rovnic $\tilde{\boldsymbol{X}}\boldsymbol{\beta} = \tilde{\boldsymbol{y}}$, kde

$$(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{y}}) = \boldsymbol{U}\,\tilde{\boldsymbol{\Sigma}}\,\boldsymbol{V}^\mathsf{T}, \quad \tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\sigma_1,\ldots,\sigma_k,0).$$

Důkaz. Proč se díváme na matici $(\boldsymbol{X},\boldsymbol{y})$ plyne z toho, že řešenou soustavu (2.2) si můžeme přepsat jako

$$(\boldsymbol{X},\boldsymbol{y})(\boldsymbol{\beta}^\mathsf{T},-1)^\mathsf{T} \approx \boldsymbol{0}.$$

Matice $\boldsymbol{X}$ vznikla z matice $(\boldsymbol{X},\boldsymbol{y})$ vynecháním jednoho sloupce, tak dle věty 2.17 a použitím předpokladu $\sigma'_k > \sigma_{k+1}$ máme

$$\sigma_1 \geq \sigma'_1 \geq \ldots \geq \sigma_k \geq \sigma'_k > \sigma_{k+1} \geq 0.$$

Tyto nerovnosti nám napovídají o hodnosti zkoumaných matic. Důsledek 2.12 nám říká, že když $\sigma'_k > 0$, tak je matice $\boldsymbol{X}$ plné hodnosti k (z definované struktury zkoumaného problému předpokládáme, že má matice $\boldsymbol{X}$ více řádků než sloupců). Stejně tak víme, že hodnost matice $(\boldsymbol{X}, \boldsymbol{y})$ je alespoň k.

Za předpokladu, že hodnost matice $(\boldsymbol{X}, \boldsymbol{y})$ je $k+1$, tak nám věta 2.14 umožní najít její nejbližší aproximaci o hodnosti k ve Frobeniově normě

$$\min_{\mathrm{h}((\boldsymbol{X}',\boldsymbol{y}'))=k} \|(\boldsymbol{X}, \boldsymbol{y}) - (\boldsymbol{X}', \boldsymbol{y}')\|_F = \|(\boldsymbol{X}, \boldsymbol{y}) - (\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})\|_F,$$

kde

$$(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}) = \boldsymbol{U}\,\tilde{\boldsymbol{\Sigma}}\,\boldsymbol{V}^\mathsf{T}, \quad \tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0).$$

Protože hodnost $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$ je k, tak existuje jediné řešení $\hat{\boldsymbol{\beta}}$ takové, že $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{y}}$.

Zbývá nám ověřit, že $\tilde{\boldsymbol{\beta}}$ odpovídá opravdu řešení (2.7).

Protože

$$(\boldsymbol{U}\,\tilde{\boldsymbol{\Sigma}})\,\boldsymbol{V}^\mathsf{T} = \begin{pmatrix} \sigma_1 u_{1,1} & \sigma_2 u_{1,2} & \ldots & \sigma_k u_{1,k} & 0 \\ \sigma_1 u_{2,1} & \sigma_2 u_{2,2} & \ldots & \sigma_k u_{2,k} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_n u_{n,1} & \sigma_2 u_{n,2} & \ldots & \sigma_k u_{n,k} & 0 \end{pmatrix} \begin{pmatrix} v_{1,1} & v_{2,1} & \ldots & v_{k+1,1} \\ v_{1,2} & v_{2,2} & \ldots & v_{k+1,2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1,k+1} & v_{2,k+1} & \ldots & v_{k+1,k+1} \end{pmatrix},$$

tak

$$\tilde{\boldsymbol{y}} = \begin{pmatrix} \sigma_1 u_{1,1} v_{k+1,1} + \sigma_2 u_{1,2} v_{k+1,2} + \ldots + \sigma_k u_{1,k} v_{k+1,k} + 0 \\ \sigma_1 u_{2,1} v_{k+1,1} + \sigma_2 u_{2,2} v_{k+1,2} + \ldots + \sigma_k u_{2,k} v_{k+1,k} + 0 \\ \vdots \\ \sigma_n u_{n,1} v_{k+1,1} + \sigma_2 u_{n,2} v_{k+1,2} + \ldots + \sigma_k u_{n,k} v_{k+1,k} + 0 \end{pmatrix}$$

a řádek m matice $\tilde{\boldsymbol{X}}$ je

$$\tilde{\boldsymbol{x}}_{m,\cdot}^\mathsf{T} = \begin{pmatrix} \sigma_1 u_{m,1} v_{1,1} + \sigma_2 u_{m,2} v_{1,2} + \ldots + \sigma_k u_{m,k} v_{1,k} + 0 \\ \sigma_1 u_{m,1} v_{2,1} + \sigma_2 u_{m,2} v_{2,2} + \ldots + \sigma_k u_{m,k} v_{2,k} + 0 \\ \vdots \\ \sigma_1 u_{m,1} v_{k,1} + \sigma_2 u_{m,2} v_{k,2} + \ldots + \sigma_k u_{m,k} v_{k,k} + 0 \end{pmatrix}, \quad m = 1, \ldots, n.$$

Tedy

$$\tilde{\boldsymbol{x}}_{m,\cdot}\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \sigma_1 u_{m,1} v_{1,1} + \sigma_2 u_{m,2} v_{1,2} + \ldots + \sigma_k u_{m,k} v_{1,k} \\ \sigma_1 u_{m,1} v_{2,1} + \sigma_2 u_{m,2} v_{2,2} + \ldots + \sigma_k u_{m,k} v_{2,k} \\ \vdots \\ \sigma_1 u_{m,1} v_{k,1} + \sigma_2 u_{m,2} v_{k,2} + \ldots + \sigma_k u_{m,k} v_{k,k} \end{pmatrix}^\mathsf{T} \begin{pmatrix} v_{1,k+1} \\ v_{2,k+1} \\ \vdots \\ v_{k,k+1} \end{pmatrix} \cdot \frac{-1}{v_{k+1,k+1}}$$

$$= \frac{-1}{v_{k+1,k+1}} \left(\sigma_1 u_{m,1} \sum_{j=1}^{k} v_{j,1} v_{j,k+1} + \sigma_2 u_{m,2} \sum_{j=1}^{k} v_{j,2} v_{j,k+1} + \ldots \right.$$

$$\left. \ldots + \sigma_k u_{m,k} \sum_{j=1}^{k} v_{j,k} v_{j,k+1} \right)$$

$$= \frac{-1}{v_{k+1,k+1}} \left(\sigma_1 \, u_{m,1} \sum_{j=1}^{k+1} v_{j,1} \, v_{j,k+1} + \sigma_2 \, u_{m,2} \sum_{j=1}^{k+1} v_{j,2} \, v_{j,k+1} + \dots \right.$$
$$\left. \dots + \sigma_k \, u_{m,k} \sum_{j=1}^{k+1} v_{j,k} \, v_{j,k+1} - \sum_{i=1}^{k+1} \sigma_i \, u_{m,i} \, v_{k+1,i} \, v_{k+1,k+1} \right)$$
$$= 0 + \dots + 0 + \sum_{i=1}^{k+1} \sigma_i \, u_{m,i} \, v_{k+1,i}.$$

Což odpovídá m-tému členu vektoru $\tilde{\boldsymbol{y}}$.

Pokud je hodnost matice $(\boldsymbol{X}, \boldsymbol{y})$ rovna hodnosti matice $\boldsymbol{X}$, pak nalezneme $\boldsymbol{\beta}$, pro které přímo $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$. Tedy v takovém případě nemáme co aproximovat, protože máme přesné řešení. Protože hodnost $(\boldsymbol{X}, \boldsymbol{y})$ je k, tak $\sigma_{k+1} = 0$ a platí

$$\boldsymbol{U} \, \boldsymbol{\Sigma} \, \boldsymbol{V}^\mathsf{T} = (\boldsymbol{X}, \boldsymbol{y}) = (\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}) = \boldsymbol{U} \, \tilde{\boldsymbol{\Sigma}} \, \boldsymbol{V}^\mathsf{T}.$$

Že řešení (2.7) splňuje $\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{y}$ se pak ověří analogicky jako výše. □

V předchozí větě jsme neřešili situaci, že je $v_{k+1,\,k+1}$ rovno nule. Ani jsme nemuseli, protože platí následující ekvivalence.

Věta 2.19. Nechť singulární rozklady matic pozorování jsou opět $\boldsymbol{X} = \boldsymbol{U'} \, \boldsymbol{\Sigma'} \, \boldsymbol{V'}^\mathsf{T}$ a $(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{U} \, \boldsymbol{\Sigma} \, \boldsymbol{V}^\mathsf{T}$. Pak

$$\sigma'_k > \sigma_{k+1} \iff \sigma_k > \sigma_{k+1} \text{ a } v_{k+1,\,k+1} \neq 0.$$

Důkaz. Viz Van Huffel a Vandewalle (1991), Dodatek 3.4. □

Poznámka 2.20. Jak moc velkým omezením je předpoklad $\sigma'_k > \sigma_{k+1}$? Van Huffel a Vandewalle (1991) ukazuje, že v reálných datech není tato podmínka splněna jen zřídka. V třetí kapitole knihy však lze nalézt diskuzi o případech, kdy předpoklad splněn není.

Řešení úplně nejmenších čtverců si lze vyjádřit ve formě, která připomíná řešení klasických nejmenších čtverců.

Věta 2.21. Nechť jsou splněny předpoklady věty 2.18. Pak

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} - \sigma_{k+1}^2 \boldsymbol{I})^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{y}$$

Důkaz. Viz Van Huffel a Vandewalle (1991), Věta 2.7. □

2.3.1 TLS úloha obsahující intercept

Zatím jsme se zaměřovali na úlohy úplně nejmenších čtverců, pro které jsme předpokládali, že všechny pozorované hodnoty v matici $\boldsymbol{X}$ jsme získali nepřesně. Může se ale stát, že řešíme jakousi směsici mezi modelem nejmenších čtverců a modelem úplně nejmenších čtverců. Některé sloupce matice jsme schopni změřit přesně, zatímco jiné ne.

Definice 2.22 (Základní úloha smíšeného modelu nejmenších a úplně nejmenších čtverců)**.** Nechť máme přeurčenou soustavu n rovnic (2.2)

$$\boldsymbol{y}_{n\times 1} \approx \boldsymbol{X}_{n\times k}\, \boldsymbol{\beta}_{k\times 1}.$$

Rozdělme si matici $\boldsymbol{X}$ na dvě, $\boldsymbol{X}_{n\times k} = (\boldsymbol{X}_{1_{n\times k_1}}, \boldsymbol{X}_{2_{n\times k_2}})$. $\boldsymbol{X}_1$ bude obsahovat sloupce, o kterých předpokládáme, že je měříme přesně. $\boldsymbol{X}_2$ pak obsahuje sloupce získané s chybou.

Naší úlohou je najít taková $\tilde{\boldsymbol{y}}$ a $\tilde{\boldsymbol{X}} = (\boldsymbol{X}_1,\, \tilde{\boldsymbol{X}}_2)$, která:

1. minimalizují Frobeniovu normu

$$\min_{(\tilde{\boldsymbol{X}},\, \tilde{\boldsymbol{y}})\in\mathbb{R}^{n\times(k+1)}} \|(\boldsymbol{X},\, \boldsymbol{y}) - (\tilde{\boldsymbol{X}},\, \tilde{\boldsymbol{y}})\|_F = \min_{(\tilde{\boldsymbol{X}}_2,\, \tilde{\boldsymbol{y}})\in\mathbb{R}^{n\times(k_2+1)}} \|(\boldsymbol{X}_2,\, \boldsymbol{y}) - (\tilde{\boldsymbol{X}}_2,\, \tilde{\boldsymbol{y}})\|_F,$$

2. $\tilde{\boldsymbol{y}}$ leží v prostoru generovaném sloupci matice $\tilde{\boldsymbol{X}}$

$$\hat{\boldsymbol{y}} \in R(\tilde{\boldsymbol{X}}) = R(\boldsymbol{X_1},\, \tilde{\boldsymbol{X}}_{\boldsymbol{2}}).$$

Řešením pak je libovolné $\tilde{\boldsymbol{\beta}}$, které splňuje

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}. \tag{2.8}$$

Předchozí úloha je velmi podstatná, protože speciálním případem smíšeného modelu je každý model úplně nejmenších čtverců obsahující intercept. V takovém případě $\boldsymbol{X_1} = (1, \ldots, 1)^\mathsf{T}$.

Při řešení úlohy smíšeného modelu se snažíme najít matici, která je co nejblíže matici pozorování. Přičemž sloupce odpovídající matici $\boldsymbol{X_1}$ zachováváme konstantní a pouze upravujeme hodnoty ve sloupcích matice $\boldsymbol{X_2}$ a vektoru $\boldsymbol{y}$. K tomu, jakým způsobem naleznout upravené hodnoty, využijeme následující větu.

Definice 2.23. Mějme matici $\boldsymbol{W}_{n\times k}$ a její singulární rozklad $\boldsymbol{W} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^\mathsf{T}$. Nechť $\mathrm{h}(\boldsymbol{W}) = m$. Operátorem H_r rozumíme

$$H_r(\boldsymbol{W}) = \begin{cases} \boldsymbol{U}\,\tilde{\boldsymbol{\Sigma}}\,\boldsymbol{V}^\mathsf{T},\ \tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0), & \text{pokud } r \le m. \\ \boldsymbol{W}, & \text{jinak.} \end{cases}$$

Věta 2.24 (Zobecnění věty Eckart-Young-Mirsky.)**.** Nechť matice $\boldsymbol{Z} = (\boldsymbol{Z_1}, \boldsymbol{Z_2})$, $\mathrm{h}(\boldsymbol{Z}) = m$ a $\mathrm{h}(\boldsymbol{Z_1}) = l$. Označme P ortogonální projekci na prostor generovaný sloupci podmatice $\boldsymbol{Z_1}$ a $P^{\perp}$ ortogonální projekci na ortogonální komplement tohoto prostoru v $\mathbb{R}^m$.

Pokud $l \leq r$, pak matice

$$\hat{\boldsymbol{Z}}_{\boldsymbol{2}} = P\boldsymbol{Z_2} + H_{r-l}(P^{\perp}\boldsymbol{Z_2})$$

splňuje

$$\inf_{\mathrm{h}(\boldsymbol{Z_1}, \boldsymbol{Z_2'}) \leq r} \|(\boldsymbol{Z_1}, \boldsymbol{Z_2'}) - (\boldsymbol{Z_1}, \boldsymbol{Z_2})\|_F = \|(\boldsymbol{Z_1}, \hat{\boldsymbol{Z}}_{\boldsymbol{2}}) - (\boldsymbol{Z_1}, \boldsymbol{Z_2})\|_F.$$

Důkaz. Viz Golub a kol. (1987). □

Věta 2.24 nám říká, že si úlohu smíšeného modelu můžeme rozdělit na dvě části. Nejdříve se zaměříme na veličiny, které nejsme schopni změřit přesně. Tyto veličiny si zobrazíme do prostoru kolmého na prostor generovaný veličinami, které schopni změřit jsme. Tím zaručíme, že zkoumané proměnné jsou očištěné o vliv těch veličin, kterými se zatím nezabýváme. Takto zúžený model pak odpovídá klasickému modelu TLS – a také ho jako klasický model TLS můžeme řešit.

Po vyřešení části odpovídající modelu TLS nám zbývá odhadnout parametry, které se vztahují k veličinám, u nichž předpokládáme, že jejich pozorování získáváme bez chyby. Což zase odpovídá hledání řešení v modelu OLS.

Z výše uvedeného je zřejmé, proč zkoumaný problém označujeme jako smíšený model. Pro odhad některých parametrů aplikujeme metodu TLS, zatímco jiné odhadujeme pomocí metody OLS.

