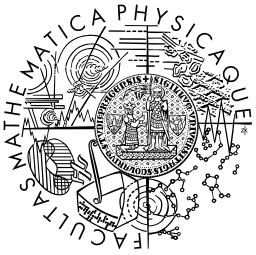


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Denisa Dočekalová

Regresní hloubka a podobné metody

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Stanislav Nagy, Ph.D.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2022

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

Tímto bych chtěla především poděkovat mému vedoucímu práce, Mgr. Stanislavu Nagymu, Ph.D. za vedení práce, za čas, který mi věnoval a za jeho nekonečnou trpělivost. Kromě toho bych chtěla také dále poděkovat mým rodičům za jejich dlouholetou podporu, a to jak ve studijním, tak i v běžném životě.

Název práce: Regresní hloubka a podobné metody

Autor: Denisa Dočekalová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Stanislav Nagy, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Zatímco poloprostorový medián jakožto robustní odhad střední hodnoty získává v posledních letech čím dál více na popularitě, regresní hloubka se i přes to, že je založena na podobném konceptu, stále řadí mezi relativně neznámé metody. Hlavním cílem této práce bylo tak především čtenáři přiblížit koncept robustní hloubky, ilustrovat její geometrickou interpretaci, a poskytnout alespoň základní přehled poznatků, ke kterým v rámci jednotlivých výzkumů došlo. Na závěr byla pak provedena malá simulační studie, která srovnává odhad metodou regresní hloubky s vybranými, v praxi běžně používanými odhady, a to konkrétně s odhadem metodou nejmenších absolutních odchylek a s odhadem metodou nejmenších čtverců.

Klíčová slova: regresní hloubka, hloubka, robustní regrese

Title: Regression depth and related methods

Author: Denisa Dočekalová

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Stanislav Nagy, Ph.D., Department of Probability and Mathematical Statistics

Abstract: While the halfspace depth has gained more and more popularity in the recent years as a robust estimator of the mean, regression depth, despite being based on a similar concept, is still a relatively unknown method. The main goal of this paper was therefore to introduce the concept of robust depth to the reader, illustrate its geometric interpretation, and provide at least a basic overview of the findings that occurred within the individual researches. Finally, a small simulation study was conducted comparing the deepest regression method with other selected methods commonly used in practice, namely the method of least absolute deviations and ordinary least squares method.

Keywords: regression depth, depth, robust regression

Obsah

Úvod

Jak jednou řekl George Box: "Všechny modely jsou špatné, ale některé jsou užitečné." Toto rčení se ve statistice nese již několik desítek let, i přesto je však stále pravdivé. Statistické modely mají za úkol co nejvíce zjednodušovat popis reálného světa, za což ale v některých případech platíme vysokou daň v podobě silných předpokladů. Nevyřčená přání, že pouze malé odchýlení od těchto předpokladů bude mít za důsledek zároveň malé odchýlení od hledaných výsledků, se však bohužel často ukazují jako mylná (viz Tukey (1960)). Z tohoto důvodu se časem oddělilo odvětví tzv. *robustní statistiky*, které se zabývá převážně metodami se zeslabenými předpoklady. Odhady získané těmito metodami tak sice nedisponují těmi nejlepšími teoretickými vlastnostmi, zato ale lépe odrážejí chování reálného světa. Jedním z hlavních předpokladů u většiny metod je, aby všechna pozorování pocházela z rozdělení, jehož charakteristiku odhadujeme. V praxi se však ukazuje, že většina datových souborů obsahuje 1-10 % pozorování, která pochází z nějakého *jiného* rozdělení, a tato pozorování jsou tedy v nějakém směru *chybná*.

Pro lineární regresní modely existují dva základní přístupy, jak se s chybnými pozorováními můžeme zkusit vypořádat. První z nich je založen na detekci podezřelých pozorování, která pak dále zkoumáme, a pokud jsou uznána za chybná, z datového souboru je vyřadíme. Tento přístup se nazývá *regresní diagnostika* a dodnes je považován za jeden z prvních kroků k získání robustnějšího odhadu. Samotné používání tohoto přístupu však s sebou nese řadu problémů. V první řadě se nejedná o ryze matematický aparát na jaké jsme zvyklí z klasické statistiky. I když se totiž některá pozorování mohou na první pohled skutečně jevit jako chybná, většinou neexistuje způsob, kterým bychom mohli ověřit, zda tomu tak skutečně je. Velmi lehce se proto může stát, že kromě chybných vyřadíme z datového souboru i správná pozorování (anebo tam naopak nějaká chybná ponecháme), což může vést ke znehodnocení našeho modelu.

Kritéria založená na regresní diagnostice jsou sice odvozena pouze za platnosti předpokladů metody nejmenších čtverců, těší se však velké oblibě i při používání celé řady dalších metod. Jejich použití pro tyto metody však není nijak podloženo, a není tedy jasné, co přesně daná kritéria říkají. I z toho důvodu se postupem času čím dále více statistiků začalo klonit ke druhému přístupu, a tím je oslabení předpokladů používaných metod v tom smyslu, že přímo nějaké procento chybných pozorování ve výběru připustíme. To vedlo k vytvoření celé třídy nových metod, jejichž základy položil Huber v roce 1964 a nazval je *robustní regresní metody*. Jednou z takových novějších takových metod je pak i tzv. *metoda regresní hloubky*, která je hlavním tématem této práce.

Metoda regresní hloubky byla poprvé představena v článku *Regression Depth* (Rousseeuw a Hubert, 1999). Její zavedení bylo inspirováno konceptem Tukeyho poloprostorové hloubky, a to konkrétně pojmem *zanoření*. Hlavním cílem této práce je tak především čtenáři přiblížit koncept robustní hloubky, ilustrovat její geometrickou interpretaci, a poskytnout alespoň základní přehled poznatků, ke kterým v rámci jednotlivých výzkumů došlo. V první části si nejprve zavedeme základní pojmy, se kterými budeme pracovat, a dále si podrobněji rozebereme vlastnosti odhadů, které budeme později porovnávat (a to s důrazem na vlast-

nosti popisující *robustnost*, tj. chování odhadu v situaci, kdy jsou v datovém souboru obsažena chybná pozorování). Ve druhé části se pak zaměříme na jednotlivé metody, kde největší prostor bude věnovaný metodě regresní hloubky, jakožto stěžejnímu tématu této práce. Na závěr bude provedena malá simulační studie, která porovná výsledky vybraných metod při různých typech a stupních kontaminace datového souboru.

1. Lineární model

V celé této práci budeme předpokládat, že máme k dispozici náhodný výběr $\boldsymbol{Z}_1 = (Y_1, \boldsymbol{X}_1^\top)^\top, \ldots, \boldsymbol{Z}_n = (Y_n, \boldsymbol{X}_n^\top)^\top$ z rozdělení určeného generickým vektorem $\boldsymbol{Z} = (Y, \boldsymbol{X}^\top)^\top$ s distribuční funkcí $H \in \mathcal{H}$, kde $\mathcal{H}$ značí množinu všech distribučních funkcí, Y je jednorozměrná náhodná veličina a $\boldsymbol{X}$ je $(p-1)$-rozměrný náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_{p-1})^\top$; $p \in \mathbb{N}$.

Poznámka. Zkráceně budeme psát $\boldsymbol{Y} := (Y_1, \ldots, Y_n)^\top$, $\mathbb{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$, kde $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p-1})^\top$ a $\mathbb{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\top$.

Definice 1. *Náhodnou veličinu Y budeme nazývat **odezvou**, náhodný vektor $\boldsymbol{X}^j = (X_{1,j}, \ldots, X_{n,j})^\top$, kde $j = 1, \ldots, p-1$ **j-tým regresorem**, matici $\mathbb{X}$ **regresní maticí** a matici $\widetilde{\mathbb{X}} = (\mathbf{1}_n, \mathbb{X})$ **regresní maticí s absolutním členem**.*

Poznámka. Vektor $\mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ se někdy také nazývá *absolutní člen* a často bývá automaticky zaveden jako 0-tý regresor. Kvůli zachování standardního značení u regresní hloubky v kapitole 5 však nebudeme tento způsob používat, i když koeficient příslušný absolutního členu v lineárním modelu budeme vždy předpokládat.

Předpoklad 1. ***Předpoklady lineárního regresního modelu*** *budeme rozumět následující podmínky:*

(1) $n > p$,

(2) $rank(\widetilde{\mathbb{X}}) = p$,

(3) $\boldsymbol{Z}_1 = (Y_1, \boldsymbol{X}_1^\top)^\top, \ldots, \boldsymbol{Z}_n = (Y_n, \boldsymbol{X}_n^\top)^\top$ *jsou nezávislé, stejně rozdělené náhodné vektory (zkr. i.i.d.),*

(4) rozdělení všech regresorů je ordinální a

(5) podmíněné rozdělení $Y|\boldsymbol{X}$ je absolutně spojité.