Ukažme si konkrétně, co výše napsané znamená pro hledání řešení TLS modelu s interceptem. Jako TLS model s interceptem označujeme model, kde všechna pozorování měříme s chybou a zároveň hodnoty vysvětlované proměnné jsou posunuty od počátku souřadnic o neznámou konstantu. Tedy

$$\boldsymbol{y}_{n\times 1} \approx \alpha \mathbf{1}_{n\times 1} + \boldsymbol{X}_{n\times k}\, \boldsymbol{\beta}_{k\times 1}.$$

Nejprve se zaměříme na matici $(\boldsymbol{X}, \boldsymbol{y})$. Označme si jako $\bar{x}_{n,i}$ průměr i-tého sloupce matice $\boldsymbol{X}$. Tj.

$$\bar{x}_{n,i} = \frac{1}{n}\sum_{j=1}^{n} x_{j,i}.$$

Stejně tak si označme průměr vektoru $\boldsymbol{y}$ jako

$$\bar{y}_n = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

Podle věty 2.24 promítáme $(\boldsymbol{X}, \boldsymbol{y})$ na prostor generovaný vektorem $\mathbf{1}$, respektive na komplement tohoto prostoru. V notaci věty si označme tuto projekci jako P,

respektive $P^{\perp}$. Protože projekce je generovaná vektorem jedniček, tak $P(\boldsymbol{X}, \boldsymbol{y})$ znamená, že každý prvek matice projektujeme na průměr odpovídajícího sloupce. A $P^{\perp}(\boldsymbol{X}, \boldsymbol{y})$ pak způsobí, že si matici $(\boldsymbol{X}, \boldsymbol{y})$ centrujeme

$$P(\boldsymbol{X}, \boldsymbol{y}) = \begin{pmatrix} \bar{x}_{n,1} & \bar{x}_{n,2} & \dots & \bar{x}_{n,k} & \bar{y}_n \\ \bar{x}_{n,1} & \bar{x}_{n,2} & \dots & \bar{x}_{n,k} & \bar{y}_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \bar{x}_{n,1} & \bar{x}_{n,2} & \dots & \bar{x}_{n,k} & \bar{y}_n \end{pmatrix} =: (\boldsymbol{X}_{\text{prum}}, \boldsymbol{y}_{\text{prum}}),$$

$$P^{\perp}(\boldsymbol{X}, \boldsymbol{y}) = \begin{pmatrix} x_{1,1} - \bar{x}_{n,1} & \dots & x_{1,k} - \bar{x}_{n,k} & y_1 - \bar{y}_n \\ x_{2,1} - \bar{x}_{n,1} & \dots & x_{2,k} - \bar{x}_{n,k} & y_2 - \bar{y}_n \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} - \bar{x}_{n,1} & \dots & x_{n,k} - \bar{x}_{n,k} & y_n - \bar{y}_n \end{pmatrix} =: (\boldsymbol{X}_{\text{centr}}, \boldsymbol{y}_{\text{centr}}).$$

Parametry $\boldsymbol{\beta}$ odhadujeme metodou TLS v centrované úloze. Vlastně jsme se vektoru jedniček „zbavili“ pomocí ortogonální projekce

$$\begin{gathered} P^{\perp}\boldsymbol{y} \approx P^{\perp}\alpha\mathbf{1} + P^{\perp}\boldsymbol{X}\,\boldsymbol{\beta} \\ \Downarrow \\ \boldsymbol{y}_{\text{centr}} \approx \boldsymbol{X}_{\text{centr}}\,\boldsymbol{\beta}_{k\times 1}. \end{gathered} \tag{2.9}$$

Z modelu (2.9) získáme TLS odhad $\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{X}}_{\text{centr}}$ a $\tilde{\boldsymbol{y}}_{\text{centr}}$, pro které platí

$$\tilde{\boldsymbol{y}}_{\text{centr}} = \tilde{\boldsymbol{X}}_{\text{centr}}\tilde{\boldsymbol{\beta}}.$$

Věta 2.24 nám říká, jakým způsobem přejít z centrovaného modelu zpátky k necentrovanému. Odhad $\tilde{\boldsymbol{\beta}}$ zůstává a platí $\tilde{\boldsymbol{X}} = \boldsymbol{X}_{\text{prum}} + \tilde{\boldsymbol{X}}_{\text{centr}}$ a $\tilde{\boldsymbol{y}} = \boldsymbol{y}_{\text{prum}} + \tilde{\boldsymbol{y}}_{\text{centr}}$.

Zbývá nám odhadnout parametr α. Protože chceme, aby $\tilde{\boldsymbol{y}} = \tilde{\alpha}\mathbf{1} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}$, tak α odhadneme z rovnice

$$\begin{aligned} \boldsymbol{y}_{\text{prum}} + \tilde{\boldsymbol{y}}_{\text{centr}} &= \tilde{\alpha}\mathbf{1} + \boldsymbol{X}_{\text{prum}}\,\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{X}}_{\text{centr}}\,\tilde{\boldsymbol{\beta}} \\ \boldsymbol{y}_{\text{prum}} - \boldsymbol{X}_{\text{prum}}\,\tilde{\boldsymbol{\beta}} &= \tilde{\alpha}\mathbf{1} - \underbrace{(\tilde{\boldsymbol{y}}_{\text{centr}} - \tilde{\boldsymbol{X}}_{\text{centr}}\,\tilde{\boldsymbol{\beta}})}_{0} \\ \boldsymbol{y}_{\text{prum}} - \boldsymbol{X}_{\text{prum}}\,\tilde{\boldsymbol{\beta}} &= \tilde{\alpha}\mathbf{1} \\ &\Downarrow \\ \tilde{\alpha} &= \bar{y}_n - \sum_{j=1}^{k} \bar{x}_{n,j}\tilde{\beta}_j. \end{aligned}$$

Závěrem si shrňme postup hledání řešení modelu s interceptem.

1. Interceptu se nejdříve zbavíme tím, že si pozorované hodnoty vycentrujeme. Poté najdeme TLS řešení centrovaného modelu.
2. Vrátíme se k necentrovanému modelu. Intercept si odhadneme tak, aby nám model platil pro upravené hodnoty pozorování získané v bodu 1.

Poznámka 2.25. Naznačme si, jak bychom postupovali v situaci, kdy má matice $\boldsymbol{X_1}$ větší počet sloupců než jeden. Tedy v modelu uvažujeme více veličin, o kterých předpokládáme, že jsme jejich hodnoty získali přesně.

Přepišme si model z definice 2.22 tak, abychom zdůraznili jaké parametry odpovídají matici $\boldsymbol{X_1}$ a jaké matici $\boldsymbol{X_2}$

$$\boldsymbol{y}_{n\times 1} \approx \boldsymbol{X}_{1_{n\times k_1}} \boldsymbol{\alpha}_{k_1\times 1} + \boldsymbol{X}_{2_{n\times k_2}} \boldsymbol{\beta}_{k_2\times 1}.$$

Stejně jako v TLS modelu s interceptem si model modifikujeme tak, abychom nejdříve řešili pouze veličiny obsažené v $\boldsymbol{X_2}$. Předchozí model jsme o intercept očistili tím, že jsme si matice pozorování vycentrovali. Nyní využijeme projekční matici $\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{X_1}(\boldsymbol{X_1}^\mathsf{T}\boldsymbol{X_1})^{-1}\boldsymbol{X_1}^\mathsf{T}$. Původní model si pomocí ní upravíme na

$$\begin{aligned} \boldsymbol{R}\boldsymbol{y} &\approx \boldsymbol{R}\boldsymbol{X}_1\boldsymbol{\alpha} + \boldsymbol{R}\boldsymbol{X}_2\,\boldsymbol{\beta} \\ &\Downarrow \\ \boldsymbol{R}\boldsymbol{y} &\approx \boldsymbol{R}\boldsymbol{X}_2\,\boldsymbol{\beta}. \end{aligned} \tag{2.10}$$

Protože $\boldsymbol{R}^\mathsf{T} = \boldsymbol{R}$, tak si snadno ověříme, že $\boldsymbol{R}^\mathsf{T}\boldsymbol{R} = \boldsymbol{R}$. Řešení modelu (2.10) si pak můžeme pomocí věty 2.21 vyjádřit jako

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\boldsymbol{X}_2^\mathsf{T}\boldsymbol{R}^\mathsf{T}\boldsymbol{R}\,\boldsymbol{X}_2 - \sigma^2_{k_2+1,\boldsymbol{R}(\boldsymbol{X_2},\boldsymbol{y})}\,\boldsymbol{I})^{-1}\boldsymbol{X}_2^\mathsf{T}\boldsymbol{R}^\mathsf{T}\boldsymbol{R}\,\boldsymbol{y} \\ &= (\boldsymbol{X}_2^\mathsf{T}\boldsymbol{R}\,\boldsymbol{X}_2 - \sigma^2_{k_2+1,\boldsymbol{R}(\boldsymbol{X_2},\boldsymbol{y})}\boldsymbol{I})^{-1}\boldsymbol{X}_2^\mathsf{T}\boldsymbol{R}\,\boldsymbol{y}, \end{aligned} \tag{2.11}$$

kde $\sigma^2_{k_2+1,\boldsymbol{R}(\boldsymbol{X_2},\boldsymbol{y})}$ pochází ze singulárního rozkladu matice $\boldsymbol{R}\,(\boldsymbol{X_2},\boldsymbol{y})$.

Nyní se zaměříme na veličiny v $\boldsymbol{X_1}$. Model si „očistíme“ o $\boldsymbol{X_2}$ následujícím způsobem

$$\boldsymbol{y} - \boldsymbol{X_2}\,\tilde{\boldsymbol{\beta}} \approx \boldsymbol{X_1}\,\boldsymbol{\alpha}.$$

OLS odhad parametrů $\boldsymbol{\alpha}$ si můžeme vyjádřit jako

$$\tilde{\boldsymbol{\alpha}} = (\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}\,(\boldsymbol{y} - \boldsymbol{X_2}\,\tilde{\boldsymbol{\beta}}) \tag{2.12}$$

Výsledné řešení získáme spojením odhadů (2.11) a (2.12).

2.4 Statistické vlastnosti odhadu úplně nejmenších čtverců

Zatím jsme se zabývali pouze hledáním řešení úlohy úplně nejmenších čtverců. V této části se podívejme, jaké má řešení vlastnosti. Ukážeme si, za jakých předpokladů je odhad TLS konzistentní. Důležité pro nás bude i asymptotické rozdělení odhadu – téma, jež bude náplní třetí kapitoly.

Budeme se zabývat základní úlohou úplně nejmenších čtverců dle definice 2.9. Zopakujme, že se omezujeme na modely, kde hodnoty vysvětlujících proměnných

jsou fixní hodnoty. Tj. předpokládáme *functional relationship* dle diskuze na začátku kapitoly. Zároveň budeme předpokládat, že chyby v modelu jsou způsobeny pouze nepřesným měřením.

Přepišme si úlohu tak, abychom si zdůraznili, kde se v modelu vyskytuje náhodnost. Zabýváme se modelem

$$\begin{aligned} \boldsymbol{Y}_{n\times 1} &= \boldsymbol{Z}_{n\times k}\,\boldsymbol{\beta}_{k\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \\ \boldsymbol{X}_{n\times k} &= \boldsymbol{Z}_{n\times k} + \boldsymbol{U}_{n\times k}. \end{aligned} \tag{EIV}$$

Pozorované hodnoty vysvětlujících proměnných máme v matici $\boldsymbol{X}$. Skutečné hodnoty $\boldsymbol{Z}$ neznáme, od změřených hodnot se liší o náhodnou chybu $\boldsymbol{U}$. Stejně tak se vysvětlovaná proměnná $\boldsymbol{Y}$ liší od skutečných hodnot o chybu $\boldsymbol{\varepsilon}$. Nadále budeme předpokládat, že matice neznámých konstant $\boldsymbol{Z}$ je plné hodnosti $\mathrm{h}(\boldsymbol{Z}) = k$.

Upravili jsme notaci $\boldsymbol{y}$ na $\boldsymbol{Y}$, abychom ilustrovali změnu našeho úhlu pohledu na úlohu. V předchozím textu jsme se zaměřovali na algebraickou stránku hledání řešení – vektor $\boldsymbol{y}$ jsme projektovali do vhodného prostoru generovaného upravenými sloupci matice $\boldsymbol{X}$. Nyní chápeme $\boldsymbol{Y}$ jako *náhodný* vektor a budeme se dívat na vlastnosti odhadu jako náhodné veličiny. Přepsali jsme si úlohu do formy, která více odpovídá statistickým modelům lineární regrese. Vyzdvihněme, co jsme zatím mimoděk předpokládali, že pracujeme na konkrétním pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$.

Vektor chyb $(\boldsymbol{U}_{m,.}, \varepsilon_m)$, $m \in \{1, \ldots, n\}$, nám říká, v jakém vztahu jsou chyby odpovídající jednomu pozorování. Budeme předpokládat, že chyby mají střední hodnotu rovnou nule, nejsou mezi sebou korelované a mají shodný rozptyl. O řádcích matice chyb $(\boldsymbol{U}, \boldsymbol{\varepsilon})$ pak budeme předpokládat, že jsou navzájem nezávislé a stejně rozdělené. Zároveň budeme potřebovat, aby sdružené rozdělení $(\boldsymbol{U}, \boldsymbol{\varepsilon})$ bylo absolutně spojité vzhledem k Lebesgueově míře.

Shrňme si předpoklady

$$\begin{aligned} &\mathsf{E}(\boldsymbol{U}_{m,.}, \varepsilon_m) = \boldsymbol{0} \quad \& \quad \mathsf{var}(\boldsymbol{U}_{m,.}, \varepsilon_m) = \sigma^2 \boldsymbol{I}_{(k+1)\times(k+1)}, \\ &\qquad (\boldsymbol{U}_{m,.}, \varepsilon_m) \text{ jsou } iid, \quad \forall m \in \{1, \ldots, n\}, \\ &\text{rozdělení } (\boldsymbol{U}, \boldsymbol{\varepsilon}) \text{ je absol. spojité vzhledem k Lebesqueově míře.} \end{aligned} \tag{2.13}$$

Uveďme si ještě větu, jak získat řešení modelu (EIV) metodou maximální věrohodnosti. Ukáže nám, že MLE-odhad je ekvivalentní s řešením z oddílu 2.3.

Věta 2.26 (MLE-odhad.)**.** Hledejme řešení modelu (EIV). Nechť jsou splněny předpoklady (2.13) a $\mathrm{h}(\boldsymbol{X}) = k$. Nechť $(\boldsymbol{U}_{m,.}, \varepsilon_m)$ má vícerozměrné normální rozdělení. Potom řešením metodou maximální věrohodnosti je

$$\tilde{\boldsymbol{\beta}} = \left[\boldsymbol{X}^\mathsf{T}\boldsymbol{X} - \sigma^2_{k+1,(\boldsymbol{X},\boldsymbol{Y})}\boldsymbol{I}\right]^{-1} \boldsymbol{X}^\mathsf{T}\boldsymbol{Y}.$$

Důkaz. Viz Gallo (1982a). □

Poznámka 2.27. V předpokladech (2.13) jsme se omezovali na kovarianční matici ve formě $\mathsf{var}(\boldsymbol{U}_{m,.}, \varepsilon_m) = \sigma^2 \boldsymbol{I}$. Předpoklad však můžeme zobecnit, aniž by to mělo velký vliv na naší analýzu. Nechť

$$\mathsf{var}(\boldsymbol{U}_{m,.}, \varepsilon_m) = \sigma^2 \boldsymbol{\Sigma}_0, \quad \boldsymbol{\Sigma}_0 \text{ je známá a pozitivně definitní.}$$

Upravme si původní data lineární transformací

$$(\boldsymbol{X}', \boldsymbol{Y}') = (\boldsymbol{X}, \boldsymbol{Y})\, \boldsymbol{\Sigma}_0^{-1/2}.$$

Pokud se zaměříme na model s modifikovanými daty $(\boldsymbol{X}', \boldsymbol{Y}')$, tak jeho kovariační matice chyb je již formy $\mathsf{var}(\boldsymbol{U}'_{m,.}, \varepsilon'_m) = \sigma^2 \boldsymbol{I}$. Tedy pro tento model můžeme najít řešení podle diskuze výše. Gleser (1981) pak ukazuje, jak od řešení pro $(\boldsymbol{X}', \boldsymbol{Y}')$ přejít k řešení původního modelu.

2.4.1 Omezení na matici $\boldsymbol{Z}$

Pro konzistenci TLS odhadu budeme potřebovat, aby matice odpovídající vysvětlujícím proměnným měla určitou strukturu. Uveďme si dvě varianty omezujících podmínek.

A) Silnější předpoklad:

$$\boldsymbol{\Delta} = \lim_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} \tag{2.14}$$

existuje a je pozitivně definitní.

B) Slabší předpoklady:

$$\begin{aligned} &\text{(i)} \qquad n \to \infty : \quad \frac{1}{\sqrt{n}} \lambda_{min}(\boldsymbol{Z}^\mathsf{T} \boldsymbol{Z}) \to \infty, \\ &\text{(ii)} \qquad n \to \infty : \quad \frac{\lambda^2_{min}(\boldsymbol{Z}^\mathsf{T} \boldsymbol{Z})}{\lambda_{max}(\boldsymbol{Z}^\mathsf{T} \boldsymbol{Z})} \to \infty. \end{aligned} \tag{2.15}$$

Jako $\lambda_{min}(\boldsymbol{W})$, $\lambda_{max}(\boldsymbol{W})$ označujeme nejmenší respektive největší vlastní číslo matice $\boldsymbol{W}$.

Předpoklad A) je silnější ve smyslu, že pokud je splněn, platí i oba předpoklady z B).

Věta 2.28. Nechť je pro čtvercovou matici $\boldsymbol{W} = \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z}$ splněn předpoklad (2.14). Potom $\boldsymbol{W}$ splňuje i (2.15).

Důkaz. Z definice vlastního čísla víme, že pokud je λ vlastní číslo matice $\boldsymbol{U}$, pak pro každou konstantu $c \in \mathbb{R}, c \neq 0$ je $c\lambda$ vlastní číslo matice $c\boldsymbol{U}$.

$$\boldsymbol{U} v = \lambda\, v \quad \Leftrightarrow \quad (c\boldsymbol{U})\, v = (c\lambda)\, v.$$

Zároveň pokud je $\boldsymbol{U} = \frac{1}{n}\boldsymbol{W}$ pozitivně definitní matice, pak všechny její vlastní čísla jsou větší jak nula [viz Horn a Johnson (2012), Věta 4.1.10].

Označme $a := \lambda_{min}(U) > 0$. Potom

(i)
$$\frac{1}{\sqrt{n}}\lambda_{min}(\boldsymbol{W}) = \frac{1}{\sqrt{n}}\lambda_{min}(n\,\boldsymbol{U}) = \sqrt{n}\,\lambda_{min}(\boldsymbol{U}) \underset{n\to\infty}{\to} \infty.$$

(ii) Podobně jako v (i) můžeme dokázat, že

$$\lambda_{min}(\boldsymbol{W}) \underset{n\to\infty}{\to} \infty.$$

Protože matice $\boldsymbol{U}$ má všechna vlastní čísla kladná, tak

$$\frac{\lambda_{min}(\boldsymbol{W})}{\lambda_{max}(\boldsymbol{W})} = \frac{n\,\lambda_{min}(\boldsymbol{U})}{n\,\lambda_{max}(\boldsymbol{U})} = b > 0.$$

Spojením pak dostaneme, že

$$\lambda_{min}(\boldsymbol{W})\,\frac{\lambda_{min}(\boldsymbol{W})}{\lambda_{max}(\boldsymbol{W})} \to \infty, \quad n \to \infty.$$

□

Co nám vlastně předpoklady (2.15) říkají o struktuře matice?

První předpoklad můžeme chápat tak, že chceme mezi nezávisle proměnnými dostatečnou variabilitu v jejich pozorováních. Jinými slovy chceme, aby nám další pozorování přinášela novou informaci – pokud budeme získávat stále stejné hodnoty pro všechny proměnné, nejsme schopni odhadnout vliv jednotlivých veličin. Není pak podstatné jestli máme 10 nebo 1000 pozorování. A tedy limitní chování odhadu nám nefunguje, jak bychom chtěli.

V druhém předpokladu se vyskytuje poměr $\frac{\lambda_{max}}{\lambda_{min}}$, který odpovídá podmíněnosti matice $\boldsymbol{Z}^\top\boldsymbol{Z}$. Pokud by byla podmíněnost příliš vysoká, odhad $\tilde{\boldsymbol{\beta}}$ by byl náchylný i na malou změnu v hodnotách nezávislých veličin. Vysoká podmíněnost napovídá o tom, že se v matici $\boldsymbol{Z}$ vyskytuje multikolinearita – jeden sloupec se dá velmi dobře aproximovat lineární kombinací ostatních sloupců. (Podrobné vysvětlení vztahu podmíněnosti a multikolinearity viz např. Belsley a kol. (1980).) Druhý předpoklad si tedy můžeme interpretovat tak, že nechceme ve sloupcích matice $\boldsymbol{Z}$ multikolinearitu.

2.4.2 Konzistence TLS odhadu

Konzistence znamená, že pro velké množství pozorování se odhad přibližuje správným hodnotám parametrů v modelu. V předchozím oddíle jsme se zabývali dvěma druhy omezení na matici $\boldsymbol{Z}$ (respektive $\boldsymbol{Z}^\top\boldsymbol{Z}$). Nyní si odůvodníme z jakého účelu. Ukážeme si, že pro slabou konzistenci odhadu nám budou stačit volnější předpoklady (2.15). Pro silnou konzistenci pak budeme potřebovat přísnější předpoklad (2.14).

Slabá a silná konzistence se od sebe odlišuje v tom, jakým způsobem měříme „blízkost“ odhadu k parametru. Slabá konzistence je konvergence v pravděpodobnosti, silná konvergence skoro jistě.

Definice 2.29 (Konvergence posloupnosti náhodných vektorů)**.** Nechť $(\Omega, \mathcal{A}, \mathsf{P})$ je pravděpodobnostní prostor. $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ je posloupnost reálných náhodných vektorů $\boldsymbol{X}_n : \Omega \to \mathbb{R}^k$, $n \in \mathbb{N}$.

(i) Řekneme, že posloupnost $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ konverguje k náhodnému vektoru $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ *skoro jistě*, pokud

$$\mathsf{P}(\omega \in \Omega : \lim_{n\to\infty} \boldsymbol{X}_n(\omega) = \boldsymbol{X}(\omega)) = 1.$$

Značíme

$$\boldsymbol{X}_n \xrightarrow{s.j.} \boldsymbol{X}, \quad n \to \infty.$$

(ii) Řekneme, že posloupnost $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ konverguje k náhodnému vektoru $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ v *pravděpodobnosti*, pokud

$$\forall \varepsilon \in \mathbb{R},\, \varepsilon > 0 : \lim_{n\to\infty} \mathsf{P}(\omega \in \Omega : \|\boldsymbol{X}_n(\omega) - \boldsymbol{X}(\omega)\|_2 > \varepsilon) = 0.$$

Značíme

$$\boldsymbol{X}_n \xrightarrow{\mathsf{P}} \boldsymbol{X}, \quad n \to \infty.$$

Věta 2.30 (Silná konzistence)**.** Hledejme řešení modelu (EIV). Nechť náhodné chyby splňují předpoklady (2.13). Nechť pro matici $\boldsymbol{Z}$ platí (2.14). Pak TLS odhad konverguje ke skutečné hodnotě parametru skoro jistě

$$\tilde{\boldsymbol{\beta}} \xrightarrow{s.j.} \boldsymbol{\beta}, \quad n \to \infty.$$

Navíc

$$\frac{1}{n}\, \sigma^2_{k+1,(\boldsymbol{X},\boldsymbol{Y})} \xrightarrow{s.j.} \sigma^2, \quad n \to \infty.$$

Důkaz. Viz Gleser (1981), Lemma 3.3 a Důsledek 3.1. □

Věta 2.31 (Slabá konzistence)**.** Hledejme řešení modelu (EIV). Nechť náhodné chyby splňují předpoklady (2.13) a náhodný vektor $(\boldsymbol{U}_{m,.}, \varepsilon_m)$, $m \in \{1, \ldots, n\}$, má rozdělení s konečným čtvrtým momentem. Nechť pro matici $\boldsymbol{Z}$ platí (2.15). Pak TLS odhad konverguje ke skutečné hodnotě parametru v pravděpodobnosti

$$\tilde{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}, \quad n \to \infty.$$

Důkaz. Viz Gallo (1982b), Věta 2.1. □

2.4.3 Asymptotická normalita TLS odhadu

Na závěr této kapitoly se podívejme na asymptotické rozdělení TLS odhadu.

Nejdříve si definujme konvergenci v distribuci pomocí konvergence distribučních funkcí.

Definice 2.32 (Konvergence v distribuci)**.** Mějme posloupnost náhodných vektorů $\boldsymbol{X}_n : \Omega \to \mathbb{R}^k$ na $(\Omega, \mathcal{A}, \mathsf{P})$, $n \in \mathbb{N}$. Buď $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ náhodný vektor definovaný na stejném pravděpodobnostním prostoru. Označme jako $F_{\boldsymbol{X}} : \mathbb{R}^k \to [0, 1]$ distribuční funkci vektoru $\boldsymbol{X} = (X_1, \ldots, X_k)$, tj.

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathsf{P}(X_1 \leq x_1, \ldots, X_k \leq x_k), \quad \boldsymbol{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k .$$

Řekneme, že posloupnost $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ konverguje v distribuci k vektoru $\boldsymbol{X}$, pokud

$$\lim_{n\to\infty} F_{\boldsymbol{X}_n}(\boldsymbol{x}) = F_{\boldsymbol{X}}(\boldsymbol{x}),$$

ve všech bodech spojitosti funkce $F_{\boldsymbol{X}}$.

Značíme

$$\boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{X}.$$

Poznámka 2.33. Alternativně se dá konvergence v distribuci definovat pomocí slabé konvergence indukovaných pravděpodobnostních měr $\mathsf{P}_{\boldsymbol{X}_n} \xrightarrow{w} \mathsf{P}_{\boldsymbol{X}}$. Indukovanou mírou rozumíme rozdělení náhodného vektoru: $\mathsf{P}_{\boldsymbol{X}}(B) = \mathsf{P}(\omega : \boldsymbol{X}(\omega) \in B)$. Proto se někdy místo konvergence v distribuci používá termín slabá konvergence.

Věta 2.34 (Asymptotická normalita)**.** Hledejme řešení modelu (EIV). Nechť náhodné chyby splňují předpoklady (2.13) a náhodný vektor $(\boldsymbol{U}_{m,.}, \varepsilon_m)$, $m \in \{1, \ldots, n\}$, má rozdělení s konečným čtvrtým momentem. Nechť pro matici $\boldsymbol{Z}$ platí (2.14). Pak $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ má asymptoticky vícerozměrné náhodné rozdělení s nulovou střední hodnotou.

Pokud navíc rozdělení řádku chyb $(\boldsymbol{U}_{m,.}, \varepsilon_m)$, $m \in \{1, \ldots, n\}$ má stejný třetí a čtvrtý moment jako odpovídající normální rozdělení, pak

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathbb{N}\left(\boldsymbol{0}, \sigma^2(1 + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta})\left[\boldsymbol{\Delta}^{-1} + \sigma^2\boldsymbol{\Delta}^{-1}\left[(\boldsymbol{I}, \boldsymbol{\beta})(\boldsymbol{I}, \boldsymbol{\beta})^{\mathsf{T}}\right]^{-1}\boldsymbol{\Delta}^{-1}\right]\right). \tag{2.16}$$

Důkaz. Viz Gallo (1982b), Věta 3.3. □

Pokud nepředpokládáme omezení na třetí a čtvrtý moment, pak rozptyl asymptotického vícerozměrného normálního rozdělení má složitou formu, viz Gallo (1982b), kapitola 3.4.

Avšak ani když omezení na vyšší momenty platí, nemůžeme použít rozdělení (2.16) pro konstrukci intervalů spolehlivosti, případně testování hypotéz. Rozptyl je definován pomocí neznámých hodnot parametrů modelu. Klasicky v těchto případech můžeme rozptyl v rozdělení nahradit jeho konzistentním odhadem. Získáváme tak intervaly spolehlivosti Waldova typu – jejich korektnost plyne z Cramér-Slutskyho věty [viz např. van der Vaart (2000), Lemma 2.8].

Velmi hrubě si naznačme, jak bychom mohli ke konzistentnímu odhadu rozptylu dospět. V rozptylu se vyskytují parametry $\sigma^2, \boldsymbol{\beta}, \boldsymbol{\Delta}$. Z věty 2.30 známé konzistentní odhady $\sigma^2, \boldsymbol{\beta}$. Gleser (1981) pak v Lemma 3.4 ukazuje i konzistentní odhad $\boldsymbol{\Delta}$

$$\frac{1}{n}(\boldsymbol{X}^\top \boldsymbol{X} - \sigma^2_{k+1,(\boldsymbol{X},\boldsymbol{Y})}\boldsymbol{I}) \xrightarrow{s.j.} \boldsymbol{\Delta}.$$

Kombinace všech těchto odhadů a aplikace věty 1.12 by nás pak mohla dostat ke konzistentnímu odhadu celého rozptylu.

Těžkosti s vyjádřením rozptylu asymptotického rozdělení nás motivují k tomu, abychom se pokusili najít jinou metodu, jakým způsobem intervaly spolehlivosti konstruovat. Tato metoda bude obsahem následující kapitoly.

Kapitola 3

Metoda bootstrap a její aplikace pro TLS úlohu

Ve větě 2.34 jsme si odvodili, že TLS odhad má asymptoticky vícerozměrné normální rozdělení. Odvodit kovarianční matici tohoto rozdělení je však poměrně složité. Pokud přidáme omezení na třetí a čtvrtý moment rozdělení $(\boldsymbol{U}_{m,.}, \varepsilon_m)$, dospějeme k o něco „snazší“ formě

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathbb{N}\left(\mathbf{0},\, \sigma^2(1+\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta})\left[\boldsymbol{\Delta}^{-1} + \sigma^2\boldsymbol{\Delta}^{-1}\left[(\boldsymbol{I},\boldsymbol{\beta})(\boldsymbol{I},\boldsymbol{\beta})^\mathsf{T}\right]^{-1}\boldsymbol{\Delta}^{-1}\right]\right).$$

Protože je asymptotické rozdělení komplikované, hodilo by se nám teoretickému odvození jeho charakteristik nějakým způsobem vyhnout. Pro sestrojení intervalů spolehlivosti parametrů $\boldsymbol{\beta}$ můžeme použít neparametrický bootstrap. V této kapitole si nejdříve metodu popíšeme a ukážeme, jakým způsobem se dá využít pro řešení našeho problému. Nakonec budeme simulovat její chování na náhodně generovaných datech.

3.1 Neparametrický bootstrap

3.1.1 Koncept

Neparametrický bootstrap je metoda, která nám umožňuje odhadovat rozdělení (charakteristiky) náhodného vektoru, aniž bychom si na vektor dopředu kladli velká omezení. Např. že rozdělení vektoru patří do rodiny normálních rozdělení. Metoda je založená na kombinaci dvou principů

(i) Monte Carlo princip,

(ii) substituční princip.

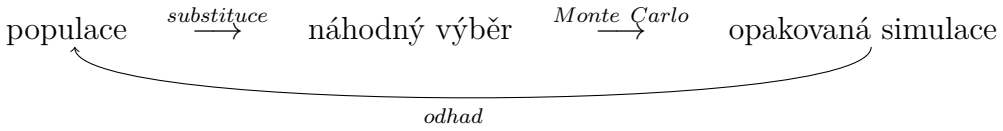
Máme nějakou populaci. Pokud bychom znali všechny její členy, tak bychom samozřejmě hned i věděli, jak je rozdělená. Což však není reálné. Monte Carlo metody jsou využitelné v situacích, kdy můžeme z populace generovat mnoho

náhodných výběrů. Princip metody pak tkví v tom, že pokud máme různých pozorování dostatek, můžeme rozdělení populace aproximovat dostatečně přesně empiricky.

Klasicky ale máme k dispozici pouze jeden náhodný výběr, na jehož základě se snažíme o populaci získat určitou představu. Pokud předpokládáme, že populace sleduje specifický model, tak například odhadujeme nějaký jeho parametr. (V naší situaci máme model (EIV) a hledáme jeho TLS odhad.) O charakteristice populace rozhodujeme na základě náhodného výběru. *Nahrazujeme* si populaci výběrem – proto substituční princip.

Neparametrický bootstrap tedy funguje tak, že z populace získáme náhodný výběr. Protože nemáme velkou znalost o rozdělení tohoto náhodného výběru, snažíme se získat bližší představu pomocí metody Monte Carlo. Například tak, že si replikujeme nové vzorky z původního náhodného výběru. Na základě Monte Carlo simulací pak odhadujeme charakteristiku populace, která nás v analýze zajímá.

Schematicky si neparametrický bootstrap můžeme zobrazit takto



V metodě se dopouštíme aproximace na dvou místech. Pro přesnost aproximace u Monte Carlo jsme omezeni jen dostupnou výpočetní kapacitou. Kolik vzorků je náš „stroj“ schopen v rozumném čase zreplikovat a procesovat. Chyby v přechodu od populace k náhodnému výběru se však zbavit nemůžeme. Nemáme k dispozici jinou informaci než pozorovaná data. Nové znalosti bychom mohli získat jen zopakováním experimentu, což v některých případech není možné.

Jakým způsobem budeme metodu bootstrap používat si ukážeme v oddíle 3.1.3. Nejdříve si ale zadefinujeme, co myslíme pod pojmem podmíněná pravděpodobnost.

3.1.2 Podmíněná pravděpodobnost

Při metodě bootstrap replikujeme vzorky z pozorovaných dat. Tedy simulace probíhá v závislosti na tom, jaká data jsme pozorovali. Rozdělení bootstrap je podmíněné. Bude tak důležité si vyjasnit, co přesně podmiňováním myslíme. Definujme si podmíněnou střední hodnotu a podmíněnou pravděpodobnost pomocí pod-σ-algebry.

Poznámka 3.1. Předpokládáme, že pracujeme s reálnými náhodnými vektory. Tedy, že se jedná o měřitelné zobrazení z pravděpodobnostního prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ do měřitelného prostoru $(\mathbb{R}^k, \mathcal{B}^k)$. Kde jako $\mathcal{B}^k$ označujeme borelovskou σ-algebru na $\mathbb{R}^k$. Výklad bychom ale mohli bez větších obtíží zobecnit na libovolný měřitelný prostor $(\mathbb{E}, \mathcal{E})$.

Definice 3.2 (Podmíněná střední hodnota)**.** Nechť $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ je náhodný vektor na $(\Omega, \mathcal{A}, \mathsf{P})$ s konečnou střední hodnotou, tedy $\mathsf{E}\, X_j < \infty,\ j \in \{1, \ldots, k\}$. Buď $\mathcal{B} \subset \mathcal{A}$ σ-algebra. Jako podmíněnou střední hodnotu $\mathsf{E}(\boldsymbol{X}|\mathcal{B})$ při podmínce $\mathcal{B}$ označíme jakýkoliv náhodný vektor splňující

(i) $\mathsf{E}(\boldsymbol{X}|\mathcal{B}) : \Omega \to \mathbb{R}^k$ je definovaná na $(\Omega, \mathcal{B}, \mathsf{P}/_{\mathcal{B}})$,

(ii) pro každou $B \in \mathcal{B}$ platí

$$\int_B \mathsf{E}(\boldsymbol{X}|\mathcal{B})(\omega)\, \mathsf{dP}(\omega) = \int_B \boldsymbol{X}(\omega)\, \mathsf{dP}(\omega).$$

Definice 3.3 (Podmiňování jinou náhodnou veličinou)**.** Nechť $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ a $\boldsymbol{Y} : \Omega \to \mathbb{R}^k$ jsou náhodné vektory na $(\Omega, \mathcal{A}, \mathsf{P})$. $\boldsymbol{X}$ má konečnou střední hodnotou. Pak definujeme

$$\mathsf{E}(\boldsymbol{X}|\boldsymbol{Y}) := \mathsf{E}(\boldsymbol{X}|\sigma(\boldsymbol{Y})).$$

Kde $\sigma(\boldsymbol{Y})$ je σ-algebra generovaná náhodným vektorem $\boldsymbol{Y}$. Pod $\sigma(\boldsymbol{Y})$ myslíme σ-algebru generovanou množinami

$$\{[\boldsymbol{Y} \in B], B \in \mathcal{B}^k\},$$

kde $[\boldsymbol{Y} \in B] = \{\omega \in \Omega : \boldsymbol{Y}(\omega) \in B\}$.

Pomocí podmíněné střední hodnoty můžeme definovat podmíněnou pravděpodobnost náhodného jevu.

Definice 3.4 (Podmíněná pravděpodobnost)**.** Nechť $(\Omega, \mathcal{A}, \mathsf{P})$ je pravděpodobnostní prostor a $\mathcal{B} \subset \mathcal{A}$ je σ-algebra. Pro jev $A \in \mathcal{A}$ definujeme podmíněnou pravděpodobnost jevu A při podmínce $\mathcal{B}$ jako

$$\mathsf{P}(A|\mathcal{B}) = E(\mathbb{I}_A|\mathcal{B}).$$

3.1.3 Popis metody bootstrap

Mějme náhodný výběr reálných náhodných vektorů $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ s distribuční funkcí F. $[\boldsymbol{X}_i : \Omega \to \mathbb{R}^k,\ i = 1, \ldots, n]$.