Každá z výše uvedených podmínek má svoji specifickou roli:

- Podmínky *(1)* a *(2)* jsou nutné podmínky k tomu, aby odhad hledané charakteristiky existoval a byl určen jednoznačně.
- Podmínka *(3)* nám říká, že všechna pozorování jsou nezávislá a pocházejí ze stejného rozdělení. Většinou se jedná o konkrétní typ rozdělení, jejíž charakteristiku chceme dále zkoumat, resp. odhadnout.
- Podmínka *(4)* nám říká, že napozorované hodnoty každého z regresorů je možné uspořádat. (Jedná se o nutnou podmínkou pro smysluplnou definici regresní hloubky, viz kapitola 5).
- Podmínku *(5)* není nezbytně nutné uvažovat, takové zobecnění však již značně přesahuje rozsah této práce.

Regresní modely se sestavují za jedním ze dvou účelů, buďto na základě hodnot regresorů chceme predikovat hodnoty odezvy, anebo je naším cílem přímo určit a popsat vztah mezi nimi. Pokud bychom pozorování nepovažovali za náhodné veličiny, ale pouze za deterministické konstanty, odpovídal by druhý úkol nalezení takového zobrazení $g\colon \mathbb{R}^{n\times(p-1)} \longrightarrow \mathbb{R}^n$, pro které by platilo $\boldsymbol{Y} = g(\mathbb{X})$. Zde ale vyvstávají dva zásadní problémy. Prvním problémem je vůbec existence takového zobrazení, protože se v praxi může velmi lehce stát, že pro stejné hodnoty regresorů dostaneme různé hodnoty odezvy. Tento problém je však možné vyřešit právě přechodem k náhodným veličinám. Druhý problém je pak určení hledaného zobrazení, protože bez jakéhokoliv omezení může takových zobrazení existovat nekonečně mnoho. Tomu se ale lze také vyvarovat, a to například omezením se pouze na soustavy *lineárních* rovnic:

Definice 2. *Nechť jsou splněny podmínky z předpokladu 1,* ***lineárním (regresním) modelem*** *budeme rozumět vztah, kdy je pro všechna $i \in \{1, \ldots, n\}$ splněno*

$$Y_i = \beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}_X + \varepsilon_i, \tag{1.1}$$

kde $\beta_0 \in \mathbb{R}$ je neznámý parametr, $\boldsymbol{\beta}_X = (\beta_1, \ldots, \beta_{p-1})^\top \in \mathbb{R}^{p-1}$ je neznámý vektorový parametr příslušný regresorům a $\varepsilon_i \in \mathbb{R}$ je náhodná veličina.

Poznámka. Parametry β_0 a $\boldsymbol{\beta}_X$ nejsou dle definice 2 určeny jednoznačně, jejich jednoznačnost zajistíme vždy pro konkrétní metodu později, a to pomocí přidání dalšího předpokladu (viz předpoklad 2 v sekci 3.1 a předpoklad 3 v sekci 4.1).

Poznámka I. Pokud nebude explicitně uvedeno jinak, budeme vždy předpokládat platnost lineárního modelu.

Poznámka II. Často budeme používat značení $\boldsymbol{\beta} := (\beta_0, \boldsymbol{\beta}_X^\top)^\top$. Ve vektorovém zápisu je pak možné soustavu (1.1) psát jako

$$\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \mathbb{X}\boldsymbol{\beta}_X + \boldsymbol{\varepsilon} = \widetilde{\mathbb{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.2}$$

kde $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $\widetilde{\boldsymbol{X}}_i^\top := (1, \boldsymbol{X}_i^\top)$ a $\widetilde{\mathbb{X}} = (\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_n)^\top$.

Poznámka III. Vektor $\boldsymbol{\beta}$ budeme nazývat *vektor koeficientů*, resp. β_0 bude *koeficient příslušný absolutnímu členu* a $\boldsymbol{\beta}_X$ *vektor koeficientů příslušný regresorům.*

Definice 3. *Náhodnou veličinu $\varepsilon_i = Y_i - \beta_0 - \boldsymbol{X}_i^\top \boldsymbol{\beta}_X$ z definice 2 budeme nazývat* ***i-tým chybovým členem*** *a vektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$* ***chybovým vektorem.***

Předpokládejme nyní, že již máme *nějaký* odhad vektoru koeficientů $\boldsymbol{\beta}$ k dispozici a uvažujme následující definice:

Definice 4. *Nechť $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, (\hat{\boldsymbol{\beta}}_X)^\top)^\top \in \mathbb{R}^p$ je nějaký odhad vektoru koeficientů $\boldsymbol{\beta}$, a předpokládejme, že existuje zobrazení $\boldsymbol{T} = (T_0, \ldots, T_{p-1})^\top\colon (\mathbb{R}^n, \mathbb{R}^{n\times p}) \longrightarrow \mathbb{R}^p$ takové, že platí $\hat{\boldsymbol{\beta}} := \boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}})$. Pak pro všechna $i \in \{1, \ldots, n\}$ budeme odhad*

$$\hat{Y}_i(\hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_X$$

nazývat ***i-tou predikovanou hodnotou*** *a*

$$U_i(\hat{\boldsymbol{\beta}}) = Y - \hat{Y}_i(\hat{\boldsymbol{\beta}})$$

nazývat ***i-tým reziduem.***

2. Vlastnosti odhadu

Volba vhodné metody záleží většinou na tom, s jakým typem dat pracujeme a jaké vlastnosti od daného odhadu požadujeme. Povaha problému a typ dat nás tedy v první řadě donutí omezit se pouze na určitou skupinu metod, které se hodí k řešení daného problému, a až z nich následně vybrat tu nejvhodnější podle toho, jaké vlastnosti od daného odhadu požadujeme.

Vlastnosti odhadů si rozdělíme do dvou hlavních skupin:

1. První skupinu budou představovat vlastnosti **teoretického charakteru**. Především se tedy bude jednat o ty vlastnosti, které popisují, *co*, a *jak dobře*, daný odhad vlastně odhaduje. Mezi tyto vlastnosti budeme řadit *nestrannost, rozptyl/eficienci, konzistenci,* a případně (pokud je známé) *asymptotické rozdělení* odhadu.

2. Druhou skupinu vlastností pak budeme vnímat jako spíše **praktického charakteru**, a budeme je dále dělit do dvou samostatných skupin:

   - První skupinu budou tvořit přípustné *transformace pozorování*, tj. jaké operace můžeme s daty provádět (změna měřítka, centrování dat, atd.), aniž by se náš odhad nějak zásadně změnil.
   - Druhou, a pro nás důležitější, budou tvořit vlastnosti popisující tzv. *robustnost* odhadu. Tyto vlastnosti popisují určitou *stabilitu* odhadu v situaci, kdy pracujeme s kontaminovaným datovým souborem, tj. v datovém souboru jsou přítomna nějaká chybná pozorování.

Definice vlastností jako nestrannost, rozptyl/eficience, konzistence a asymptotické rozdělení jsou dobře známé z klasické statistiky, a nebudeme si je zde proto podrobněji uvádět (pro podrobné definice viz např. Anděl (2007)). Přejdeme proto nyní přímo k definicím praktických vlastností.

2.1 Transformace pozorování

Stejně jako v definici 4 budeme i nyní předpokládat, že již máme *nějaký* odhad vektoru parametrů $\boldsymbol{\beta}$ k dispozici, a že existuje zobrazení $\boldsymbol{T}\colon (\mathbb{R}^n, \mathbb{R}^{n\times p}) \longrightarrow \mathbb{R}^p$ takové, že platí $\hat{\boldsymbol{\beta}} := \boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}}) = \boldsymbol{T}(\boldsymbol{Y}, (\mathbf{1}_n, \mathbb{X}))$. Následující definice jsou převzaté z knihy Rousseeuw a Leroy (1987), str. 116:

Definice 5. *Řekneme, že odhad $\hat{\boldsymbol{\beta}}$ je **regresně invariantní**, pokud platí*

$$T(\boldsymbol{Y} + \widetilde{\mathbb{X}}\boldsymbol{h}, \widetilde{\mathbb{X}}) = T(\boldsymbol{Y}, \widetilde{\mathbb{X}}) + \boldsymbol{h},$$

kde $\boldsymbol{h} \in \mathbb{R}^p$ je libovolný p-rozměrný vektor.

Definice 6. *Řekneme, že odhad $\hat{\boldsymbol{\beta}}$ je **škálově invariantní**, pokud platí*

$$T(c\boldsymbol{Y}, \widetilde{\mathbb{X}}) = cT(\boldsymbol{Y}, \widetilde{\mathbb{X}}),$$

kde $c \in \mathbb{R}, c \neq 0$ je libovolná reálná konstanta.

Definice 7. *Řekneme, že odhad* $\hat{\boldsymbol{\beta}}$ *je* ***afinně invariantní****, pokud platí*

$$T(\boldsymbol{Y}, \widetilde{\mathbb{X}}A) = A^{-1}\,T(\boldsymbol{Y}, \widetilde{\mathbb{X}})$$

kde $A \in \mathbb{R}^{p\times p}$ *je libovolná regulární matice.*

Pozorování. Pokud je odhad afinně invariantní, pak libovolná afinní transformace regresorů vektor predikovaných hodnot neovlivní:

$$\hat{\boldsymbol{Y}} = \widetilde{\mathbb{X}}\hat{\boldsymbol{\beta}} = \widetilde{\mathbb{X}}AA^{-1}\,T(\boldsymbol{Y}, \widetilde{\mathbb{X}}) = \widetilde{\mathbb{X}}A\,T(\boldsymbol{Y}, \widetilde{\mathbb{X}}A) = \widetilde{\mathbb{X}}^*\,T(\boldsymbol{Y}, \widetilde{\mathbb{X}}^*),$$

kde $\widetilde{\mathbb{X}}^* := \widetilde{\mathbb{X}}A$.

Poznámka. Výše uvedené vlastnosti se oproti vlastnostem jako nestrannost či konzistence nemusí na první pohled zdát jako příliš důležité, přesto však někdy můžou hrát zásadní roli. Například regresní invariance bývá častý předpoklad pro množství důkazů, které implicitně předpokládají možnost *posunu* hledaného parametru do počátku (tj. předpoklad BÚNO $\boldsymbol{\beta} := \mathbf{0}_p$). Škálová či afinní invariance odhadu se pak ukazují jako užitečné vlastnosti do praxe, kde např. afinní invariance odhadu může výrazně zredukovat množství potřebných výpočtů v situacích, kdy konstruujeme více modelů, které si liší pouze v afinní transformaci.

2.2 Robustnost

Předpokládejme nyní, že jsme v situaci, kdy je datový soubor kontaminovaný, tj. obsahuje určité procento pozorování, která jsou v určitém směru chybná. Tato pozorování se mohou v datovém souboru vyskytovat z několika různých důvodů:

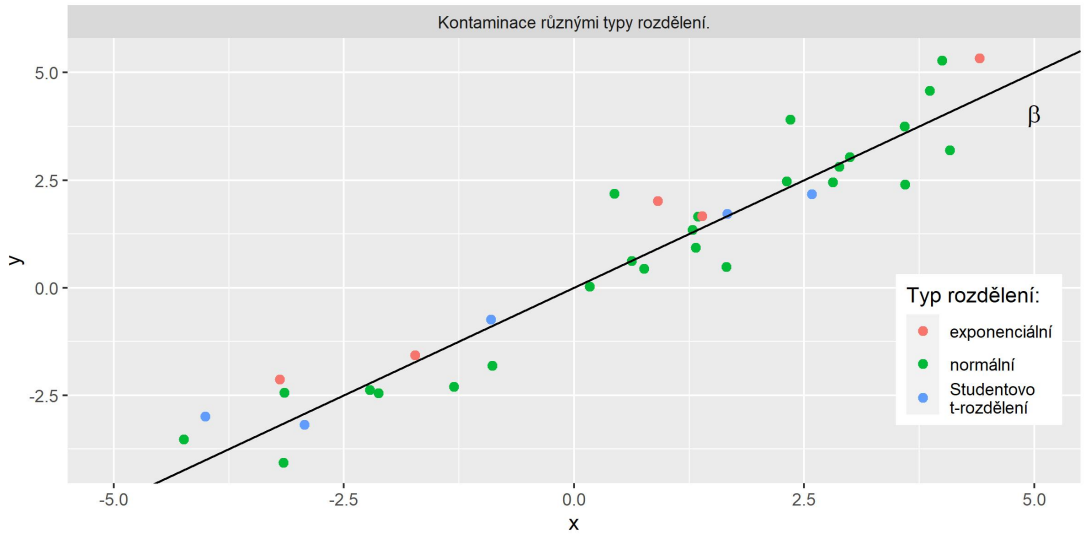
- byla do datového souboru zahrnuta omylem,
- byla u nich špatně umístěná desetinná čárka,
- došlo k chybě při jejich přepisu,
- došlo k chybě při jejich měření v důsledku nečekaných vlivů, atd.

Robustnost v takovém případě představuje soubor vlastností, které popisují, jak moc se může odhad změnit (resp. vychýlit) právě v případě, kdy jsou v datovém souboru taková chybná pozorování obsažena. Kontaminace výběru chybnými pozorováními totiž znamená, že je porušen předpoklad *i.i.d.* z předpokladů lineárního modelu, a není tedy na první pohled jasné, jak se odhad v takovém případě zachová.

Jak již bylo zmíněno v úvodu, pokud jsou v datovém souboru chybná pozorování přítomna, máme dvě možnosti, jak se s nimi zkusit vypořádat:

- V první řadě se můžeme pokusit tato pozorování detekovat. Jak se však ukazuje, to může být v některých případech náročné, někdy i takřka neproveditelné (viz obrázek 2.1). Pokud se nám však chybná pozorování nepodaří odhalit, tak v lepším případě hrozí, že budeme odhadovat charakteristiku nikoliv správného, ale kontaminovaného rozdělení. V takovém případě však narážíme na první problém, a tím je interpretace odhadu. V horším případě se pak může stát, že díky porušení předpokladů metody nebudeme ani schopni určit, co vlastně odhadujeme.

- Druhá možnost je připustit, že budeme přímo odhadovat charakteristiku kontaminovaného rozdělení. V takovém případě pokud zajistíme, aby se charakteristiky založené na správném a kontaminovaném rozdělení od sebe *příliš* nelišily, pak tento přístup nabízí východisko pro získání *smysluplného* odhadu i pro kontaminované datové soubory, a to bez nutnosti detekce chybných pozorování.



Obrázek 2.1: Kontaminace datového souboru různými typy rozdělení. Na obrázku vidíme směs pozorování pocházející ze 3 různých typů rozdělení, kde pozorování z normálního rozdělení představují správná pozorování, a pozorování pocházející z exponenciálního/Studentova t-rozdělení dvě různá rozdělení chybných pozorování. Protože se však nosiče všech rozdělení silně překrývají, nejsme schopni chybná pozorování odhalit.

Co znamená *nelišit se příliš* není v tuto chvíli jasné, protože neklademe žádné podmínky na hodnoty chybných pozorování. Pokud však předem vyloučíme extrémní situace, kdy samotná konstrukce odhadu nedává smysl (např. když je počet chybných pozorování větší než počet správných), nabízí se uvažovat ty odhady, které nám o hledaném parametru *vždy* poskytnou *alespoň nějakou* informaci.

Definice 8 (Neformální)**.** *Řekneme, že* $\hat{\boldsymbol{\beta}}$ *je* ***smysluplný odhad*** *vektorového parametru* $\boldsymbol{\beta}$*, pokud existuje konstant* $0 < k < \infty$ *taková, že při konstrukci odhadu* $\hat{\boldsymbol{\beta}}$ *na základě kontaminovaného výběru platí*

$$|| E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})||_2 < k, \tag{2.1}$$

kde $|| \cdot ||_2$ *značí euklidovskou normu.*

Když máme tedy nyní alespoň nějakou představu, které odhady bychom mohli považovat za smysluplné, nabízí se hledat odpovědi na následující otázky:

1. Jaké maximální procento chybných pozorování může být ve výběru obsaženo, abychom i tak stále dostali smysluplný odhad?

2. Jsme schopni nějakým způsobem popsat vztah mezi odhadem a hodnotou chybného pozorování, tj. jak moc se bude konkrétní odhad měnit v závislosti na jeho hodnotách?

Odpovědi na předchozí otázky nám poskytuje tzv. *infinitezimální* přístup, do kterého spadají pojmy jako *breakdown-point* či *influenční funkce*. Obecně není možné říci, která z těchto charakteristik je k popisu robustnosti důležitější, každá z nich totiž popisuje chování odhadu v jiném smyslu. Zatímco *breakdown-point* lze považovat za *globální* charakteristiku, *influenční funkce* je charakteristika *lokální*. Vždy je tedy potřeba zaměřit se na obě z nich.