Definujme si empirickou distribuční funkci jako

$$F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\boldsymbol{X}_i \leq \boldsymbol{x}],$$

kde

$$\mathbb{I}[\boldsymbol{X}_i \leq \boldsymbol{x}] = \begin{cases} 1, & X_{i,1} \leq x_1 \ \&\ \ldots\ \&\ X_{i,k} \leq x_k, \\ 0, & \text{jinak.} \end{cases}$$

Když jsme se seznamovali s metodou bootstrap, mluvili jsme o kombinaci substitučního a Monte Carlo principu. Empirická distribuční funkce patří k substituční částí. Od neznámého rozdělení F přecházíme k rozdělení náhodného výběru, tj. k empirické distribuční funkci.

Proč k empirické distribuční funkci můžeme přejít nám ukazuje následující věta. Obecně nám také dokazuje validitu metody Monte Carlo. Získáním velkého počtu výběrů z neznámého rozdělení jsme schopni toto rozdělení aproximovat empirickou distribuční funkcí.

Věta 3.5 (Věta Glivenko-Cantelli)**.** Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ jsou *iid* k-rozměrné náhodné vektory s distribuční funkcí $F : \mathbb{R}^k \to [0, 1]$. Potom

$$\sup_{\boldsymbol{y} \in \mathbb{R}^k} |F_n(\boldsymbol{y}) - F(\boldsymbol{y})| \xrightarrow{s.j.} 0, \quad n \to \infty,$$

kde $F_n(\boldsymbol{y})$ je empirická distribuční funkce

$$F_n(\boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\boldsymbol{Y}_i < \boldsymbol{y}].$$

Důkaz. Viz van der Vaart (2000), Věta 19.1. □

Nyní si popíšeme, jak aplikovat neparametrický bootstrap pro náhodný výběr výše.

Naším úkolem je získat představu o nějaké charakteristice rozdělení F. Označme si $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) = \boldsymbol{\theta}(F)$. Pro odhad $\boldsymbol{\theta}$ používáme statistiku $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. Rozdělení $\hat{\boldsymbol{\theta}}_n$ nám napoví o vhodnosti statistiky pro odhad. Z tohoto důvodu si pro větší názornost můžeme $\hat{\boldsymbol{\theta}}_n$ ještě standardizovat pomocí vhodné funkce $\boldsymbol{g}_n$, např. $\boldsymbol{g}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, viz odvozené asymptotické rozdělení TLS odhadu (2.16). Potom nás zajímá rozdělení náhodného vektoru

$$\boldsymbol{R}_n = \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}).$$

Označme si jeho distribuční funkci jako

$$H_n(\boldsymbol{x}) = \mathsf{P}(\boldsymbol{R}_n \leq \boldsymbol{x}).$$

Jakým způsobem zapojit metodu Monte Carlo, když máme k dispozici pouze jeden náhodný výběr? I když rozdělení F neznáme, můžeme se zaměřit na empirickou distribuční funkci F_n. Náhodných výběrů z F_n si totiž můžeme vygenerovat podle libosti. Jednoduše tak, že si realizované hodnoty vektorů $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ zamícháme a budeme z nich náhodně vybírat s opakováním. Získáme tím nový výběr $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$, v němž se některé hodnoty $\boldsymbol{X}_i$ mohou opakovat a naopak některé nemusí vůbec objevit. Každý jeden takový výběr je pak z definice výběr z rozdělení F_n.

V notaci používané výše si označme odhad $\boldsymbol{\theta}$ založený na bootstrapovém výběru jako $\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n^*(\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*)$. A označme si rozdělení $\boldsymbol{R}_n^* = \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n^*, \hat{\boldsymbol{\theta}}_n)$

$$H_n^*(\boldsymbol{x}) = \mathsf{P}(\boldsymbol{R}_n^* \leq \boldsymbol{x} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

(Všimněme si, že se jedná o podmíněné rozdělení. Bootstrap replikace provádíme z jednoho konkrétního náhodného výběru.)

Obecně v metodě bootstrap chceme ukázat, že $\boldsymbol{R}_n^* = \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n^*, \hat{\boldsymbol{\theta}}_n)$ je dobrou aproximací $\boldsymbol{R}_n = \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta})$. Protože představu o rozdělení $\boldsymbol{R}_n^*$ můžeme získat jednoduše z bootstrapových výběrů. Na základě $\boldsymbol{R}_n^*$ pak můžeme získat odhad parametru $\boldsymbol{\theta}$, vytvářet intervaly spolehlivosti, testovat hypotézy.

Doplňme si ještě značení konkrétních bootstrapových výběrů, na jejichž základě získáváme představu o rozdělení $\boldsymbol{R}_n^*$. Budeme mít celkem B bootstrapových výběrů z F_n

$$\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*, \quad b = 1, \ldots, B.$$

Označme si jako $\hat{\boldsymbol{\theta}}_{n,b}^*$ odhad $\boldsymbol{\theta}$ založený na $\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*$. Analogicky označme

$$\boldsymbol{R}_{n,b}^* = \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_{n,b}^*, \hat{\boldsymbol{\theta}}_n).$$

Buď $H_{n,B}^*(\boldsymbol{x})$ empirická distribuční funkce výběru $\boldsymbol{R}_{n,b}^*$, $b = 1, \ldots, B$,

$$H_{n,B}^*(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}[\boldsymbol{R}_{n,b}^* < \boldsymbol{x}].$$

Na následujícím schématu se pokusme názorně shrnout metodu bootstrap popsanou slovně výše.

náhodný výběr		teoretický bootstrap		empirický bootstrap
$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$		$\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$		$\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*$
$\hat{\boldsymbol{\theta}}_n$		$\hat{\boldsymbol{\theta}}_n^*$		$\hat{\boldsymbol{\theta}}_{n,b}^*$
$\boldsymbol{R}_n$		$\boldsymbol{R}_n^*$		$\boldsymbol{R}_{n,b}^*$
H_n	Pešta (2013) $\xleftarrow[(ii)]{}$	H_n^*	Glivenko-Cantelli $\xleftarrow[(i)]{}$	$H_{n,B}^*$

Smyslem metody je co nejvíce využít informace, které nám poskytují pozorovaná data (realizace náhodného výběru $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$). I když nemůžeme získat z dat informaci, která v nich obsažená není, bootstrap replikace nám přinášejí nový užitečný pohled. Pokud navíc platí, že je rozdělení bootstrap odhadu dostatečně blízko rozdělení odhadu z náhodného výběru, můžeme na jeho základě učinit závěry o modelu jako celku.

Pro validitu metody však musí být správné aproximace, kterých se dopouštíme na dvou místech.

(i) První krokem je ukázat, že empirická distribuční funkce $H_{n,B}^*(\boldsymbol{x})$ je dostatečně blízko k rozdělení bootstrap výběru $H_n^*(\boldsymbol{x})$. Podle věty 3.5 vidíme, že toto obecně platí

$$\sup_{\boldsymbol{x} \in \mathbb{R}^k} |H_{n,B}^*(\boldsymbol{x}) - H_n^*(\boldsymbol{x})| \overset{s.j.}{\to} 0.$$

(ii) Zbývá nám se ještě podívat, zda $H_n^*(\boldsymbol{x})$ je správnou aproximací $H_n(\boldsymbol{x})$. Pro důkaz správnosti se klasicky využívá nějaká metrika na prostoru distribučních funkcí a ukazuje se, že pokud se vzdálenost v této metrice mezi $H_n^*(\boldsymbol{x})$ a $H_n(\boldsymbol{x})$ blíží k nule, funkce k sobě konvergují v distribuci. Pro TLS odhad v našem (EIV) modelu je situace obtížnější problém. Využijeme poznatků, které dokázal vedoucí této diplomové práce. Jeho závěry si uvedeme v další sekci.

3.2 Aplikace bootstrap pro TLS model

3.2.1 Podmíněná slabá konvergence dvou náhodných posloupností

Zobecněme si naši definici konvergence v distribuci 2.32. V definici předpokládáme, že existuje nějaké limitní rozdělení $\boldsymbol{X}$ posloupnosti $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$. Tedy předpokládáme nějaký náhodný vektor, ke kterému může posloupnost konvergovat. Což může být v některých situacích omezující.

Mezi výhody neparametrického bootstrapu patří, že neklademe příliš velká omezení na to, jaké rozdělení mají zkoumané veličiny. Bude pozitivní, pokud stejnou „volnost“ zachováme i pro rozdělení limitní.

Zobecnění a definice převezmeme z Belyaev a Luna (2000). Konvergenci v distribuci jsme si definovali pomocí konvergence distribučních funkcí. Pro další výklad bude vhodné, pokud si slabou konvergenci vyjádříme jinou ekvivalentní formou.

Věta 3.6. Mějme posloupnost náhodných vektorů $\boldsymbol{X}_n : \Omega \to \mathbb{R}^k$ na $(\Omega, \mathcal{A}, \mathsf{P})$, $n \in \mathbb{N}$. Buď $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ náhodný vektor definovaný na stejném pravděpodobnostním prostoru. Pak jsou následující tvrzení ekvivalentní

(i)
$$\boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{X},$$

(ii)
$$\lim_{n\to\infty} \mathsf{E}\, f(\boldsymbol{X}_n) = \mathsf{E}\, f(\boldsymbol{X}), \text{ pro každou spojitou omezenou funkci } f : \mathbb{R}^k \to \mathbb{R}.$$

Důkaz. Viz van der Vaart (2000), Věta 2.2. (Jedná se o jednu část z tzv. Portmanteau lemma.) □

Nyní se už můžeme vrátit k zobecnění konvergence v distribuci.

Definice 3.7 (Slabá konvergence dvou posloupností náhodných vektorů)**.** Nechť $(\Omega, \mathcal{A}, \mathsf{P})$ je pravděpodobnostní prostor. Mějme dvě posloupnosti náhodných vektorů $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ a $\{\boldsymbol{Y}_n\}_{n=1}^{\infty}$. $\boldsymbol{X}_n, \boldsymbol{Y}_n : \Omega \to \mathbb{R}^k$, $n \in \mathbb{N}$. Pak řekneme, že posloupnosti $\boldsymbol{X}_n$ a $\boldsymbol{Y}_n$ se k sobě *blíží v distribuci*, pokud

$$\lim_{n\to\infty} \mathsf{E}\, f(\boldsymbol{X}_n) - \mathsf{E}\, f(\boldsymbol{Y}_n) = 0, \text{ pro každou spojitou omezenou funkci } f : \mathbb{R}^k \to \mathbb{R}.$$

Značíme

$$\boldsymbol{X}_n \xleftrightarrow[n\to\infty]{\mathcal{D}} \boldsymbol{Y}_n.$$

Přepišme si definici výše do podmíněné verze. Protože podmíněná střední hodnota je náhodná veličina budeme mluvit o podmíněné slabé konvergenci skoro jistě a v pravděpodobnosti.

Definice 3.8 (Podmíněná slabá konvergence dvou posloupností náhodných vektorů)**.** Nechť $(\Omega, \mathcal{A}, \mathsf{P})$ je pravděpodobnostní prostor. Nechť $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$, $\{\boldsymbol{Y}_n\}_{n=1}^{\infty}$ a $\{\boldsymbol{Z}_n\}_{n=1}^{\infty}$ jsou posloupnosti náhodných vektorů. $\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{Z}_n : \Omega \to \mathbb{R}^k$, $n \in \mathbb{N}$.

(i) Řekneme, že posloupnost $\boldsymbol{X}_n$ se *blíží v distribuci k* $\boldsymbol{Y}_n$ *skoro jistě podle* $\boldsymbol{Z}_n$, pokud pro každou spojitou omezenou funkci $f : \mathbb{R}^k \to \mathbb{R}$

$$\mathsf{E}(f(\boldsymbol{X}_n)|\boldsymbol{Z}_n) - \mathsf{E}\, f(\boldsymbol{Y}_n) \xrightarrow[n\to\infty]{s.j.} 0.$$

Značíme

$$\boldsymbol{X}_n|\boldsymbol{Z}_n \xleftrightarrow[n\to\infty]{\mathcal{D}([\mathsf{P}]\text{-}s.j.)} \boldsymbol{Y}_n.$$

(ii) Řekneme, že posloupnost $\boldsymbol{X}_n$ se *blíží v distribuci k* $\boldsymbol{Y}_n$ *v pravděpodobnosti podle* $\boldsymbol{Z}_n$, pokud pro každou spojitou omezenou funkci $f : \mathbb{R}^k \to \mathbb{R}$

$$\mathsf{E}(f(\boldsymbol{X}_n)|\boldsymbol{Z}_n) - \mathsf{E}\, f(\boldsymbol{Y}_n) \xrightarrow[n\to\infty]{\mathsf{P}} 0.$$

Značíme

$$\boldsymbol{X}_n|\boldsymbol{Z}_n \xleftrightarrow[n\to\infty]{\mathcal{D}(\mathsf{P})} \boldsymbol{Y}_n.$$

Poznámka 3.9. Pokud $\boldsymbol{Y}_n = \boldsymbol{Y}_0$ pro $\forall\, n \geq m, m \in \mathbb{N}$, tak $\boldsymbol{Y}_0$ je limitní rozdělení, kterému jsme se v definicích výše snažili vyhnout. Pokud však existuje, nic nám nebrání nahradit posloupnost $\boldsymbol{Y}_n$ náhodným vektorem $\boldsymbol{Y}_0$ a přeformulovat definice tak, aby odpovídaly nepodmíněné variantě z věty 3.6. Místo toho, že posloupnost $\boldsymbol{X}_n$ se *blíží v distribuci k* $\boldsymbol{Y}_n$ *skoro jistě podle* $\boldsymbol{Z}_n$, bychom řekli, že $\boldsymbol{X}_n$ *konverguje v distribuci k* $\boldsymbol{Y}_0$ *skoro jistě podle* $\boldsymbol{Z}_n$.

3.2.2 Limitní rozdělení bootstrap odhadu

Následující věta nám dokáže legitimitu bootstrapové metody pro náš model.

Věta 3.10 (Shoda bootstrap rozdělení)**.** Řešme model (EIV), pro který platí předpoklady (2.13). Nechť matice $\boldsymbol{Z}$ splňuje (2.14).

Nechť dále platí

$$\sup_{n\in\mathbb{N}} z_{n,j}^2 < \infty, \quad j \in \{1,\dots,k\},$$

$$\sup_{n\in\mathbb{N}} \mathsf{E}\,|U_{n,j}|^8 < \infty, \quad j \in \{1,\dots,k\},$$

$$\sup_{n\in\mathbb{N}} \mathsf{E}\,|\varepsilon|^8 < \infty.$$

Pokud existuje taková pozitivně definitní matice $\boldsymbol{\Upsilon}$, že

$$\liminf_{n\to\infty} \frac{1}{n} \mathsf{var}\left[(\boldsymbol{X},\boldsymbol{Y})^\mathsf{T}(\boldsymbol{X},\boldsymbol{Y})\begin{pmatrix}\boldsymbol{\beta}\\ -1\end{pmatrix}\right] = \boldsymbol{\Upsilon} > \mathbf{0},$$

pak

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}^* - \tilde{\boldsymbol{\beta}})|\,(\boldsymbol{X},\boldsymbol{Y}) \overset{\mathcal{D}(\mathsf{P})}{\underset{n\to\infty}{\longleftrightarrow}} \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{3.1}$$

Důkaz. Viz Pešta (2013), Věta 2.5. □

Kapitola 4

Simulace

V této kapitole si na simulovaných datech ukážeme, jestli dává bootstrap metoda výsledky, které bychom od ní očekávali. V předchozích kapitolách jsme si uvedli asymptotické vlastnosti metody. Nyní se podíváme, jak se bootstrap odhad chová pro různý konečný počet pozorování. Tím získáme bližší představu o rychlosti konvergence a tedy i o tom, jak je metoda v praxi využitelná.

Pešta (2010) simuloval vlastnosti pro jednorozměrný parametr β. My se z jednoho rozměru posuneme do roviny a zaměříme se na dvourozměrný parametr $\boldsymbol{\beta}$.

4.1 *Statistical depth function*

Bootstrap je metoda, která nám dává poměrně jednoduchý způsob, jak odhadnout rozdělení odhadu (samozřejmě pokud je metoda v dané situaci platná). Tedy je dobře využitelná pro konstrukci intervalových odhadů a testování hypotéz.

V jednorozměrném případě si lze jednoduše seřadit jednotlivé bootstrap odhady od nejmenšího po největší. Intervalový odhad si pak můžeme zkonstruovat například pomocí kvantilů takto seřazených odhadů. Označme α-kvantil jako $\theta^*(\alpha)$, pak intervalovým odhadem o spolehlivosti α může být

$$(\theta^*(\alpha/2), \theta^*(1-\alpha/2)).$$

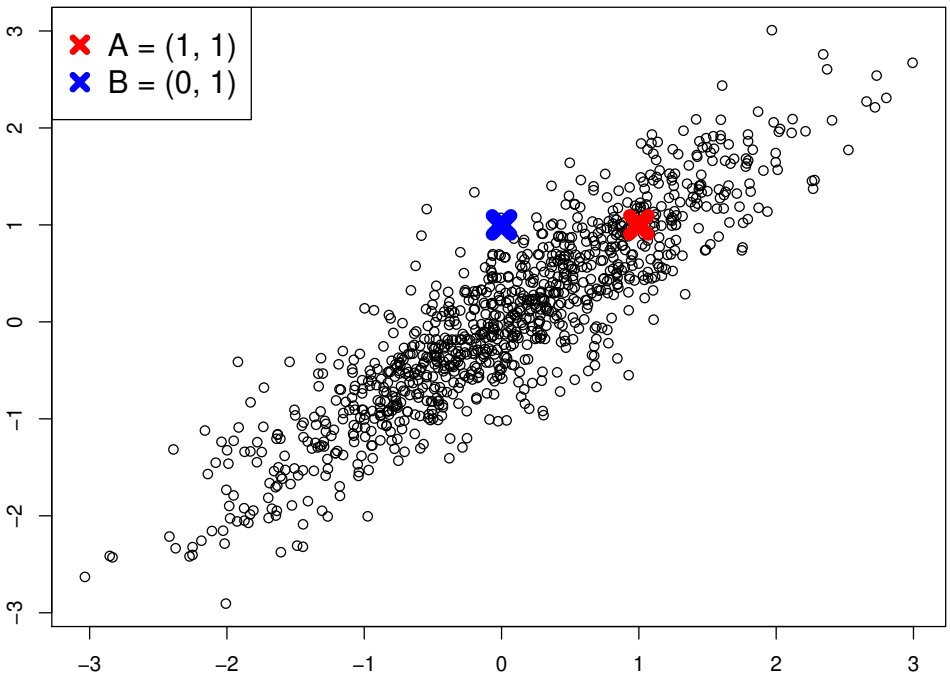
Pokud odhadujeme vícerozměrný parametr, je obtížnější navzájem seřadit jednotlivé odhady. Je větší $\boldsymbol{\beta}_1 = (1,2)$ nebo $\boldsymbol{\beta}_2 = (2,0)$?