2.2.1 Breakdown-point

„Jaké maximální procento chybných pozorování může být ve výběru obsaženo, abychom i tak stále dostali smysluplný odhad?“

Definice 9. *Označme si $\boldsymbol{Y}_m$, $\mathbb{X}_m$ náhodný vektor/matici $\boldsymbol{Y}$, $\mathbb{X}$, kde došlo k nahrazení m-pozorování libovolnými hodnotami.* ***Breakdown-pointem*** *(zkráceně pouze BP) odhadu $\boldsymbol{T}$ při náhodném výběru $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ pak rozumíme statistiku*

$$\varepsilon_n^* := \varepsilon_n^*(\boldsymbol{T}, \mathbb{Z}) = \varepsilon_n^*(\boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) = \min_{m \in \mathbb{N}} \left\{ \frac{m}{n};\ \Delta(m, \boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) = \infty \right\}, \tag{2.2}$$

kde

$$\Delta(m, \boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) := \sup_{\boldsymbol{Y}_m, \mathbb{X}_m} \left|\left| \boldsymbol{T}(\boldsymbol{Y}_m, \widetilde{\mathbb{X}}_m) - \boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}}) \right|\right|_2. \tag{2.3}$$

Pozorování I. Pokud bychom měli k dispozici nestranný odhad parametru $\boldsymbol{\beta}$, pak hodnota funkce Δ odhaduje vychýlení odhadu $\hat{\boldsymbol{\beta}}$ při nejhorším možném scénáři.

Pozorování II. Rozdělení pozorování na odezvu a regresory může být pro zápis rovnice (2.2) velice výhodné, v takovém případě je totiž možné při zafixování druhé složky zkoumat robustnost odhadu jak vzhledem k odezvě, tak vzhledem k regresorům samostatně. Velké množství odhadů tak sice není robustních, ale jsou robustní alespoň vzhledem k odezvě.

Poznámka I. Definice 9 je převzata od Donoho a Huber (1983), někdy se však můžeme také setkat i s mírně odlišnými definicemi BP. V první řadě může dojít k malé úpravě v rovnici (2.2), a to snížením hodnoty BP o $1/n$ (viz Hampel (1997), str. 98). V takovém případě pak hodnota BP představuje maximální přípustné poměrné zastoupení chybných pozorování v datovém souboru, při kterých ještě dostáváme smysluplný odhad, tj.

$$\varepsilon_n^{**} := \varepsilon_n^{**}(\boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) = \max_{m \in \mathbb{N}} \left\{ \frac{m}{n};\ \Delta(m, \boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) < \infty \right\}. \tag{2.4}$$

Další obměna může spočívat ve způsobu zařazení chybných pozorování. Označme si $\boldsymbol{Y}_{n+m}$, $\mathbb{X}_{n+m}$ náhodný vektor/matici po *přidání* m chybných pozorování k již existujícím pozorováním ve výběru. V takovém případě BP definujeme jako

$$\varepsilon_n^* := \varepsilon_n^*(\boldsymbol{T}, \mathbb{Z}) = \varepsilon_n^*(\boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) = \min_{m \in \mathbb{N}} \left\{ \frac{m}{n+m};\ \Delta(m, \boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) = \infty \right\}, \tag{2.5}$$

kde

$$\Delta(m, \boldsymbol{T}, \boldsymbol{Y}, \mathbb{X}) := \sup_{\boldsymbol{Y}_{n+m}, \mathbb{X}_{n+m}} \left|\left|\boldsymbol{T}(\boldsymbol{Y}_{n+m}, \widetilde{\mathbb{X}}_{n+m}) - \boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}})\right|\right|_2. \tag{2.6}$$

Mezi definicí pomocí nahrazení a pomocí přidání existují přímé vztahy (viz Zuo (2001)), a výběr konkrétní z nich se pak často odvíjí od toho, se kterou definicí se autorovi lépe pracuje.

Poznámka II. V neposlední řadě je dobré si povšimnout, že BP dle definice 9 je založen na náhodném výběru, a proto se také někdy nazývá *BP pro konečný rozsah výběru*. Existuje však i jeho obecnější definice, která je založena na distribuční funkci a Lévy-Prochorovově metrice (viz Prohorov (1956)). Tato definice je však výrazně komplikovanější (viz Hampel (1997), str. 97), a protože v rozumných případech platí

$$\varepsilon_n^* \xrightarrow{s.j.} \varepsilon^*,$$

kde ε^* je hodnota BP z obecnější definice, rozhodli jsme se zde tuto definici neuvádět.

Naše definice BP tedy představuje minimální poměrné zastoupení chybných pozorování v náhodném výběru, při kterém může dojít ke změně odhadu koeficientů regresní přímky libovolným způsobem. Čím vyšší je jeho hodnota, tím je metoda vůči chybným pozorováním odolnější. Minimální možná hodnotě BP je $1/n$, maximální pak $1/2$. Bohužel, pro množství běžně používaných metod platí, že $BP = 1/n$ (viz Rousseeuw a Leroy (1987)), a tak jediné chybné pozorování může způsobit, že se získaný odhad stane zcela bezcenným.

Pozorování. Jak je vidět z předchozí části, terminologie v oblasti BP je značně nejednoznačná, a je tedy dobré si vždy předem ujasnit, s jakou definicí autor zrovna pracuje.

2.2.2 Influenční funkce

„*Jsme schopni nějakým způsobem popsat vztah mezi odhadem a hodnotou chybného pozorování, tj. jak moc se bude konkrétní odhad měnit v závislosti na jeho hodnotách?*“

Definice vznikla kombinací definic z knihy Hampel (1997), str. 84 a 226.

Definice 10. *Nechť H je distribuční funkce náhodného vektoru $\boldsymbol{Z} = (Y, \boldsymbol{X}^\top)^\top$ a předpokládejme, že existuje funkcionál $\widetilde{\boldsymbol{T}}(H) = (\widetilde{T}_0(H), \ldots, \widetilde{T}_{p-1}(H))^\top$ takový, že platí $\boldsymbol{\beta} = \widetilde{\boldsymbol{T}}(H)$. **Influenční funkcí** funkcionálu $\widetilde{\boldsymbol{T}}$ pak budeme rozumět zobrazení*

$$\mathsf{IF}(\boldsymbol{z}) := \mathsf{IF}(\boldsymbol{z}, \widetilde{\boldsymbol{T}}, H) = (\mathsf{IF}(\boldsymbol{z}, \widetilde{T}_0, H), \ldots, \mathsf{IF}(\boldsymbol{z}, \widetilde{T}_{p-1}, H)), \tag{2.7}$$

v těch bodech $\boldsymbol{z} \in \mathbb{R}^p$, ve kterých pro všechna $j \in \{0, \ldots, p-1\}$ existuje limita

$$\mathsf{IF}(\boldsymbol{z}, \widetilde{T}_j, H) := \lim_{h \to 0^+} \frac{\widetilde{T}_j((1-h)\,H + h\delta_{\boldsymbol{z}}) - \widetilde{T}_j(H)}{h}, \tag{2.8}$$

kde $0 < h < 1$ a $\delta_{\boldsymbol{z}}$ označuje vícerozměrné Diracovo rozdělení v bodě $\boldsymbol{z} \in \mathbb{R}^p$.

Poznámka. Součet $(1-h)\,H + h\delta_{\boldsymbol{z}}$ označuje distribuční funkci příslušnou pravděpodobnostní míře $P_{\delta_{\boldsymbol{z}}}$, kde $P_{\delta_{\boldsymbol{z}}}$ je definována jako

$$P_{\delta_{\boldsymbol{z}}}(B) = (1-h)P_H(B) + hI(\boldsymbol{Z} \in B)$$

pro všechny borelovské množiny $B \subset \mathbb{R}^p$ a P_H značí pravděpodobnostní míru příslušnou distribuční funkci H.

Pozorování. Na rozdíl od BP (dle definice 9) představuje influenční funkce pouze teoretický koncept, a nemusí tak pro některé typy odhadů vůbec existovat.

Influenční funkce (původně nazývaná influenční křivka, viz Hampel (1974)) tak popisuje velikost změny odhadu při „nekonečně malé kontaminaci“ v bodě $\boldsymbol{z} \in \mathbb{R}^p$.

Vztah mezi influenční funkcí a asymptotickým rozdělením M-odhadů

Tvar influenční funkce lze kromě teorie robustních odhadů nalézt i v jiné oblasti statistiky, a to konkrétně ve vzorci asymptotického rozdělení některých typů odhadů. Díky tomu je pak možné odvodit tvar influenční funkce i jiným způsobem než přímo z definice 10. Pro nás bude důležitý především vztah pro tzv. M-odhady (mezi které patří jak metoda OLS, tak metoda LAD).

Definice 11. *Řekneme, že $\hat{\boldsymbol{\beta}}$ je **M-odhadem** vektorového parametru $\boldsymbol{\beta}$, pokud existuje funkce $\rho\colon \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ taková, že platí*

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}).$$

Pozorování. Pokud $\mathsf{E}\ \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})$ existuje a je konečná, pak ze silného zákona velkých čísel vyplývá, že

$$\frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}) \xrightarrow{s.j.} \mathsf{E}\ \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}),$$

a (za určitých podmínek) tedy i

$$\operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}) \xrightarrow{s.j.} \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathsf{E}\ \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}).$$

Poznámka. Často platí, že ρ je možné přímo zapsat jako funkci chybových členů ε_i. V takovém případě existuje funkce $\rho^*\colon \mathbb{R} \times \mathbb{R}^p \longrightarrow \mathbb{R}$ taková, že platí

$$\rho^*(\varepsilon_i) = \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}),$$

a proto se také někdy funkce ρ nazývá **ztrátová funkce**.

Poznámka II. Pokud budeme navíc dále předpokládat, že ρ má druhou derivaci podle $\boldsymbol{\beta}$, a střední hodnota této derivace existuje a je konečná, můžeme si zavézt značení

$$\psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}) := \frac{\partial \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \quad \Gamma(\boldsymbol{\beta}) := \mathsf{E}\ \frac{\partial \psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}.$$

Vektoru $\psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}) = (\frac{\partial \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})}{\partial \beta_0}, \ldots, \frac{\partial \rho(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})}{\partial \beta_{p-1}})^\top$ se pak říká **vektor skórů**.

Tvrzení 12. *Předpokládejme, že jsou splněny předpoklady regularity (Omelka (2020), str. 44) a* $\boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}})$ *z definice 11 je konzistentním odhadem vektorového parametru* $\boldsymbol{\beta} = \widetilde{\boldsymbol{T}}(H)$*, pak*

$$\sqrt{n}(\boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}}) - \widetilde{\boldsymbol{T}}(H)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathsf{IF}((Y_i, \boldsymbol{X}_i), \widetilde{\boldsymbol{T}},\ H), \tag{2.9}$$

kde

$$\mathsf{IF}((Y_i, \boldsymbol{X}_i), \widetilde{\boldsymbol{T}},\ H) = -\Gamma^{-1}(\boldsymbol{\beta})\psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta}), \tag{2.10}$$

a platí, že

$$\sqrt{n}(\boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}}) - \widetilde{\boldsymbol{T}}(H)) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}_p, \Gamma^{-1}(\boldsymbol{\beta})\, \Sigma(\boldsymbol{\beta})\, \Gamma^{-1}(\boldsymbol{\beta})\right),$$

kde

$$\Sigma(\boldsymbol{\beta}) := \mathsf{var}(\psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})) = \mathsf{E}\ \psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})\psi(Y_i, \widetilde{\boldsymbol{X}}_i, \boldsymbol{\beta})^\top.$$

Poznámka. Tvrzení 12 platí kromě M-odhadů např. také pro L-odhady nebo R-odhady (Rousseeuw a Leroy, 1987), a jak již bylo zmíněno, často se tak využívá k získání influenční funkce. Bohužel existují ale i takové typy odhadů, pro které tvrzení 12 neplatí (viz regresní hloubka, Van Aelst a Rousseeuw (2000)), a pro ty je pak nutné odvodit tvar jejich influenční funkce přímo z definice 10.

Pozorování. Důkaz tvrzení 12 lze nalézt např. ve skriptech Omelka (2020) s tím rozdílem, že influenční funkce je již přímo definována pouze pro M-odhady pomocí rovnice (2.10). Vztah mezi influenční funkcí dle definice 10 můžeme tedy alespoň zkusit nahlédnout z následující aproximace (Hampel, 1997):

Důkaz (náznak). Předpokládejme, že je možné zkonstruovaný odhad přepsat jako funkcionál do tvaru $\boldsymbol{T}(\boldsymbol{Y}, \widetilde{\mathbb{X}}) = \widetilde{\boldsymbol{T}}(\hat{H}_n)$, kde $\hat{H}_n$ je empirická distribuční funkce H. Z rovnice (2.8) pak pomocí von Misesova rozvoje[1] prvního řádu dostáváme aproximaci

$$\boldsymbol{T}(\boldsymbol{Y}, \mathbb{X}) = \widetilde{\boldsymbol{T}}(\hat{H}_n) \approx \widetilde{\boldsymbol{T}}(H) + \int_{\mathbb{R}^p} \mathsf{IF}\,(\boldsymbol{z}, \widetilde{\boldsymbol{T}}, H)\, d\, \hat{H}_n(\boldsymbol{z}),$$

a protože $\hat{H}_n$ je empirická distribuční funkce, je možné aproximaci přepsat do tvaru

$$\sqrt{n}(\widetilde{\boldsymbol{T}}(\hat{H}_n) - \widetilde{\boldsymbol{T}}(H)) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathsf{IF}\,(\boldsymbol{Z}_i, \widetilde{\boldsymbol{T}}, H),$$

což odpovídá vyjádření rovnice (2.9).

□

[1] Viz Taylorův rozvoj pro funkcionály, Hampel (1997).

3. Metoda OLS

Metoda nejmenších čtverců, zkráceně pouze OLS (z *a.j. ordinary least squares*) je jednou z nejstarších metod používaných k získání odhadu parametru $\boldsymbol{\beta}$ v lineárních modelech. Její použití se datuje až do roku 1805, a často bývá také chybně označována jako první použitá metoda vůbec (tou byla však metoda LAD). Zároveň není přehnané tvrdit, že se jedná i o metodu nejznámější a nejpopulárnější, a to z toho důvodu, že na rozdíl od jiných metod existuje její přesné analytické řešení.

Metoda OLS sice nepatří mezi robustní metody, přesto si ji zde však uvedeme, protože je běžně používána jako *benchmark* k porovnávání výsledků získaných pomocí metody regresní hloubky (a použijeme ji tedy pro porovnání s metodou regresní hloubky v simulační studii, viz kapitola 6). Pokud totiž pracujeme se správnými daty (tj. datový soubor neobsahuje chybná pozorování), odhad získaný metodou OLS disponuje těmi nejlepšími teoretickými vlastnostmi.

Definice 13. ***Odhadem vektoru koeficientů β metodou OLS*** *budeme rozumět odhad získaný na základě následujícího minimalizačního kritéria:*

$$\hat{\boldsymbol{\beta}}^{LS} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, SS(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_0 - \boldsymbol{X}_i^\top \boldsymbol{\beta}_X)^2, \tag{3.1}$$

kde funkce $SS(\boldsymbol{\beta}) : \mathbb{R}^p \longrightarrow \mathbb{R}_0^+$ se nazývá ***součet čtverců modelu****.*

Pozorování. Uvažujme definici 11, pak zřejmě odhad OLS je M-odhadem pokud jako ztrátovou funkci ρ zvolíme

$$\rho(y, \widetilde{\boldsymbol{x}}, \boldsymbol{\beta}) := (y - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\beta})^2 = (y - \beta_0 - \boldsymbol{x}^\top \boldsymbol{\beta}_X)^2.$$

Tvrzení 14. *Nechť jsou splněny podmínky z předpokladu 1 (tj. předpoklady lineárního modelu), rovnice (3.1) má pak* ***právě jedno*** *řešení, a lze jej psát ve tvaru*

$$\hat{\boldsymbol{\beta}}^{LS} = (\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}})^{-1} \widetilde{\mathbb{X}}^\top \boldsymbol{Y}. \tag{3.2}$$

Důkaz. Součet čtverců modelu z rovnice (3.1) lze zapsat ve vektorovém tvaru jako

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \widetilde{\mathbb{X}}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \widetilde{\mathbb{X}}\boldsymbol{\beta}).$$

Minimum pak můžeme nalézt pomocí parciálních derivací vzhledem k $\boldsymbol{\beta}$, které všechny položíme rovny 0. Že se skutečně jedná o minimum vyplývá z druhých parciálních derivací, protože matice $\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}}$ je pozitivně definitní. Z předpokladu o plné hodnosti matice $\widetilde{\mathbb{X}}$ nakonec víme, že k matici $\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}}$ existuje matice inverzní a řešení je tedy možné vyjádřit ve tvaru z rovnice (3.2):

$$\begin{aligned} \frac{\partial SS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \mathbf{0}_p \\ -2\widetilde{\mathbb{X}}^\top (\boldsymbol{Y} - \widetilde{\mathbb{X}}\boldsymbol{\beta}) &= \mathbf{0}_p \\ \widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}} \boldsymbol{\beta} &= \widetilde{\mathbb{X}}^\top \boldsymbol{Y}. \end{aligned}$$

□

3.1 Teoretické vlastnosti odhadu OLS

Předpoklad 2. *Řekneme, že je splněn* ***předpoklad linearity ve střední hodnotě****, pokud platí rovnost*

$$\mathsf{E}(Y \mid \widetilde{\boldsymbol{X}}) = \widetilde{\boldsymbol{X}}^\top \boldsymbol{\beta}^{MEAN}, \tag{3.3}$$

kde $\boldsymbol{\beta}^{MEAN} = (\beta_0^{MEAN}, (\boldsymbol{\beta}_X^{MEAN})^\top)^\top$.