Zároveň narážíme na další problém, který si budeme ilustrovat na obrázku 4.1. Nechť náhodný vektor $\boldsymbol{U}$ má vícerozměrné normální rozdělení

$$\boldsymbol{U} \sim \mathsf{N}\left(\boldsymbol{\mu} = (0{,}0),\, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right).$$

Na obrázku vidíme tisíc náhodných výběrů z $\boldsymbol{U}$. Červeným křížkem je označený bod $A = (1,1)$ a modrým bod $B = (0,1)$. I když je bod B blíže (v Eukleidovské

metrice) střední hodnotě rozdělení, je ze simulace patrné, že bod A je v oblasti, kde se náhodné výběry více shlukují. Hustota $\boldsymbol{U}$ je větší v bodě $(1,1)$ než v $(0,1)$. Potřebujeme najít určitý návod, jak určit, zda je bod v „centru“ náhodného výběru nebo na jeho „okraji“.



Obrázek 4.1: Náhodný výběr z $\boldsymbol{U}$ – vícerozměrného normálního rozdělení.

K tomu můžeme využít koncept tzv. *statistical depth function*. Z nedostatku českého ekvivalentu budeme používat anglickou terminologii – český překlad by byl podobný výrazu *statistická hloubková funkce*. Logika za depth function je přiřadit určitému vícerozměrnému bodu hodnotu, jak *hluboko* je daný bod ve zkoumaném rozdělení.

Definice 4.1 (*Statistical depth function*)**.** Označme jako $\mathscr{P}$ třídu rozdělení na $(\mathbb{R}^k, \mathcal{B}^k)$. Buď $D(\cdot;\cdot) : \mathbb{R}^k \times \mathscr{P} \to \mathbb{R}^1$ omezená nezáporná funkce a $\mathsf{P}_{\boldsymbol{X}} \in \mathscr{P}$ rozdělení náhodného vektoru $\boldsymbol{X}$. Řekneme, že funkce D je *statistical depth function*, pokud splňuje

(i) [funkce je invariantní pro afinní transformace]
$D(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}; \mathsf{P}_{\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}}) = D(\boldsymbol{x}; \mathsf{P}_{\boldsymbol{X}}) \quad \forall \boldsymbol{A} \in \mathrm{M}(k \times k),\ |\boldsymbol{A}| \neq 0,\ \forall \boldsymbol{b} \in \mathbb{R}^k,$

(ii) [funkce dosahuje maximální hodnoty v bodech symetrie rozdělení]
$\forall\, \mathsf{P}_{\boldsymbol{Z}} \in \mathscr{P},\ \forall \boldsymbol{\vartheta} \in \mathbb{R}^k$ kolem kterých je rozdělení $\mathsf{P}_{\boldsymbol{Z}}$ symetrické:
$D(\boldsymbol{\vartheta}; \mathsf{P}_{\boldsymbol{Z}}) = \sup_{\boldsymbol{x} \in \mathbb{R}^k} D(\boldsymbol{x}; \mathsf{P}_{\boldsymbol{Z}})$,

(iii) [funkce je monotónní ve směru od nejhlubšího bodu]
$\forall\, \mathsf{P}_{\boldsymbol{Z}} \in \mathscr{P}$ & $\forall \boldsymbol{\eta} \in \mathbb{R}^k$ splňující $D(\boldsymbol{\eta}; \mathsf{P}_{\boldsymbol{Z}}) = \sup_{\boldsymbol{x} \in \mathbb{R}^k} D(\boldsymbol{x}; \mathsf{P}_{\boldsymbol{Z}})$, platí

$$D(\boldsymbol{x}; \mathsf{P}_{\boldsymbol{Z}})) \leq D(\boldsymbol{\eta} + \alpha(\boldsymbol{x} - \boldsymbol{\eta}; \mathsf{P}_{\boldsymbol{Z}})), \quad \alpha \in [0{,}1]\,,$$

(iv) [funkce se v nekonečnu blíží k nule]
$\forall \mathsf{P}_{\boldsymbol{Z}} \in \mathscr{P} : \|\boldsymbol{x}\|_2 \to \infty \Rightarrow D(\boldsymbol{x}; \mathsf{P}_{\boldsymbol{Z}}) \to 0.$

Zopakujme si k čemu má nám funkce sloužit. Naše data se skládají z vícerozměrných pozorování. Depth function nám umožní určit, jak relativně „hluboko“ je jedno dané pozorování oproti ostatním. Tj. pokud pozorujeme jeden shluk dat, ve kterém leží naše pozorování, chceme aby zde funkce dosahovala relativně vysoké hodnoty (vlastnost (*ii*)). Pokud se od shluku dat vzdalujeme, chceme aby hodnota funkce klesala (vlastnost (*iii*)) až v nekonečnu dosáhla své minimální hodnoty (vlastnost (*iv*)). Zároveň pokud změníme měřítko, nechceme, aby se hodnoty funkce měnily (vlastnost (*i*)).

Napravme nyní jeden dluh z definice 4.2. Pracovali jsme s pojmem symetrické rozdělení a pojmem bod symetrie. Myšlenkově si bod symetrie můžeme ztotožnit se shlukem dat v odstavci výše. (Platí pro symetrická unimodální rozdělení, pro symetrická bimodální rozdělení nám bod symetrie leží mezi dvěma shluky.)

Definice 4.2 (Symetrická rozdělení)**.** Nechť $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ je náhodný vektor. Řekneme, že rozdělení vektoru $\boldsymbol{X}$ je kolem bodu $\boldsymbol{\vartheta} \in \mathbb{R}^k$

(i) *C-symetrické*, pokud
$$\boldsymbol{X} - \boldsymbol{\vartheta} \overset{\mathcal{D}}{=} \boldsymbol{\vartheta} - \boldsymbol{X},$$

(ii) *A-symetrické*, pokud
$$\frac{\boldsymbol{X} - \boldsymbol{\vartheta}}{\|\boldsymbol{X} - \boldsymbol{\vartheta}\|_2} \overset{\mathcal{D}}{=} \frac{\boldsymbol{\vartheta} - \boldsymbol{X}}{\|\boldsymbol{X} - \boldsymbol{\vartheta}\|_2},$$

(iii) *H-symetrické*, pokud
$$\forall H \text{ uzavřené poloprostory } \mathbb{R}^k, \boldsymbol{\vartheta} \in H : \mathsf{P}(\boldsymbol{X} \in H) \geq 1/2.$$

(Pro náhodné vektory $\boldsymbol{X}, \boldsymbol{Y}$ značení $\boldsymbol{X} \overset{\mathcal{D}}{=} \boldsymbol{Y}$ znamená rovnost v distribuci.)

Z definic je zřejmé, že *C-symetrie* $\Rightarrow$ *A-symetrie* $\Rightarrow$ *H-symetrie*. Klasicky symetrické rozdělení chápeme pouze ve smyslu *C-symetrie*. Zobecnění na další typy symetrie je vhodné z toho důvodu, abychom definici depth function mohli vztáhnout k co nejširší škále rozdělení.

Podívejme se nyní na konkrétní funkce, které můžeme použít jako statistical depth funciton. Dobrý přehled lze nalézt v Zuo a Sering (2000). My v této práci budeme používat Tukey a Mahalanobis depth function.

Definice 4.3 (*Tukey a Mahalanobis depth function*)**.** Nechť $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ je náhodný vektor a $\mathsf{P}_{\boldsymbol{X}}$ jeho rozdělení. Nechť $\boldsymbol{x}$ je bod v $\mathbb{R}^k$.

(i) Definujme *Tukey depth function* jako
$$HD(\boldsymbol{x}, \mathsf{P}_{\boldsymbol{X}}) := \inf \left\{ \mathsf{P}_{\boldsymbol{X}}(H) : H \text{ je uzavřený poloprostor } \mathbb{R}^k, \boldsymbol{x} \in H \right\}.$$

(ii) Označme *Mahanalobisovu vzdálenost bodů* $\boldsymbol{x}$ *a* $\boldsymbol{y}$ *vzhledem k matici* $\boldsymbol{A}$

$$d^2_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^{\mathsf{T}} \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{y}),$$

$\boldsymbol{A} \in M(k \times k), \boldsymbol{A} > 0, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^k$.

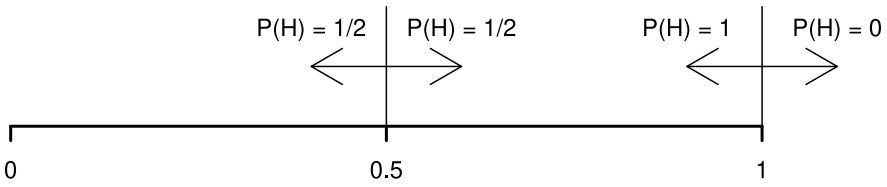
Nechť všechny složky náhodného vektoru mají konečný druhý absolutní moment, tj. $\mathsf{E}\,|X_i|^2 < \infty, i \in \{1, \ldots, k\}$. Pak definujeme *Mahanalobis deph function* jako

$$MD(\boldsymbol{x}, \mathsf{P}_{\boldsymbol{X}}) := \frac{1}{1 + d^2_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{\mu})},$$

kde $\boldsymbol{\mu} = \mathsf{E}\boldsymbol{X}, \boldsymbol{\Sigma} = \mathsf{var}\,\boldsymbol{X}$.

Pro hodnotu Tukey funkce v bodě $\boldsymbol{x}$ hledáme poloprostor s nejmenší pravděpodobností, který bod $\boldsymbol{x}$ obsahuje. Logika za tím je taková, že pokud je bod $\boldsymbol{x}$ ve středu rozdělení, bude pravděpodobnost poloprostorů obsahující $\boldsymbol{x}$ vysoká, bez ohledu jakým směrem poloprostor konstruujeme. A naopak pokud bude bod $\boldsymbol{x}$ na kraji, poloprostor ve směru od rozdělení bude mít nízkou pravděpodobnost.

Znázornění pro rovnoměrné rozdělení $\mathsf{R}(0,1)$ vidíme na obrázku 4.2. Pro bod $x_1 = 0.5$ ležící „uprostřed“ je $HD(x_1, \mathsf{R}) = 0.5$. Pro bod $x_2 = 1$ na „okraji“ pak $HD(x_2, \mathsf{R}) = 0$.



Obrázek 4.2: Tukey depth function pro rovnoměrné rozdělení na intervalu (0,1).

Myšlenku za Mahanalobisovou funkcí si pak můžeme opět ukázat na obrázku 4.1. Měříme vzdálenost bodu $\boldsymbol{x}$ od středu rozdělení (střední hodnoty), ale tuto vzdálenost převážíme ve smyslu kovariance mezi jednotlivými složkami náhodného vektoru. Dále v textu si ještě ukážeme vztah (ekvivalenci) mezi Mahanalobisovou vzdáleností a vrstevnicemi hustoty normálního rozdělení.

Každá z obou definovaných funkcí má své výhody a své nevýhody. Hlavní výhodou Mahalanobis funkce je její relativně nízká výpočetní náročnost. Naproti tomu Tukey funkce je „čistá“ depth function. Má všechny vlastnosti, které jsme si uvedli v definici, aniž bychom museli přidávat omezující předpoklady.

Věta 4.4. Tukey depth function je statistical depth function dle definice 4.2.

Důkaz. Viz Zuo a Sering (2000), Věta 2.1. □

U Mahanalobis depth function však z definice musíme předpokládat konečné druhé momenty. Zároveň potřebujeme, aby i střední hodnota a kovarianční matice byly invariantní pro afinní transformace.

Věta 4.5. Nechť náhodný vektor $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ má symetrické rozdělení $\mathsf{P}_{\boldsymbol{X}}$, jeho složky mají konečný druhý absolutní moment, tj. $\mathsf{E}\,|X_i|^2 < \infty,\ i \in \{1, \ldots, k\}$, a střední hodnota odpovídá bodu symetrie. Nechť $\forall \boldsymbol{A} \in \mathrm{M}(k \times k),\ |\boldsymbol{A}| \neq 0$, $\forall \boldsymbol{b} \in \mathbb{R}^k$ platí

$$\mathsf{E}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\,\mathsf{E}\boldsymbol{X} + \boldsymbol{b},$$

$$\mathsf{var}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}(\mathsf{var}\,\boldsymbol{X})\boldsymbol{A}^{\mathsf{T}}.$$

Potom Mahanalobis depth function je statistical depth function dle definice 4.2.

Důkaz. Viz Zuo a Sering (2000), Věta 2.10. □

Zbývá nám si ujasnit poslední věc, než se pustíme do vlastních simulací. Budeme pracovat s realizací náhodného výběru, ale zatím jsme si pouze zadefinovali teoretické statistical depth function. Je podstatné, jestli teoretické verze uvažovaných funkcí můžeme nahradit výběrovými verzemi.

Definice 4.6 (*Výběrové verze depth function*)**.** Nechť $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ je náhodný výběr z rozdělení $\mathsf{P}_{\boldsymbol{X}}$, $\boldsymbol{X} : \Omega \to \mathbb{R}^k$. Nechť $\boldsymbol{y}$ je bod v $\mathbb{R}^k$. Označme jako $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ realizace náhodného výběru. Pak definujme

(i) *Výběrovou Tukey depth function*

$$SHD(\boldsymbol{y}, \mathsf{P}_n) := \min\left\{\mathsf{P}_n(H) : H \text{ je uzavřený poloprostor } \mathbb{R}^k, \boldsymbol{y} \in H\right\},$$

kde

$$\mathsf{P}_n(H) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\boldsymbol{x}_i \in H],$$

(ii) *Výběrovou Mahanalobis deph function*

$$SMD(\boldsymbol{y}, \mathsf{P}_n) := \frac{1}{1 + d^2_{\hat{\boldsymbol{\Sigma}}}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})},$$

kde $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^{\mathsf{T}}$.

Definice výběrových funkcí odpovídají logice teoretických funkcí. Pro výběrovou Tukey depth function hledáme poloprostor, který obsahuje nejméně bodů z realizovaného náhodného výběru. Pro výběrovou Mahanalobis funkci si neznámé momenty rozdělení nahrazujeme výběrovým průměrem a výběrovou kovarianční maticí.

Věta 4.7 (Stejnoměrná konvergence výběrových statistical depth function)**.** Nechť $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ je náhodný výběr z rozdělení $\mathsf{P}_{\boldsymbol{X}}$, $\boldsymbol{X} : \Omega \to \mathbb{R}^k$.

(i) Pro výběrovou Tukey depth function platí

$$\sup_{\boldsymbol{y} \in \mathbb{R}^k} |HD(\boldsymbol{y}, \mathsf{P}_{\boldsymbol{X}}) - SHD(\boldsymbol{y}, \mathsf{P}_n)| \xrightarrow{s.j.} 0.$$

(ii) Nechť jsou splněny předpoklady věty 4.5, potom

$$\sup_{\boldsymbol{y} \in \mathbb{R}^k} |MD(\boldsymbol{y}, \mathsf{P}_{\boldsymbol{X}}) - SMD(\boldsymbol{y}, \mathsf{P}_n)| \xrightarrow{s.j.} 0.$$

Důkaz.

(i) Viz Donoho a Gasko (1992), Sekce 6.

(ii) Viz Dyckerhoff (2016), Důsledek 4.1.

□

4.2 Uvažované simulace

Na simulacích budeme zkoumat chování bootstrap odhadů pro model (EIV), kde chyby v modelu splňují předpoklady (2.13). Budeme pracovat s dvěma vysvětlujícími proměnnými, tedy model si můžeme schematicky zobrazit jako

$$Y \sim \beta_1 X_1 + \beta_2 X_2.$$

Odhadujeme dvourozměrný parametr $\boldsymbol{\beta} = (\beta_1, \beta_2)$. Jednotlivé bootstrap simulace $\tilde{\boldsymbol{\beta}}^*$ si v prostoru budeme řadit pomocí Tukey a Mahalanobis depth funkce.

Abychom si nasimulovali rychlost asymptotické konvergence, budeme uvažovat čtyři různé hodnoty počtu pozorování $n \in \{20, 50, 100, 1000\}$.

Závislost chování odhadu na pravé hodnotě parametru $\boldsymbol{\beta}$, budeme simulovat na dvou uvažovaných hodnotách parametru.

Pro korektní asymptotické chování bootstrap odhadu potřebujeme ještě dle věty 3.10, aby rozdělení chyb mělo konečný osmý absolutní moment. Budeme uvažovat dvě varianty rozdělení. Kromě vícerozměrného normálního rozdělení budeme v simulacích využívat vícerozměrné t-rozdělení (o 9 stupních volnosti). Abychom viděli, jaký vliv na odhad má rozdělení s „těžšími konci“. Tedy kde je větší pravděpodobnost odlehlých pozorování.

Zároveň se podíváme, jestli bude mít na odhad vliv rozptyl rozdělení. Z každého rozdělení si vygenerujeme náhodné výběry pro dvě různé hodnoty σ^2.

Poslední předpoklad, který musí být splněn, je požadavek na strukturu matice $\boldsymbol{Z}$. Matice pravých (nepozorovaných) hodnot musí splňovat předpoklad (2.14).

Pohybujeme se v dvourozměrném prostoru – označme si

$$\boldsymbol{Z} = \begin{pmatrix} z_{1,1} & z_{2,1} \\ z_{1,2} & z_{2,2} \\ \vdots & \vdots \\ z_{1,n} & z_{2,n} \end{pmatrix} \quad \Rightarrow \quad \frac{1}{n}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n z_{1,i}^2 & \sum_{i=1}^n z_{1,i}\, z_{2,i} \\ \sum_{i=1}^n z_{1,i}\, z_{2,i} & \sum_{i=1}^n z_{2,i}^2 \end{pmatrix}.$$

Pro validitu bootstrap odhadu i dokázanou platnost asymptotického teoretického rozdělení (2.16) potřebujeme, aby $\boldsymbol{\Delta} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ existovala a byla pozitivně definitní.

Budeme uvažovat dvě varianty matice $\boldsymbol{Z}$, které toto budou splňovat

1. $z_{1,i} = i/n,\ z_{2,i} = \sqrt{1 - 1/i}, \quad i \in \{1, \ldots, n\}$,
2. $z_{1,i} = i/n,\ z_{2,i} = i^2/n^2, \quad i \in \{1, \ldots, n\}$.

Konkrétní tvary $z_{j,i}$ jsou do určité míry technického rázu. Abychom zajistili chování matice $\boldsymbol{Z}$ ve smyslu, jak jsme již popisovali v diskuzi u předpokladů (2.15). Nicméně předpoklady můžou odpovídat i reálným situacím.

V druhé variantě říkáme, že přibývajícím n roste do nekonečna interval, na kterém pozorování můžeme získat. A zároveň potřebujeme, aby pozorování byla rozložena po celém intervalu. Což není nepřekonatelný požadavek – s tímto druhem modelu se určitě ve skutečnosti můžeme setkat.

První varianta pak říká, že jedna část pozorování se pohybuje po celé ose a druhá část pozorování se koncentruje v jedné určité oblasti. Opět situace, která v realitě může nastat.

Označme

$$\boldsymbol{\Delta}_1 = \lim_{n\to\infty} \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \frac{i^2}{n^2} & \sum_{i=1}^n \frac{i}{n}\sqrt{1 - \frac{1}{i}} \\ \sum_{i=1}^n \frac{i}{n}\sqrt{1 - \frac{1}{i}} & \sum_{i=1}^n \left(1 - \frac{1}{i}\right) \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix},$$

$$\boldsymbol{\Delta}_2 = \lim_{n\to\infty} \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \frac{i^2}{n^2} & \sum_{i=1}^n \frac{i^3}{n^4} \\ \sum_{i=1}^n \frac{i^3}{n^3} & \sum_{i=1}^n \frac{i^4}{n^4} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

Podle věty 1.6 snadno ověříme, že jsou matice $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ pozitivně definitní.

Poznámka 4.8. Pro výpočet limit jsme použili znalosti o aritmetice limit a o vzorcích pro částečný součet řad

- $\sum_{i=1}^n i = \frac{1}{2}\, n\,(n+1) \ \Rightarrow\ \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n i = \frac{1}{2}$
- $\sum_{i=1}^n i^2 = \frac{1}{6}\, n\,(n+1)\,(2n+1) \ \Rightarrow\ \lim_{n\to\infty} \frac{1}{n^3} \sum_{i=1}^n i^2 = \frac{1}{3}$
- $\sum_{i=1}^n i^3 = \frac{1}{4}\, n^2\,(n+1)^2 \ \Rightarrow\ \lim_{n\to\infty} \frac{1}{n^4} \sum_{i=1}^n i^3 = \frac{1}{4}$

- $\sum_{i=1}^{n} i^4 = \frac{1}{30}\, n\,(n+1)\,(2n+1)\,(3n^2+3n-1) \;\Rightarrow\; \lim_{n\to\infty} \frac{1}{n^5} \sum_{i=1}^{n} i^4 = \frac{1}{5}$

Součet harmonické řady lze až na Eulerovu konstantu aproximovat přirozeným logaritmem [viz Lagarias (2013)], $\lim_{n\to\infty} \left(\sum_{i=1}^{n} \frac{1}{i} - \ln n \right) = \gamma$. Pak

- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{1}{i}\right) = 1 - \lim_{n\to\infty} \frac{\ln n + \gamma}{n} = 1.$

Poslední částí je $\sum_{i=1}^{n} \frac{i}{n} \sqrt{1 - \frac{1}{i}} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{i^2 - i}$. Hledanou limitu najdeme pomocí věty o dvou policistech.

$$\begin{array}{ccccc} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \sqrt{(i-1)^2} & \leq & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \sqrt{i^2 - i} & \leq & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \sqrt{i^2} \\ \downarrow & & & & \downarrow \\ 1/2 & & & & 1/2 \end{array}$$

4.2.1 Varianty simulací

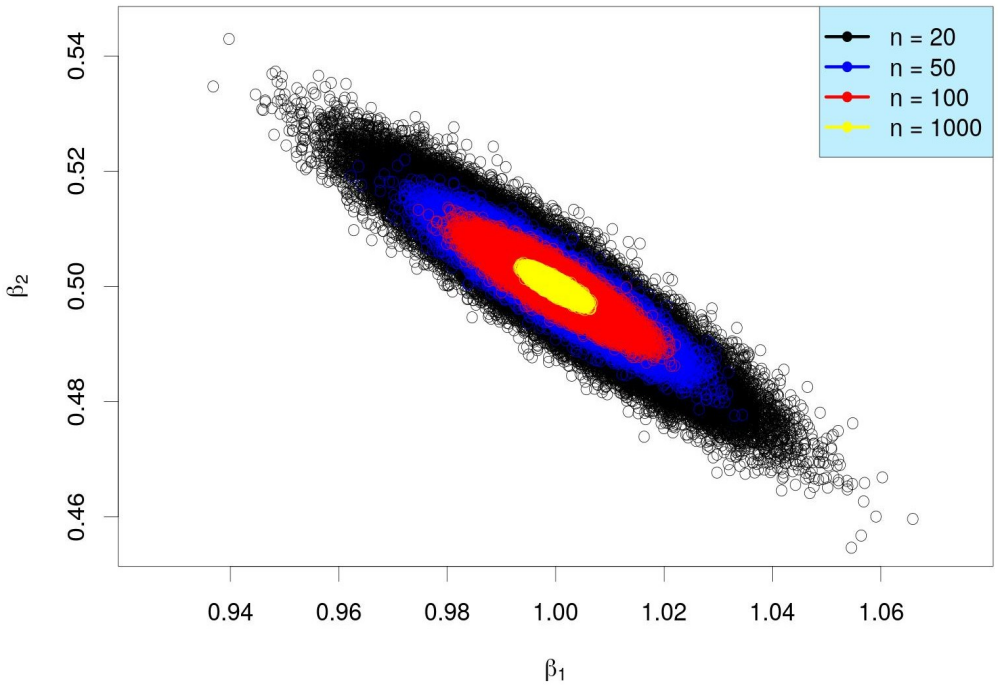
- Dvě skutečné hodnoty parametrů $\boldsymbol{\beta}_1$ a $\boldsymbol{\beta}_2$.
  - $\boldsymbol{\beta}_1 = (1, 0.5)$
  - $\boldsymbol{\beta}_2 = (2, 5)$
- Čtyři různé hodnoty počtu pozorování n_1, n_2, n_3, n_4.
  - $n_1 = 20$
  - $n_2 = 50$
  - $n_3 = 100$
  - $n_4 = 1000$
- Dvě varianty rozdělení chyb.
  - vícerozměrné normální rozdělení
  - vícerozměrné t rozdělení o 9 stupních volnosti
- Dvě varianty směrodatné odchylky každého rozdělení.
  - $\sigma_1 = 10^{-2}$
  - $\sigma_2 = 10^{-3}$
- Dvě varianty matice $\boldsymbol{Z}$.
  - $\boldsymbol{Z}_1$, $z_{1,i} = i/n,\ z_{2,i} = \sqrt{1 - 1/i}, \quad i \in \{1, \ldots, n\}$
  - $\boldsymbol{Z}_2$, $z_{1,i} = i/n,\ z_{2,i} = i^2/n^2, \quad i \in \{1, \ldots, n\}$

Poznámka 4.9. Opět všechny simulace provádíme v R Core Team (2019). Pro simulace byl nastaven náhodný generátor pomocí příkazu `set.seed(78456)`. Pro výpočet konvexního obalu využíváme balík `geometry`.

4.3 „Přesnost“ teoretického asymptotického rozdělení

Víme, že TLS odhad $\tilde{\boldsymbol{\beta}}$ má za určitých předpokladů asymptoticky vícerozměrné normální rozdělení (2.16). Ukažme si na nasimulovaných datech, jak rychle odhad k tomuto rozdělení konverguje. Výsledky si budeme moci porovnat s bootstrap simulacemi v dalších oddílech práce.

Vygenerujeme si 100 000 TLS odhadů podle uvažovaných alternativ simulací v předchozí kapitole. Budeme uvažovat pouze normálně rozdělené chyby. Postup simulace je v podstatě takový, jako kdybychom 100 000x opakovali náš TLS experiment. Tedy pseudonáhodně vygenerujeme chyby, spočteme si TLS odhad a postup opakujeme. Pro bližší představu jsou odhady pro situaci $\beta_1 = 1$, $\beta_2 = 0.5$, $\sigma = 10^{-2}$, $\boldsymbol{\Delta}_1$ zobrazené na obrázku 4.3 v závislosti na různém počtu pozorování n.



Obrázek 4.3: Simulace 100 000 odhadů TLS, $\beta_1 = 1$, $\beta_2 = 0.5$, $\sigma = 10^{-2}$, $\boldsymbol{\Delta}_1$.

Hustota d-rozměrného normálního rozdělení $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ je

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \quad \boldsymbol{x} \in \mathbb{R}^d\,.$$

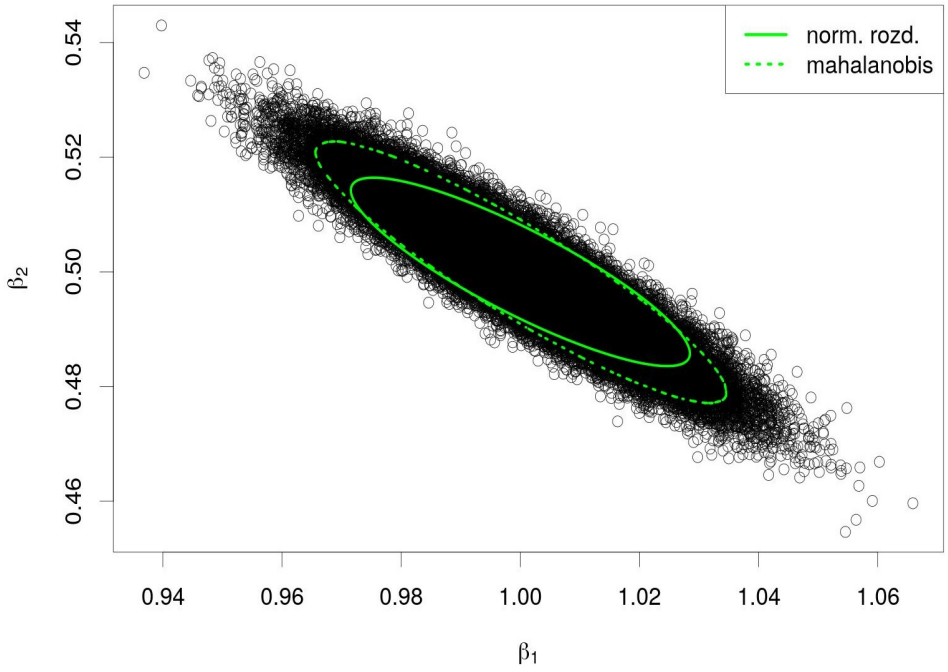
Vrstevnici bodů $\boldsymbol{x}$ se stejnou hustotou tak tvoří elipsoid

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c, \quad c \in \mathbb{R}\,,\, c > 0.$$

Což je Mahalanobisova vzdálenost z Mahalanobis depth function.

Tohoto poznatku využijeme. Budeme porovnávat vrstevnici teoretického asymptotického rozdělení s konvexním obalem, který nám dá simulace seřazená podle Mahalanobis depth function.

Příklad vidíme na obrázku 4.4. Zelená plná čára je 95% elipsa teoretického asymptotického rozdělení. Tím myslíme, že máme pravděpodobnost 95 %, že realizace náhodné veličiny s tímto rozdělením bude ležet uvnitř elipsy. Šrafovaná čára je pak vytvořena ze simulace. Je to konvexní obal opět 95 % bodů, které mají největší hodnotu Mahalanobis depth function. Podle vztahu mezi vrstevnicí a Mahanalobisovou vzdálenosti je patrné, že při shodě asymptotického rozdělení s rozdělením simulace by měly tyto čáry splývat.



Obrázek 4.4: Porovnání vrstevnice teoretického asymptotického rozdělení a konvexního obalu simulace, $\beta_1 = 1,\ \beta_2 = 0.5,\ \sigma = 10^{-2},\ \mathbf{\Delta}_1,\ n = 20$.

V tabulce 4.1 vidíme, jak se simulace chová v situaci $\beta_1 = 1,\ \beta_2 = 0.5,\ \mathbf{\Delta}_1$.

	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*obsah*	*pokrytí (%)*
$\sigma = 10^{-2}$	20	1.000	0.500	1.341	87.26
	50	1.000	0.500	1.139	92.22
	100	1.000	0.500	1.079	93.61
	1000	1.000	0.500	1.005	94.83
$\sigma = 10^{-3}$	20	1.000	0.500	1.320	87.31
	50	1.000	0.500	1.141	92.34
	100	1.000	0.500	1.081	93.54
	1000	1.000	0.500	1.006	94.83

Tabulka 4.1: Simulace asympt. rozdělení odhadu TLS, $\beta_1 = 1,\ \beta_2 = 0.5,\ \mathbf{\Delta}_1$.

Popišme si, co přesně znamenají jednotlivé sloupce v tabulce

- σ – směrodatná odchylka normálního rozdělení chyb v modelu,
- n – počet pozorování, na jejichž základě vytváříme TLS odhad,
- $\bar{\beta}_1$ – průměr první složky z nasimulovaných odhadů $\tilde{\boldsymbol{\beta}}$,
- $\bar{\beta}_2$ – průměr druhé složky z nasimulovaných odhadů $\tilde{\boldsymbol{\beta}}$,
- *obsah* – poměr obsahu konvexního mnohoúhelníku a teoretické vrstevnice,
- *pokrytí (%)* – kolik % nasimulovaných odhadů leží v 95% vrstevnici normálního rozdělení.

Pomocí sloupce *obsah* se snažíme odhadnout, jak moc se 95% oblast ze simulace odlišuje oproti 95% oblasti asymptotického rozdělení. Tedy jak moc je asymptotické rozdělení vzdálené od reality v závislosti na počtu pozorování použitých při TLS odhadu. (Plochu obrazce v dvourozměrném prostoru chápeme v klasickém smyslu. Což odpovídá Lebesguevě míře daného obrazce.) Sloupec *pokrytí* je pak doplnění *obsahu*. Ukazuje, kolik % simulovaných bodů se nachází uvnitř elipsy teoretického rozdělení. Můžeme to brát jako jakousi analogii konfidenční oblasti pro odhad.

Vraťme se k tabulce 4.1. Nepřekvapí, že průměrné odhady jsou velmi blízké skutečným hodnotám parametru $\boldsymbol{\beta}$. Pro malý počet pozorování $n = 20$ není asymptotické rozdělení příliš přesné. Pouze asi 87 % odhadů leží ve vrstevnici normálního rozdělení (kde očekáváme 95 %). Od $n = 50$ už není chyba tak výrazná.

Další tabulky 4.3, 4.4 a 4.5 obsahují stejné údaje pro jiná nastavení počáteční simulace. Ukazuje se, že pro chování asymptotického rozdělení nejsou tak podstatné hodnoty $\boldsymbol{\beta}$ nebo σ jako tvar limitní matice $\boldsymbol{\Delta}$. Vidíme, že pro $\boldsymbol{\Delta_1}$ asymptotické rozdělení plochu podhodnocuje, kdežto pro $\boldsymbol{\Delta_2}$ nadhodnocuje. Hodnota rozptylu σ^2 nemá vliv v relativním smyslu (poměru ploch simulací a asymptotického rozdělení). Samozřejmě absolutně má vliv na velikost plochy, na které se odhady koncentrují. Jak je vidět v tabulce 4.2.

		$\boldsymbol{\Delta}_1$		$\boldsymbol{\Delta}_2$	
	n	$\boldsymbol{\beta} = (1, 0.5)$	$\boldsymbol{\beta} = (2, 5)$	$\boldsymbol{\beta} = (1, 0.5)$	$\boldsymbol{\beta} = (2, 5)$
$\sigma = 10^{-2}$	20	7.3×10^{-4}	9.8×10^{-3}	3.3×10^{-3}	4.4×10^{-2}
	50	2.9×10^{-4}	3.9×10^{-3}	1.3×10^{-3}	1.8×10^{-2}
	100	1.5×10^{-4}	2.0×10^{-3}	6.6×10^{-4}	8.8×10^{-3}
	1000	1.5×10^{-5}	2.0×10^{-4}	6.6×10^{-5}	8.8×10^{-4}
$\sigma = 10^{-3}$	20	7.3×10^{-6}	9.8×10^{-5}	3.3×10^{-5}	4.4×10^{-4}
	50	2.9×10^{-6}	3.9×10^{-5}	1.3×10^{-5}	1.8×10^{-4}
	100	1.5×10^{-6}	2.0×10^{-5}	6.6×10^{-6}	8.8×10^{-5}
	1000	1.5×10^{-7}	2.0×10^{-6}	6.6×10^{-7}	8.8×10^{-6}

Tabulka 4.2: Plocha 95% elipsy teoretického asymptotického rozdělení.