Poznámka. Aby bylo v budoucnu jasné, o jakém parametru zájmu $\boldsymbol{\beta}$ mluvíme (tj. jestli předpokládáme linearitu ve střední hodnotě nebo v mediánu), budeme parametr zájmu odlišovat horním indexem dle daného předpokladu.

Důkaz následujícího tvrzení lze nalézt například ve skriptech Komárek (2019):

Tvrzení 15. *Nechť je splněn předpoklad linearity ve střední hodnotě z definice 2, pak odhad* $\hat{\boldsymbol{\beta}}^{LS}$ *je* ***nestranným*** *odhadem vektorového parametru zájmu* $\boldsymbol{\beta}^{MEAN}$.

Tvrzení 16. *Nechť je splněn předpoklad linearity ve střední hodnotě, jsou splněny předpoklady regularity pro M-odhady (Omelka, 2020) a* $\hat{\boldsymbol{\beta}}^{LS}$ *je* ***konzistentní*** *odhad parametr zájmu* $\boldsymbol{\beta}^{MEAN}$*, pak platí, že*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{LS} - \boldsymbol{\beta}^{MEAN}) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}_p, \Gamma\Sigma\Gamma\right), \tag{3.4}$$

kde

- $\Gamma := \Gamma(\boldsymbol{\beta}^{MEAN}) = \left\{2\,\mathsf{E}(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top)\right\}^{-1}$,
- $\Sigma := \Sigma(\boldsymbol{\beta}^{MEAN}) = 4\,\mathsf{E}(\sigma^2(\widetilde{\boldsymbol{X}})\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top)$, *kde* $\sigma^2(\widetilde{\boldsymbol{X}}) = \mathsf{var}(Y|\widetilde{\boldsymbol{X}})$.

Důkaz. Z pozorování za definicí 13 víme, že se jedná o M-odhad, k důkazu tedy využijeme tvrzení 12. Vyjádření vektoru skórů a matice Γ získáme přímo z parciálních derivací:

$$\psi(y, \widetilde{\boldsymbol{x}}, \boldsymbol{\beta}) = \frac{\partial\rho(y, \widetilde{\boldsymbol{x}}, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2(y - \widetilde{\boldsymbol{x}}^\top\boldsymbol{\beta})\widetilde{\boldsymbol{x}},\ \Gamma(\boldsymbol{\beta}) = \mathsf{E}\ \frac{\partial\psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top} = \mathsf{E}\ 2\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top.$$

Z předpokladu linearity ve střední hodnotě vyplývá, že $\mathsf{E}\ \psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta}^{MEAN}) = \mathbf{0}_p$, a tedy

$$\begin{aligned}\mathsf{var}\left(\psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta}^{MEAN})\right) &= 4\,\mathsf{var}\left((Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}^{MEAN})\widetilde{\boldsymbol{X}}^\top\right) \\ &= 4\,\mathsf{E}\left(\widetilde{\boldsymbol{X}}(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}^{MEAN})^2\widetilde{\boldsymbol{X}}^\top\right).\end{aligned}$$

Matici Σ v rovnici (3.4) získáme po dosazení $\boldsymbol{\beta}^{MEAN}$ do vyjádření $\Sigma(\boldsymbol{\beta})$, které dostaneme na základě podmínění náhodným vektorem $\widetilde{\boldsymbol{X}}$ z rovnosti

$$\mathsf{E}\left(\mathsf{E}\left(\widetilde{\boldsymbol{X}}(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta})^\top(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta})\widetilde{\boldsymbol{X}}^\top|\widetilde{\boldsymbol{X}}\right)\right) = \mathsf{E}\left(\widetilde{\boldsymbol{X}}\,\mathsf{E}\left((Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta})^2|\widetilde{\boldsymbol{X}}\right)\widetilde{\boldsymbol{X}}^\top\right),$$

a z označení $\sigma^2(\widetilde{\boldsymbol{X}}) := \mathsf{var}\left(Y|\widetilde{\boldsymbol{X}}\right) = \mathsf{E}\left((Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}^{MEAN})^2|\widetilde{\boldsymbol{X}}\right)$.

□

Poznámka. Pokud bychom dále předpokládali, že $\sigma^2(\widetilde{\boldsymbol{X}}) =: \sigma^2 > 0$ je neznámý parametr (tj. předpokládáme *homoskedasticitu*), a že podmíněné rozdělení $Y|\boldsymbol{X}$ je normální, tj.

$$Y|\widetilde{\boldsymbol{X}} \sim \mathcal{N}\left(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\beta}^{MEAN}, \sigma^2\right),$$

pak je možné ukázat, že odhad $\hat{\boldsymbol{\beta}}^{LS}$ je dokonce ***eficientním*** odhadem parametru zájmu $\boldsymbol{\beta}^{MEAN}$ (důkaz vychází z ekvivalence odhadu získaného na základě rovnice (3.2) a odhadu získaného pomocí metody maximální věrohodnosti, viz Komárek (2019)). Při splnění těchto podmínek tedy neexistuje odhad parametru zájmu $\boldsymbol{\beta}^{MEAN}$ s lepšími teoretickými vlastnostmi, díky čemuž se i metoda OLS používá jako již zmíněný *benchmark*.

3.2 Praktické vlastnosti odhadu OLS

Škálová a regresní invariance odhadu OLS vyplývá přímo po dosazení transformovaných veličin do rovnice (3.2), afinní invarianci pak dostaneme po malé úpravě:

$$((\tilde{\mathbb{X}}A)^\top \tilde{\mathbb{X}}A)^{-1}(\tilde{\mathbb{X}}A)^\top \boldsymbol{Y} = (A^\top \tilde{\mathbb{X}}^\top \tilde{\mathbb{X}}A)^{-1}(\tilde{\mathbb{X}}A)^\top \boldsymbol{Y} = A^{-1}(\tilde{\mathbb{X}}^\top \tilde{\mathbb{X}})^{-1}A^{-\top}A^\top \tilde{\mathbb{X}}^\top \boldsymbol{Y},$$

kde $A^{-\top} := (A^\top)^{-1}$.

Tvrzení 17. *Odhad OLS je regresně, škálově, i afinně invariantní.*

Jak už jsme poznamenali, odhad OLS má velmi dobré teoretické vlastnosti, není však robustní, a to ani vzhledem k odezvě, ani vzhledem k regresorům (vyplývá přímo z tvaru odhadu z rovnice (3.2)). Tvar influenční funkce pak vyplývá přímo z tvrzení 12 a z vyjádření vektoru ψ a matice Γ z tvrzení 16.

Tvrzení 18. ***Influenční funkci pro odhad získaný metodou OLS*** *je možné psát ve tvaru*

$$\mathit{IF}((y, \boldsymbol{x}), \hat{\boldsymbol{\beta}}^{LS}, H) = (E\ \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top)^{-1}(y - \beta_0^{MEAN} - \boldsymbol{x}^\top \boldsymbol{\beta}_X^{MEAN})(1, \boldsymbol{x}^\top)^\top.$$

Pozorování. Z regresní invariance OLS odhadu můžeme BÚNO předpokládat, že $\boldsymbol{\beta}^{MEAN} := \mathbf{0}_p$, v takovém případě se influenční funkce redukuje do tvaru

$$\mathsf{IF}\left((y, \boldsymbol{x}), \hat{\boldsymbol{\beta}}^{LS}, H\right) = (\mathsf{E}\ \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top)^{-1} y\,(1, \boldsymbol{x}^\top)^\top,$$

a zřejmě tedy není omezená ani vzhledem k odezvě, ani vzhledem k regresorům.

Pokud jsou tedy splněny potřebné předpoklady, disponuje odhad OLS nejlepšími teoretickými vlastnostmi. Pokud je však datový soubor kontaminovaný, pak může odhad OLS v závislosti na typu kontaminace dávat velmi odlišné výsledky (viz kapitola 6).

4. Metoda LAD

Metoda nejmenších absolutních odchylek, zkráceně pouze LAD (z *a.j. least absolute deviation*) byla vůbec první použitou metodou k získání odhadu parametru $\boldsymbol{\beta}$ v lineárních modelech. Poprvé ji představil fyzik Roger Joseph Boscovich v roce 1757, což bylo cca o 50 let dříve než byla publikována metoda OLS.

Metodu OLS jsme si vybrali k porovnání z důvodu nejlepších teoretických vlastností v případě nekontaminovaného datového souboru. Metoda regresní hloubky však na rozdíl od metody OLS nepředpokládá linearitu ve střední hodnotě, ale v mediánu (viz předpoklad 3), a pro asymetrická rozdělení tak odhaduje jiný parametr zájmu. Chceme-li tedy výsledky získané metodou regresní hloubky posoudit i pro jiná než symetrická rozdělení, bylo potřeba kromě metody OLS zvolit i další metodu, která by stejně jako metoda regresní hloubky předpokládala linearitu v mediánu, metodu LAD.

Definice 19. ***Odhadem vektorového parametru β metodou LAD*** *budeme rozumět odhad získaný na základě následujícího minimalizačního kritéria:*

$$\hat{\boldsymbol{\beta}}^{LAD} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left| Y_i - \beta_0 - \boldsymbol{X}_i^\top \boldsymbol{\beta}_X \right|. \tag{4.1}$$

Pozorování. Odhad LAD je opět zřejmě M-odhadem se ztrátovou funkcí

$$\rho(y, \widetilde{\boldsymbol{x}}, \boldsymbol{\beta}) := |y - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\beta}| = |y - \beta_0 - \boldsymbol{x}^\top \boldsymbol{\beta}_X|.$$

Poznámka. Na rozdíl od ztrátové funkce OLS nemá však ztrátová funkce LAD ve všech bodech derivaci, a není tak možné zkonstruovat přesné analytické řešení (proto se také v historii, když ještě nebyly počítače, přešlo na metodu OLS). Řešení optimalizační úlohy určené rovnicí (4.1) je však možné získat na základě simplexového algoritmu (viz např. Dupačová a Lachout (2011)).

4.1 Teoretické vlastnosti odhadu LAD

Předpoklad 3. *Řekneme, že je splněn* ***předpoklad linearity v mediánu****, pokud platí rovnost*

$$\mathsf{med}\,(Y \mid \widetilde{\boldsymbol{X}}) = \widetilde{\boldsymbol{X}}^\top \boldsymbol{\beta}^{MED}, \tag{4.2}$$

kde $\boldsymbol{\beta}^{MED} = (\beta_0^{MED}, (\boldsymbol{\beta}_X^{MED})^\top)^\top$ *a* $\mathsf{med}\,(\cdot)$ *značí medián podmíněného rozdělení.*

Poznámka. Jak již bylo zmíněno úvodem, metodu LAD jsme si vybrali primárně z toho důvodu, že odhaduje stejný parametr zájmu jako metoda regresní hloubky. Pokud je však podmíněné rozdělení $Y|\boldsymbol{X}$ symetrické s konečnou střední hodnotou, pak platí

$$\mathsf{E}\,(Y \mid \widetilde{\boldsymbol{X}}) = \mathsf{med}\,(Y \mid \widetilde{\boldsymbol{X}}),$$

a všechny uvažované metody tak odhadují stejný parametr, tj. $\boldsymbol{\beta}^{MEAN} = \boldsymbol{\beta}^{MED}$.

Stejně jako v případě OLS odhadu si i nyní odvodíme tvar asymptotického rozdělení (následující náznak odvození je převzat ze skript Omelka (2020)):

Tvrzení 20. *Nechť je splněn předpoklad linearity v mediánu, jsou splněny předpoklady regularity pro M-odhady (Omelka, 2020) a* $\hat{\boldsymbol{\beta}}^{LAD}$ *je* ***konzistentní*** *odhad parametr zájmu* $\boldsymbol{\beta}^{MED}$*, pak platí,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{LAD} - \boldsymbol{\beta}^{MED}) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}_p, \Gamma\Sigma\Gamma\right), \tag{4.3}$$

kde

- $\Sigma := \Sigma(\boldsymbol{\beta}^{MED}) = \mathsf{E}\, \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top$
- $\Gamma := \Gamma(\boldsymbol{\beta}^{MED}) = \left\{2\,\mathsf{E}(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top) f_\varepsilon(0)\right\}^{-1}$*, kde* f_ε *je hustota chybových členů.*

Důkaz (náznak). Opět víme, že se jedná o M-odhad, a k důkazu tak můžeme využít tvrzení 12. Vyjádření vektoru skórů je možné vyjádřit ve tvaru:

$$\psi(y, \widetilde{\boldsymbol{x}}, \boldsymbol{\beta}) = \frac{\partial \rho(y, \widetilde{\boldsymbol{x}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\operatorname{sgn}(y - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\beta})\widetilde{\boldsymbol{x}}.$$

Problém však nastává při vyjádření matice $\Gamma(\boldsymbol{\beta})$. Předpokládejme proto, že můžeme zaměnit derivaci a střední hodnotu, tj.

$$\Gamma(\boldsymbol{\beta}) = \Gamma^*(\boldsymbol{\beta}) := \frac{\partial\, \mathsf{E}\, \psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}.$$

Označme si $F := F_{Y|\boldsymbol{X}}$ podmíněné rozdělení $Y|\boldsymbol{X}$ a $f := f_{Y|\boldsymbol{X}}$ jeho příslušnou hustotu, pak z předpokladu absolutní spojitosti dostáváme, že

$$\begin{aligned}
\mathsf{E}\, \psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta}) &= -\,\mathsf{E}\left((I(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta} > 0) - I(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta} < 0))\widetilde{\boldsymbol{X}}\right) \\
&= -\,\mathsf{E}\left(\mathsf{E}\left((I(Y > \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}) - I(Y < \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}))\widetilde{\boldsymbol{X}}|\widetilde{\boldsymbol{X}}\right)\right) \\
&= -\,\mathsf{E}\left(1 - 2F(\widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta})\widetilde{\boldsymbol{X}}\right).
\end{aligned}$$

Pokud nyní tuto střední hodnotu zderivujeme podle $\boldsymbol{\beta}^\top$, z předpokladu linearity v mediánu dostáváme, že

$$\Gamma(\boldsymbol{\beta}^{MED}) = 2\,\mathsf{E}\left(f(\widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}^{MED})\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top\right) = 2\,\mathsf{E}\left(f_\varepsilon(0)\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top\right).$$

Z předpokladu linearity v mediánu dále také vyplývá, že $\mathsf{E}\, \psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta}^{MED}) = \mathbf{0}_p$, a tedy

$$\begin{aligned}
\Sigma(\boldsymbol{\beta}^{MED}) &= \mathsf{var}\, \psi(Y, \widetilde{\boldsymbol{X}}, \boldsymbol{\beta}^{MED}) \\
&= \mathsf{E}\left(-\operatorname{sgn}(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}^{MED})\widetilde{\boldsymbol{X}}\right)\left(-\operatorname{sgn}(Y - \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}^{MED})\widetilde{\boldsymbol{X}}\right)^\top \\
&= \mathsf{E}\, \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top.
\end{aligned}$$

□

Poznámka. Je nutno poznamenat, že se jedná **pouze** o náznak důkazu. Ztrátová funkce nemá derivaci v 0, a stejně tak nebyla nikde ospravedlněna záměna derivace a střední hodnoty.

Pozorování I. Hustota chybových členů f_ε stále může záviset na regresorech (nikde jsme nepředpokládali nezávislost, resp. stejné rozdělení).

Pozorování II. Pokud je podmíněné rozdělení $Y|\boldsymbol{X}$ symetrické, pak jsme schopni z vyjádření asymptotického rozptylu určit, za jakých podmínek bude odhad LAD lepší volbou než odhad OLS. Z tvrzení 16 a 20 totiž vyplývá, že pokud je rozdělení $Y|\boldsymbol{X}$ symetrické, a navíc je splněn i předpoklad homoskedasticity, pak

$$\mathsf{var}\,(\hat{\boldsymbol{\beta}}^{LAD}) < \mathsf{var}\,(\hat{\boldsymbol{\beta}}^{LS}) \quad \Longleftrightarrow \quad \frac{1}{4[f_\varepsilon(0)]^2} < \sigma^2. \tag{4.4}$$

Pokud by rozdělení $Y|\boldsymbol{X}$ bylo navíc i normální, pak by platilo $f_\varepsilon(0) = 1/\sqrt{2\pi\sigma^2}$, a nerovnost (4.4) by tedy nebyla splněna nikdy.