	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*obsah*	*pokrytí (%)*
$\sigma = 10^{-2}$	20	2.000	5.000	1.335	87.19
	50	2.000	5.000	1.141	92.25
	100	2.000	5.000	1.074	93.65
	1000	2.000	5.000	1.004	94.87
$\sigma = 10^{-3}$	20	2.000	5.000	1.339	87.14
	50	2.000	5.000	1.150	92.05
	100	2.000	5.000	1.080	93.56
	1000	2.000	5.000	1.001	94.91

Tabulka 4.3: Simulace asympt. rozdělení odhadu TLS, $\beta_1 = 2,\ \beta_2 = 5,\ \mathbf{\Delta}_1$.

	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*obsah*	*pokrytí (%)*
$\sigma = 10^{-2}$	20	1.000	0.500	0.906	96.18
	50	1.000	0.500	0.963	95.42
	100	1.000	0.500	0.976	95.27
	1000	1.000	0.500	0.991	95.04
$\sigma = 10^{-3}$	20	1.000	0.500	0.904	96.22
	50	1.000	0.500	0.962	95.44
	100	1.000	0.500	0.966	95.43
	1000	1.000	0.500	0.992	95.04

Tabulka 4.4: Simulace asympt. rozdělení odhadu TLS, $\beta_1 = 1,\ \beta_2 = 0.5,\ \mathbf{\Delta}_2$.

	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*obsah*	*pokrytí (%)*
$\sigma = 10^{-2}$	20	1.999	5.002	0.908	96.15
	50	2.000	5.000	0.958	95.52
	100	2.000	5.000	0.980	95.23
	1000	2.000	5.000	0.990	95.06
$\sigma = 10^{-3}$	20	2.000	5.000	0.903	96.25
	50	2.000	5.000	0.955	95.56
	100	2.000	5.000	0.970	95.32
	1000	2.000	5.000	0.992	95.04

Tabulka 4.5: Simulace asympt. rozdělení odhadu TLS, $\beta_1 = 2,\ \beta_2 = 5,\ \mathbf{\Delta}_2$.

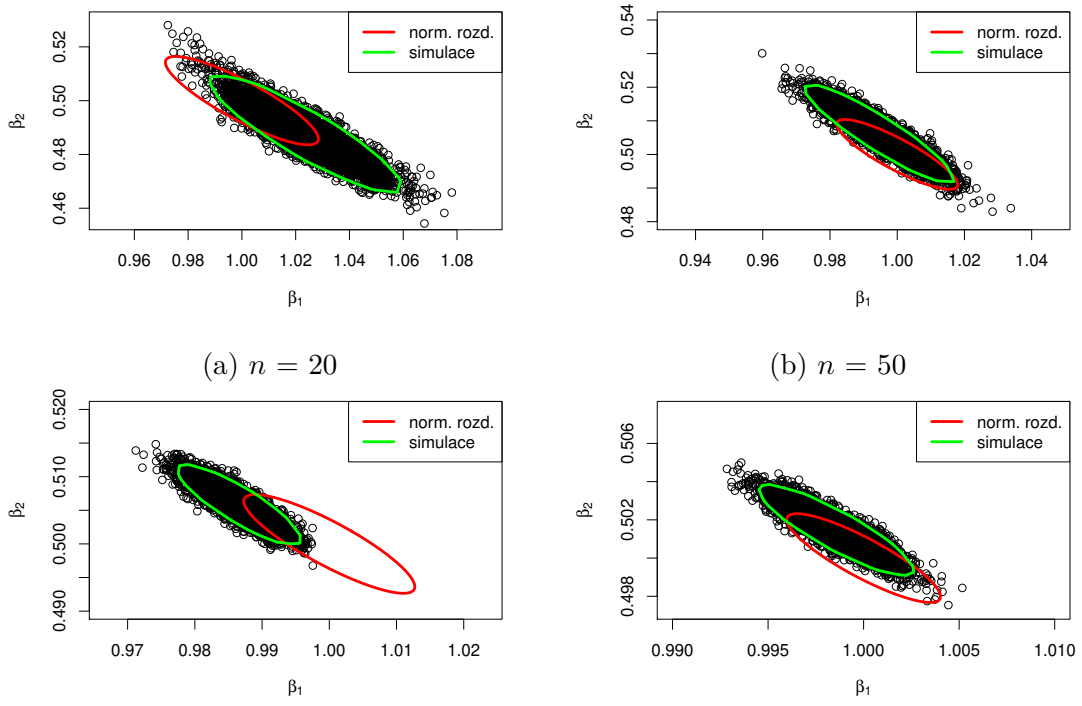
4.4 Simulace – normální rozdělení

Vlastnosti bootstrap metody si budeme ukazovat na základě 1000 simulací náhodných výběrů. Tisíckrát si vygenerujeme náhodný výběr, který odpovídá jedné variantě simulací dle shrnutí 4.2.1. Pro každý vygenerovaný náhodný výběr si vytváříme 5000 bootstrap odhadů. Tyto odhady si v rovině seřadíme podle Tukey a Mahalanobis depth function. A na základě takto seřazených odhadů si vytvoříme konfidenční oblast – např. 95% konfidenční oblast bude tvořit konvexní obal 95 % bodů s nejvyšší hodnotou depth funkce. Výsledný TLS odhad pomocí metody bootstrap vytvoříme jako aritmetický průměr z vygenerovaných 5000 bootstrap odhadů.

Budeme v simulacích zkoumat, jak velká je konfidenční oblast a jestli skutečná hodnota parametru v oblasti leží. Při validitě metody čekáme, že v 95% oblasti bude ležet cca 950 z 1000 provedených simulací. Velikost konfidenční oblasti pak porovnáme s velikostí konfidenčního regionu teoretického asymptotického rozdělení. Které nám tak bude sloužit jako referenční hodnota.

V tomto oddíle se zaměříme na normálně rozdělené chyby, tj. $(\boldsymbol{U}, \varepsilon) \sim \mathsf{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

Pro větší názornost popsaného postupu máme na obrázku 4.5 zobrazené bootstrap TLS odhady z jednoho vygenerovaného náhodného výběru. Červeně je zobrazena 95% konfidenční oblast teoretického asymptotického rozdělení pro daný počet pozorování n. Zeleně pak konvexní obal 95 % bodů s největší hodnotou Mahalanobis depth function.



(a) $n = 20$ (b) $n = 50$

(c) $n = 100$ (d) $n = 1000$

Obrázek 4.5: Jedna simulace bootstrap odhadů, $\beta_1 = 1,\ \beta_2 = 0.5,\ \boldsymbol{\Delta}_1,\ \sigma = 10^{-2}$.

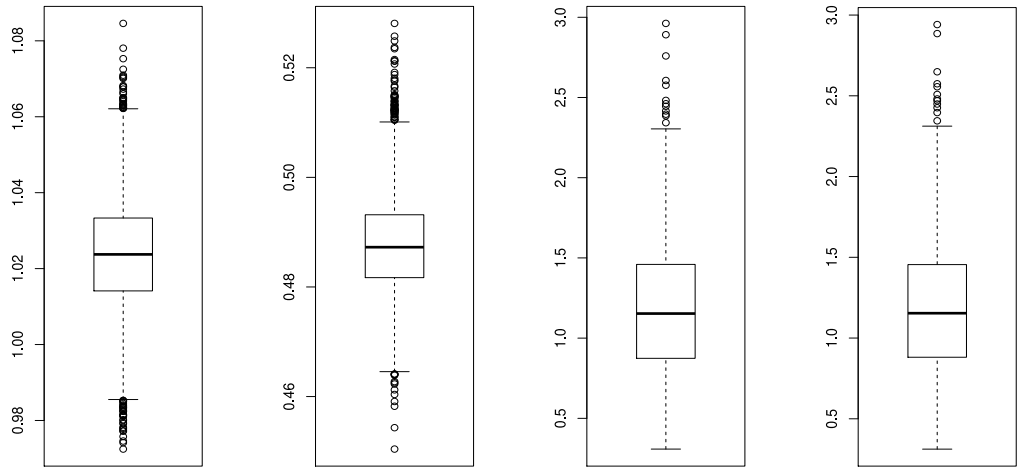
				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	1.024	0.488	7.3×10^{-4}	1.198	90.1	1.197	88.7
	50	0.995	0.506	2.9×10^{-4}	1.069	93.0	1.073	92.9
	100	0.987	0.506	1.5×10^{-4}	1.020	94.5	1.023	94.5
	1000	0.999	0.501	1.5×10^{-5}	0.974	93.9	0.978	94.2
$\sigma = 10^{-3}$	20	1.002	0.499	7.3×10^{-6}	1.196	89.9	1.196	88.7
	50	0.999	0.501	2.9×10^{-6}	1.069	92.7	1.072	92.7
	100	0.999	0.501	1.5×10^{-6}	1.019	94.2	1.023	94.2
	1000	1.000	0.500	1.5×10^{-7}	0.974	94.1	0.978	94.4

Tabulka 4.6: Simulace bootstrap odhadů TLS, $\beta_1 = 1$, $\beta_2 = 0.5$, $\mathbf{\Delta}_1$.

Podívejme se do tabulky 4.6 na výsledky simulací pro $\beta_1 = 1$, $\beta_2 = 0.5$, $\mathbf{\Delta}_1$. $\bar{\beta}_1$ a $\bar{\beta}_2$ označují průměrný TLS bootstrap odhad daného parametru. Logicky pro chyby s větším rozptylem je odhad nepřesnější. Nicméně i tak nám hodnoty napovídají, že odhady ke správným hodnotám konvergují.

Dále je v tabulce pohled na konfidenční oblast dle Mahalanobis a Tukey depth funkce. *Obsah* značí kolikrát je konfidenční region větší než konfidenční oblast referenčního teoretického asymptotického rozdělení. *Pokrytí* pak říká procentuální podíl vygenerovaných konfidenčních oblastí, v nichž se nacházela pravá hodnota parametru $\boldsymbol{\beta}$. Vidíme, že hodnoty *obsah* ani *pokrytí* nezávisí na rozptylu chyb. Od počtu pozorování $n = 50$ se *pokrytí* začíná zezdola přijatelně blížit očekávané hodnotě 95 %.

V tabulce jsou zobrazeny průměry. Lepší představu o variabilitě jednotlivých zkoumaných metrik nám dají krabicové grafy na obrázku 4.6.



(a) $\tilde{\beta}_1$ (b) $\tilde{\beta}_2$ (c) obsah Mahal. (d) obsah Tukey

Obrázek 4.6: Boxplot grafy pro $n = 20$, $\sigma = 10^{-2}$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\mathbf{\Delta}_1$.

V dalších tabulkách 4.7, 4.8 a 4.9 máme ostatní varianty simulací. Chování bootstrap odhadu je obdobné již diskutované situaci. Čeho si můžeme opět jako v oddíle 4.3 všimnout, je rozdílná velikost konfidenční oblasti pro různé varianty matice $\mathbf{\Delta}$. Znovu je velikost konfidenční oblasti pro matici $\mathbf{\Delta}_2$ odhadována menší, než je velikost této oblasti z teoretického asymptotického rozdělení. Co je pozitivní, že simulace nebootstrapového TLS odhadu z předchozí sekce se pohybuje stejným směrem jako bootstrapový odhad. Což by mohlo značit, že bootstrap odhad je méně citlivý (v kladném smyslu slova) na strukturu matice $\boldsymbol{Z}$ než teoretické asymptotické rozdělení.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	2.030	4.972	9.8×10^{-3}	1.191	89.9	1.190	89.4
	50	1.974	5.021	3.9×10^{-3}	1.064	92.5	1.068	92.5
	100	2.002	4.996	2.0×10^{-3}	1.017	93.2	1.021	93.6
	1000	1.997	5.002	2.0×10^{-4}	0.974	94.7	0.977	94.6
$\sigma = 10^{-3}$	20	2.003	4.997	9.8×10^{-5}	1.189	89.6	1.189	89.4
	50	1.998	5.002	3.9×10^{-5}	1.064	92.7	1.067	92.8
	100	2.000	5.000	2.0×10^{-5}	1.017	93.2	1.020	93.4
	1000	2.000	5.000	2.0×10^{-6}	0.974	94.9	0.977	94.6

Tabulka 4.7: Simulace bootstrap odhadů TLS, $\beta_1 = 2,\ \beta_2 = 5,\ \mathbf{\Delta}_1$.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	0.984	0.528	3.3×10^{-3}	0.855	89.1	0.861	88.6
	50	1.018	0.481	1.3×10^{-3}	0.897	92.2	0.900	91.8
	100	1.012	0.477	6.6×10^{-4}	0.922	93.7	0.925	93.4
	1000	1.003	0.497	6.6×10^{-5}	0.962	95.0	0.965	95.3
$\sigma = 10^{-3}$	20	0.998	0.503	3.3×10^{-5}	0.852	89.2	0.857	88.7
	50	1.002	0.498	1.3×10^{-5}	0.896	92.4	0.900	92.3
	100	1.001	0.498	6.6×10^{-6}	0.921	94.2	0.924	93.7
	1000	1.000	0.500	6.6×10^{-7}	0.962	95.3	0.966	95.4

Tabulka 4.8: Simulace bootstrap odhadů TLS, $\beta_1 = 1,\ \beta_2 = 0.5,\ \mathbf{\Delta}_2$.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	1.925	5.085	4.4×10^{-2}	0.850	90.9	0.856	89.3
	50	2.030	4.968	1.8×10^{-2}	0.898	93.0	0.902	92.8
	100	1.991	5.008	8.8×10^{-3}	0.920	93.4	0.923	93.4
	1000	2.000	5.000	8.8×10^{-4}	0.961	95.3	0.964	94.9
$\sigma = 10^{-3}$	20	1.992	5.009	4.4×10^{-4}	0.844	90.6	0.851	89.4
	50	2.003	4.996	1.8×10^{-4}	0.897	93.4	0.901	92.9
	100	1.999	5.001	8.8×10^{-5}	0.918	93.2	0.922	93.0
	1000	2.000	5.000	8.8×10^{-6}	0.961	94.8	0.964	94.9

Tabulka 4.9: Simulace bootstrap odhadů TLS, $\beta_1 = 2,\ \beta_2 = 5,\ \mathbf{\Delta}_2$.

4.5 Simulace – t-rozdělení

V tomto oddíle se podíváme, jaký vliv bude mít na simulace, pokud chyby budou mít jiné než normální rozdělení. Nepřesnosti měření pro vysvětlovanou proměnnou Y i vysvětlující proměnné X_1 a X_2 budou navzájem nezávislé a budou sledovat t-rozdělení o 9 stupních volnosti. Počet stupňů volnosti jsme zvolili z toho důvodu, aby rozdělení mělo konečný 8. moment. Předpoklad, jenž potřebujeme pro korektní chování bootstrap metody.

V shrnutí variant simulací jsme uváděli, že budeme pracovat s dvěma variantami odchylky chyb $\sigma_1 = 10^{-2}$ a $\sigma_2 = 10^{-3}$. Protože t-rozdělení o 9 stupních volnosti má rozptyl $\frac{9}{7}$ [viz Anděl (2011), str. 28], upravíme si t-rozdělení ještě konstantou tak, aby mělo požadovanou odchylku. Tj. vynásobíme $\sqrt{7/9}\,\sigma_1$ nebo $\sqrt{7/9}\,\sigma_2$.

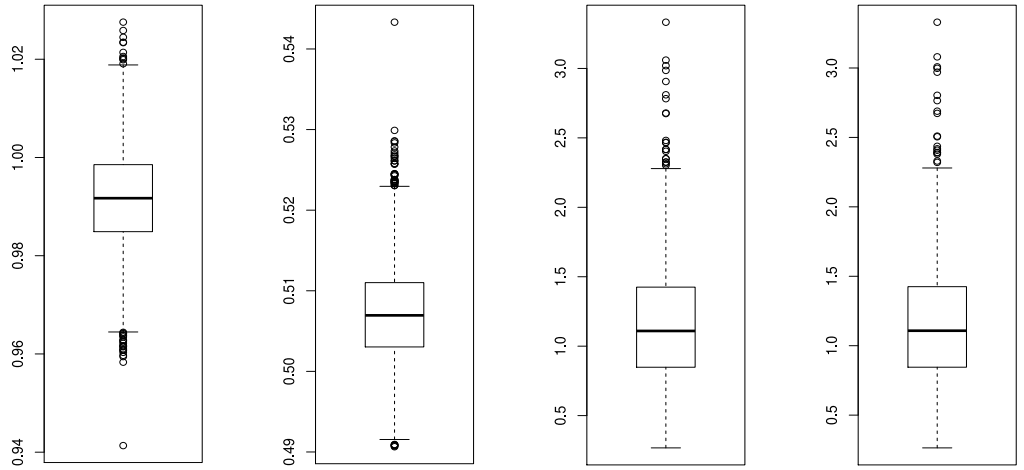
Vlastní simulace pak provádíme stejně jako v předchozím oddíle. V tabulce 4.10 vidíme výsledky pro naši základní variantu $\beta_1 = 1,\ \beta_2 = 0.5,\ \mathbf{\Delta}_1$. Je patrné, že průměrné hodnoty odhadů parametrů se již od malého počtu pozorování blíží ke skutečným hodnotám.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	0.992	0.507	7.3×10^{-4}	1.172	88.3	1.173	87.7
	50	0.995	0.502	2.9×10^{-4}	1.056	91.6	1.059	90.6
	100	1.006	0.495	1.5×10^{-4}	1.010	92.1	1.014	92.2
	1000	1.000	0.500	1.5×10^{-5}	0.970	94.7	0.973	94.7
$\sigma = 10^{-3}$	20	0.999	0.501	7.3×10^{-6}	1.170	88.4	1.171	87.3
	50	1.000	0.500	2.9×10^{-6}	1.055	91.7	1.059	91.0
	100	1.001	0.499	1.5×10^{-6}	1.009	92.1	1.013	92.5
	1000	1.000	0.500	1.5×10^{-7}	0.970	94.5	0.973	94.7

Tabulka 4.10: Simulace bootstrap odhadů TLS, t-rozdělení, $\beta_1 = 1,\ \beta_2 = 0.5,\ \mathbf{\Delta}_1$.

Zaměřme se nyní na shody a rozdíly oproti situaci s normálně rozdělenými chybami. Mahalanobis a Tukey depth function dávají opět srovnatelné výsledky. Stejně tak *obsah* je hodně podobný hodnotám v tabulce 4.6. Co je pro t-rozdělení pomalejší, je konvergence *pokrytí* k očekávané hodnotě 95 %. Což není překvapivé, protože pro t-rozdělení je větší pravděpodobnost odlehlých pozorování než u normálního rozdělení. Tedy můžeme očekávat, že vygenerovaná konfidenční oblast bude s vyšší pravděpodobností ležet mimo pravou hodnotu parametru.

Z důvodu, že častěji pozorujeme odlehlé pozorování, zobrazme si ještě detail o jednotlivých metrikách v krabicových grafech na obrázku 4.7. I když i tady pozorujeme větší rozptyl hodnot než na obrázku 4.6, nejsou rozdíly v grafech výrazné.



(a) $\tilde{\beta}_1$ (b) $\tilde{\beta}_2$ (c) obsah Mahal. (d) obsah Tukey

Obrázek 4.7: Boxplot grafy, t-rozdělení, $n = 20$, $\sigma = 10^{-2}$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\mathbf{\Delta}_1$.

V tabulkách 4.11, 4.12, 4.13 jsou uvedeny hodnoty pro ostatní varianty simulací. Závěry už ale jenom opakujeme. Ukazuje se, že bootstrap metoda funguje dobře i pro chyby, které mají t-rozdělení. Nicméně pro konstrukci správné 95% konfidenční oblasti potřebujeme vyšší počet pozorování než v situaci normálního rozdělení.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	1.988	5.015	9.8×10^{-4}	1.202	89.9	1.203	89.0
	50	1.977	5.004	3.9×10^{-4}	1.058	92.0	1.062	91.9
	100	1.990	5.011	2.0×10^{-4}	1.012	93.2	1.015	93.0
	1000	2.004	4.999	2.0×10^{-5}	0.972	93.8	0.976	94.0
$\sigma = 10^{-3}$	20	1.998	5.002	9.8×10^{-6}	1.199	89.9	1.200	88.9
	50	1.998	5.000	3.9×10^{-6}	1.057	92.0	1.061	92.0
	100	1.999	5.001	2.0×10^{-6}	1.011	93.4	1.015	93.0
	1000	2.000	5.000	2.0×10^{-7}	0.972	94.1	0.975	94.0

Tabulka 4.11: Simulace bootstrap odhadů TLS, t-rozdělení, $\beta_1 = 2$, $\beta_2 = 5$, $\mathbf{\Delta}_1$.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	1.026	0.468	3.3×10^{-3}	0.829	88.7	0.836	86.9
	50	1.002	0.495	1.3×10^{-3}	0.884	92.0	0.889	91.3
	100	0.992	0.509	6.6×10^{-4}	0.918	93.1	0.921	92.7
	1000	0.997	0.504	6.6×10^{-5}	0.957	94.6	0.961	94.3
$\sigma = 10^{-3}$	20	1.002	0.497	3.3×10^{-5}	0.825	88.4	0.832	86.5
	50	1.000	0.499	1.3×10^{-5}	0.882	91.5	0.886	90.9
	100	0.999	0.501	6.6×10^{-6}	0.916	92.3	0.919	92.1
	1000	1.000	0.500	6.6×10^{-7}	0.956	95.2	0.959	95.0

Tabulka 4.12: Simulace bootstrap odhadů TLS, t-rozdělení, $\beta_1 = 1,\ \beta_2 = 0.5$, $\mathbf{\Delta}_2$.

				obsah teor.	*Mahalanobis*		*Tukey*	
	n	$\bar{\beta}_1$	$\bar{\beta}_2$	*elipsy*	*obsah*	*pokrytí*	*obsah*	*pokrytí*
$\sigma = 10^{-2}$	20	2.040	4.959	4.4×10^{-2}	0.853	89.8	0.859	88.1
	50	2.017	4.956	1.8×10^{-2}	0.892	92.1	0.896	91.7
	100	2.069	4.916	8.8×10^{-3}	0.920	92.4	0.923	92.2
	1000	1.987	5.019	8.8×10^{-4}	0.957	93.2	0.960	93.3
$\sigma = 10^{-3}$	20	2.003	4.998	4.4×10^{-4}	0.850	90.2	0.856	88.4
	50	2.001	4.996	1.8×10^{-4}	0.890	91.8	0.894	91.6
	100	2.007	4.991	8.8×10^{-5}	0.919	92.7	0.922	92.3
	1000	1.998	5.002	8.8×10^{-6}	0.955	93.5	0.959	93.4

Tabulka 4.13: Simulace bootstrap odhadů TLS, t-rozdělení, $\beta_1 = 2,\ \beta_2 = 5$, $\mathbf{\Delta}_2$.

4.6 Shrnutí simulací

V předchozích oddílech jsme se podívali na tři situace. Simulovali jsme si chování teoretického asymptotického rozdělení (2.16). Pak jsme si generovali data pro metodu bootstrap – nejprve jsme uvažovali chyby normálně rozdělené. Poté jsme pro simulace použili rozdělení s „těžšími“ konci – vícerozměrné t-rozdělení.

Naším hlavním cílem bylo zjistit, jak se bootstrap odhady chovají pro různá nastavení modelu dle 4.2.1. Teoretické asymptotické rozdělení nám sloužilo jako vzor, s kterým metodu bootstrap srovnáváme. Ukázalo se, že při splnění předpokladů metody nám bootstrap dává dobré výsledky. Pro velikost konfidenční oblasti dokonce lepší než teoretické asymptotické rozdělení. U teoretického asymptotického rozdělení byla podstatná struktura matice $\boldsymbol{Z}$.

V uvažovaných simulacích nezáleželo, zda jsme použili Tukey nebo Mahalanobis depth function. Pozitivní je, že jsme nepozorovali vliv pravé hodnoty parametru na výsledek. Oproti normálnímu rozdělení chyb potřebujeme pro t-rozdělení více pozorovaní, abychom věřili, že konstruovaná konfidenční oblast pokrývá pravou hodnotu parametru s dostatečnou jistotou.

Závěr

Tato práce se zabývala metodou úplně nejmenších čtverců. Metodou, která slouží pro odhad parametrů v lineárních modelech.

První polovina práce se věnuje obecnému popisu metody. Při výkladu jsme postupovali od jednoduchého ke složitějšímu. Vysvětlili jsme si základní myšlenku na vzorové úloze, která nám následně posloužila jako vodítko pro pochopení exaktnějšího popisu metody.

V druhé polovině jsme se zaměřili na asymptotické chování metody a jakým způsobem metodu implementovat pro reálné využití. Vysvětlili jsme si, jak odhadovat parametry pomocí bootstrap. Simulacemi jsme pak odpovídali na otázku, pro jaký počet pozorování dává bootstrap již dobré výsledky.

Celá práce byla koncipována tak, aby byla problematika čtenáři vysvětlena srozumitelně. S ambicemi, aby práci mohli využít kolegové studenti jako svůj úvod do metody úplně nejmenších čtverců.

Hlavním přínosem práce je závěrečná kapitola se simulacemi. Kde jsme validovali použití metody bootstrap pro získání TLS odhadu parametrů modelu. Základem nám byli teoretické poznatky dokázané vedoucím této diplomové práce. Pro simulace jsme se zaměřili na dvourozměrný parametr a dívali se, jak se bootstrap TLS odhad chová pro různý počet pozorování v náhodném výběru.

Klíčovou otázkou pro simulace bylo, jakým způsobem řadit jednotlivé bootstrap odhady v rovině. Odpovědí nám byly statistical depth function. Jejich kombinace s bootstrap je právě podstatou přínosu práce. Konkrétně jsme použili Mahalanobis a Tukey depth function.

Simulace ukázaly, že metoda má očekávané chování. Již od malého počtu dostupných pozorování se odhad parametru blíží jeho reálné hodnotě. Pro správnou konstrukci konfidenční oblasti je však třeba již o něco více dat. Nemělo velký vliv, zda chyby v modelu měly normální či t-rozdělení. Jen pro t-rozdělení je vhodné mít pro korektní chování odhadu konfidenční oblasti větší počet pozorování.

Důležitou poznámkou je, že jsme simulovali, jak metoda funguje při splnění svých předpokladů. Ukázali jsme, že v těchto případech je metoda užitečná a v praxi použitelná. Logickým pokračováním pro další výzkum by bylo se zaměřit na situace, kdy předpoklady neplatí. Simulovat, jaký vliv na výsledek má který předpoklad. A tím získat bližší představu o tom, jak moc je metoda robustní.

Použité značení a zkratky

OLS	metoda nejmenších čtverců
TLS	metoda úplně nejmenších čtverců
$\mathbb{N}$	množina přirozených čísel
$\mathbb{R}$	těleso reálných čísel
$\boldsymbol{x}, \boldsymbol{x}_{(n)}$	sloupcový vektor, $\boldsymbol{x} = (x_1, \ldots, x_n)^T$
$\bar{x}$	výběrový průměr vektoru pozorování $(x_1, \ldots, x_n)^T$
s_x^2	výběrový rozptyl vektoru pozorování $(x_1, \ldots, x_n)^T$
s_{xy}^2	výběrová kovariance vektorů pozorování $(x_1, \ldots, x_n)^T$ a $(y_1, \ldots, y_n)^T$
$\mathrm{M}(n \times m)$	množina všech matic s reálnými prvky o n řádcích a m sloupcích
$\boldsymbol{X}, \boldsymbol{X}_{n\times m}$	matice $\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,m} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,m} \end{pmatrix}$
$\mathrm{tr}(\boldsymbol{X})$	stopa čtvercové matice $\boldsymbol{X}_{n\times n}$ $\mathrm{tr}(\boldsymbol{X}) = \sum_{i=1}^{n} x_{i,i}$
$R(\boldsymbol{X})$	vektorový prostor generovaný sloupci matice $\boldsymbol{X}$
$\boldsymbol{x_{m,\cdot}}$	řádek m matice $\boldsymbol{X_{n\times k}}$ $\boldsymbol{x_{m,\cdot}} = (x_{m,1}, \ldots, x_{m,k})$
$\boldsymbol{I}, \boldsymbol{I}_{n\times n}$	jednotková matice

$$\boldsymbol{I} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}}_{n} \Bigg\} n$$

$\mathbf{1}, \mathbf{1}_{m\times n}$ matice jedniček

$$\mathbf{1} = \underbrace{\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}}_{n} \Bigg\} m$$

$\mathbb{I}_A$ identifikátor množiny A

$$\mathbb{I}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Y náhodná veličina

$\boldsymbol{Y}$ náhodný vektor $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\mathsf{T}$

$\mathsf{R}(a,b)$ rovnoměrné rozdělení na intervalu (a,b)

$\mathsf{N}(\mu, \sigma^2)$ normální rozdělení o střední hodnotě μ a rozptylu σ^2

$F_n(\boldsymbol{x})$ empirická distribuční funkce vektoru pozorování $\boldsymbol{X}$

$F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\boldsymbol{X}_i \leq \boldsymbol{x}]$,

$$\mathbb{I}[\boldsymbol{X}_i \leq \boldsymbol{x}] = \begin{cases} 1, & X_{i,1} \leq x_1 \ \& \ \dots \ \& \ X_{i,k} \leq x_k, \\ 0, & \text{jinak.} \end{cases}$$

Seznam použité literatury

Aldrich, J. (1993). Reiersøl, Geary and the idea of instrumental variables. *The Economic and Social Review*, **24**(3), 243 – 273. ISSN 0012-9984. URL `http://hdl.handle.net/2262/64869`.

Anděl, J. (2011). *Základy matematické statistiky*. 3. vydání. Univerzita Karlova v Praze, Matematicko-fyzikální fakulta. ISBN 9788073781620.

Belsley, D. A., Kuh, E. a Welsch, R. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. John Wiley & Sons. ISBN 9780471058564.

Belyaev, Y. a Luna, S. (2000). Weakly approaching sequences of random distributions. *Journal of Applied Probability*, **37**, 807–822. doi: 10.1239/jap/1014842838.

Donoho, D. L. a Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, **20**(4), 1803–1827. doi: 10.1214/aos/1176348890.

Dupač, V. a Hušková, M. (2009). *Pravděpodobnost a matematická statistika*. Učební texty Univerzity Karlovy v Praze. Karolinum. ISBN 9788024600093.

Dyckerhoff, R. (2016). Convergence of depths and depth-trimmed regions. *arXiv:1611.08721v2*. URL `https://arxiv.org/pdf/1611.08721.pdf`.

Eckart, C. a Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**(3), 211–218. doi: 10.1007/BF02288367.

Fromkorth, A. a Kohler, M. (2011). Analysis of least squares regression estimates in case of additional errors in the variables. *Journal of Statistical Planning and Inference*, **141**(1), 172 – 188. ISSN 0378-3758. doi: 10.1016/j.jspi.2010.05.031.

Gallo, P. P. (1982a). Consistency of regression estimates when some variables are subject to error. *Comm. Statist. B-Theory Methods*, pages 973–893.

Gallo, P. P. (1982b). *Properties of Estimators in Errors-in-Variables Models*. PhD thesis, University of North Carolina, Chapel Hill, NC.

Gleser, L. J. (1981). Estimation in a multivariate "Errors in Variables" regression model: Large sample results. *Ann. Statist.*, **9**(1), 24–44. doi: 10.1214/aos/1176345330.

Golub, G. a Van Loan, C. (2013). *Matrix Computations*. 4. vydání. Johns Hopkins University Press. ISBN 9781421407944.

Golub, G., Hoffman, A. a Stewart, G. (1987). A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its Applications*, **88-89**, 317 – 327. ISSN 0024-3795. doi: 10.1016/0024-3795(87)90114-5.

Healy, J. D. (1975). Estimation and Tests for Unknown Linear Restriction in Multivariate Linear Models.

Horn, R. A. a Johnson, C. R. (2012). *Matrix Analysis*. 2. vydání. Cambridge University Press. ISBN 978-0521548236.

Jones, T. A. (1979). Fitting straight lines when both variables are subject to error. *Journal of the International Association for Mathematical Geology*, **11** (1), 1–25. ISSN 1573-8868. doi: 10.1007/BF01043243.

Lachout, P. (2018). Optimization theory - direct approach. Skripta k přednášce Teorie optimalizace. URL `http://www.karlin.mff.cuni.cz/~lachout/Vyuka/T-Optima/180103-TO-OptI-text.pdf`. Navštíveno 27.8.2018.

Lagarias, J. C. (2013). Euler's constant: Euler's work and modern developments. *Bulletin of the American Mathematical Society*, **50**(4), 527–628. doi: 10.1090/S0273-0979-2013-01423-X.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, **54**(1), 173–205. ISSN 0162-1459. doi: 10.2307/2282145.

Malinvaud, E. (1980). *Statistical Methods of Econometrics*. Handbook of Statistics. North-Holland Publishing Company. ISBN 9780444854735.

Markovsky, I. a Huffel, S. V. (2007). Overview of total least-squares methods. *Signal Processing*, **87**(10), 2283 – 2302. ISSN 0165-1684. doi: 10.1016/j.sigpro.2007.04.004.

Moran, P. (1971). Estimating structural and functional relationships. *Journal of Multivariate Analysis*, **1**(2), 232 – 255. ISSN 0047-259X. doi: 10.1016/0047-259X(71)90013-3.

Paige, C. C. a Wei, M. (1993). Analysis of the generalized total least squares problem $AX \approx B$ when some columns of A are free of error. *Numerische Mathematik*, **65**(1), 177–202. ISSN 0945-3245. doi: 10.1007/BF01385747.

Pešta, M. (2010). *Modern Asymptotic Pespectives on Errors-in-Variables Modeling*. PhD thesis, Charles University in Prague.

Pešta, M. (2013). Total least squares and bootstrapping with applications in calibration. *Statistics*, **47**(5), 966–991. doi: 10.1080/02331888.2012.658806.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Riggs, D. S., Guarnieri, J. A. a Addelman, S. (1978). Fitting straight lines when both variables are subject to error. *Life Sciences*, **22**(13), 1305 – 1360. ISSN 0024-3205. doi: 10.1016/0024-3205(78)90098-X.

Rousseeuw, P. J. a Ruts, I. (1996). Bivariate location depth. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **45**(4), 516–526. doi: 10.2307/2986073.

Thompson, R. (1972). Principal submatrices IX: Interlacing inequalities for singular values of submatrices. *Linear Algebra and its Applications*, **5**(1), 1 – 12. ISSN 0024-3795. doi: 10.1016/0024-3795(72)90013-4.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press. ISBN 9780521784504.

Van Huffel, S. a Vandewalle, J. (1991). *The Total Least Squares Problem*. Society for Industrial and Applied Mathematics (SIAM). ISBN 0-89871-275-0. doi: 10.1137/1.9781611971002.

Watson, G. (1998). Choice of norms for data fitting and function approximation. *Acta Numerica*, **7**, 337–377. doi: 10.1017/S0962492900002853.

Wooldridge, J. (2008). *Introductory Econometrics: A Modern Approach (with Economic Applications, Data Sets, Student Solutions Manual Printed Access Card)*. 4. vydání. South-Western College Pub. ISBN 0324581629.

Zuo, Y. a Sering, R. (2000). General notions of statistical depth function. *Annals of Statistics*, **28**(2), 461–482. doi: 10.1214/aos/1016218226.