4.2 Praktické vlastnosti odhadu LAD

Důkaz následujícího tvrzení bychom provedli analogicky, jako tomu bude v případě regresní hloubky:

Tvrzení 21. *Odhad $\hat{\boldsymbol{\beta}}^{LAD}$ je regresně, škálově a afinně invariantní.*

Tvar influenční funkce pak opět přímo vyplývá z tvrzení 12 a z vyjádření vektoru ψ a matice Γ z tvrzení 20:

Lemma 22. ***Influenční funkci pro odhad získaný metodou LAD*** *je možné psát ve tvaru*

$$\mathsf{IF}((y,\boldsymbol{x}),\hat{\boldsymbol{\beta}}^{LAD},H) = (\mathsf{E}\,\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top)^{-1}\frac{\mathrm{sgn}(y-\beta_0^{MED}-\boldsymbol{x}^\top\boldsymbol{\beta}_X^{MED})}{2f_\varepsilon(0)}(1,\boldsymbol{x}^\top)^\top. \tag{4.5}$$

Pozorování. Opět můžeme z regresní invariance LAD odhadu BÚNO předpokládat $\boldsymbol{\beta}^{MED} = \mathbf{0}_p$. V takovém případě se influenční funkce redukuje do tvaru

$$\mathsf{IF}\,((y,\boldsymbol{x}),\hat{\boldsymbol{\beta}}^{LAD},H) = (\mathsf{E}\,\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top)^{-1}\frac{\mathrm{sgn}(y)}{2f_\varepsilon(0)}(1,\boldsymbol{x}^\top)^\top.$$

Z vyjádření influenční funkce vidíme, že ať už změníme hodnoty odezvy libovolným způsobem, tato změna bude mít vždy pouze omezený vliv. To ukazuje, že odhad LAD bude robustnější než odhad OLS, alespoň tedy vzhledem k odezvě.

5. Metoda regresní hloubky

Metodu regresní hloubky, zkráceně pouze *RD* (z *a.j. regression depth*) jako jednu z novějších robustních regresních metod poprvé představili Rousseeuw a Hubert v roce 1996. Hlavní myšlenka této metody stojí na konceptu tzv. „zanoření“, kterým se autoři inspirovali u Tukeyho poloprostorové hloubky (Tukey, 1975). I přesto, že se tomuto tématu v posledních letech věnovalo i několik dalších statistiků (Van Aelst, S., Struyf, A., etc.), se regresní hloubka na rozdíl od hloubky poloprostorové nikdy nestala příliš populární. Proč se tomu tak nestalo bude, spolu s jejími vlastnostmi, podrobněji rozebráno v této kapitole.

Ještě než přejdeme k samotné definici regresní hloubky, je potřeba upozornit na pár důležitých poznatků:

- I když byla teorie regresní hloubky poprvé představena před více než 20 lety, nebyla doposud publikována žádná práce shrnující základní poznatky, ke kterým v rámci jednotlivých výzkumů došlo. Bohužel, tato skutečnost má neblahé dopady. V první řadě tak nikdy nedošlo k zavedení jednotného značení, což znamená, že značení používané v jednotlivých pracích je značně nekonzistentní, a v některých případech má za důsledek dokonce mírně odlišné definice základních pojmů (např. i samotné regresní hloubky). Protože v rámci této práce čerpáme hned z několika různých zdrojů, bylo v některých případech nutné upravit definice tak, aby byl celkový text konzistentní. Pokud k nějakým takovým úpravám v rámci dané definice docházelo, jsou tyto úpravy vždy výše/níže uvedeny, a pokud to bylo uznáno za nutné, tak i podrobněji rozebrány jejich důsledky.
- Za druhé je nutné zdůraznit, že některé důkazy nebyly pravděpodobně dostatečně ověřeny, a mohou tak obsahovat skryté chyby. Bohužel, ve většině případů úroveň náročnosti těchto důkazů značně přesahuje úroveň této práce, a nejsou proto podrobněji rozebrány. Rozhodně by však bylo vhodné, aby tvrzení, která jsou v této práci uvedena bez důkazu, byla dále pečlivěji prozkoumána a ověřena.

5.1 Motivace

Cílem regresní hloubky je podobně jako poloprostorové hloubky zobecnit pojem mediánu, tentokrát však do lineárních regresních modelů. V případě volby $p = 1$ (tj. když nemáme k dispozici žádný regresor) bychom tedy mohli očekávat, že se definice regresní hloubky bude s definicí mediánu shodovat.

Z předpokladu absolutně spojitého rozdělení odezvy vyplývá, že můžeme uvažovat uspořádaný výběr $Y_{[1]} < \cdots < Y_{[n]}$,[1] na jehož základě pak můžeme definovat odhad parametru $\boldsymbol{\beta} = \beta_0$ jako

$$\hat{\beta}_0 := \begin{cases} (Y_{[n/2]} + Y_{[n/2+1]})/2 & \text{pro } n \text{ sudé}, \\ Y_{\lfloor (n+1)/2 \rfloor} & \text{pro } n \text{ liché}. \end{cases} \tag{5.1}$$

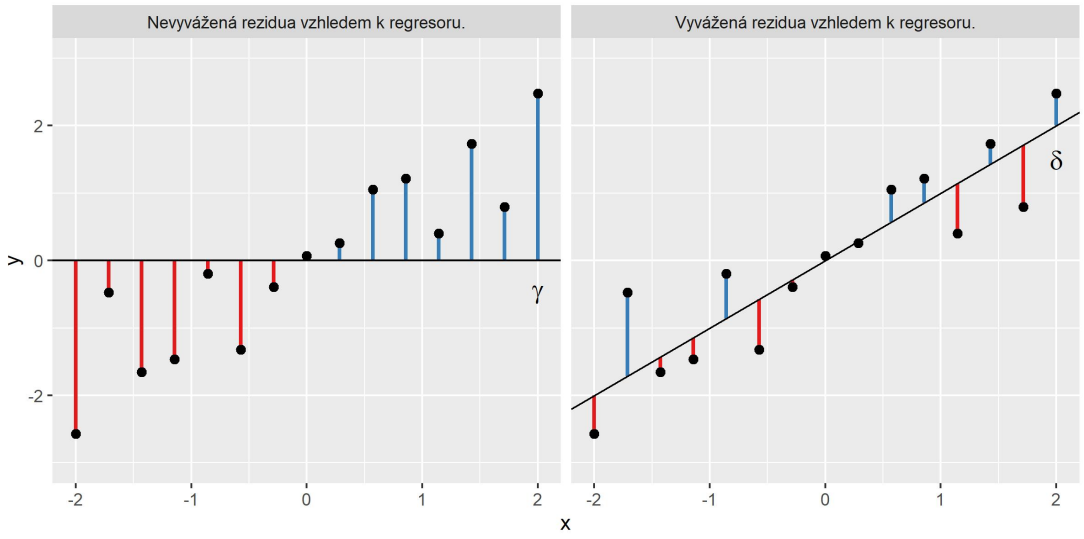
[1] kde $Y_{[i]}$ značí i-tou pořádkovou statistiku

Pro jednoduchost nyní předpokládejme, že n je liché. V takovém případě je možné odhad $\hat{\beta}_0$ ekvivalentně definovat jako řešení rovnice

$$\sum_{i=1}^{n} I(Y_i - \beta_0 \geq 0) = \sum_{i=1}^{n} I(Y_i - \beta_0 \leq 0), \tag{5.2}$$

tj. počty pozorování, jejichž rezidua mají kladné/záporné znaménko, se rovnají.

Příklad 1. Nyní se podívejme, jak se situace změní, pokud budeme navíc uvažovat jeden (netriviální) regresor. Na obrázku 5.1 je znázorněn náhodný výběr a k němu dva odhady regresní přímky, $\boldsymbol{\gamma} = (0, 0)^\top$ (vlevo) a $\boldsymbol{\delta} = (0, 1)^\top$ (vpravo). Pokud bychom nyní aplikovali analogii podmínky (5.2), zjistíme, že oba tyto odhady podmínku splňují. Takové kritérium tedy nebude postačující, protože odhad $\boldsymbol{\delta}$ představuje výrazně lepší odhad než $\boldsymbol{\gamma}$. V čem se ale navíc od sebe tyto odhady liší je to, že zatímco na obrázku vlevo jsou všechna pozorování s kladnými rezidui koncentrována na pravé straně a se zápornými na levé straně, na obrázku vpravo jsou všechna rezidua víceméně „rovnoměrně“ rozprostřena podél celé osy x. Jinak řečeno, ať už rozdělíme pozorování do dvou skupin libovolnou nadrovinou rovnoběžnou s osou y, bude zastoupení znamének reziduí v obou skupinách *vyvážené*.



Obrázek 5.1: Rozložení znamének reziduí pro dva vybrané odhady regresní přímky splňující analogii podmínky (5.2). Na obrázku můžeme vidět, že pro horší odhad dochází ke kumulaci znamének reziduí (vlevo), zatímco pro lepší odhad jsou znaménka reziduí „rovnoměrně“ rozprostřena podél osy x (vpravo).

Poznámka. Právě určitá „vyváženost“ znamének reziduí bez ohledu na volbu dělící nadroviny pro nás bude v budoucnu představovat základní myšlenku pro definici regresní hloubky.

Pozorování. Povšimněme si, že čím lépe jsou pozorování kolem odhadu regresní přímky rozložena, tím více je mezi ně přímka „zanořena“, a koncept tedy skutečně v něčem připomíná Tukeyho poloprostorovou hloubku.

5.2 Nonfit

Než přejdeme k samotné definici regresní hloubky, formalizujeme si myšlenku z motivace, a definujeme tak pojem, na jehož základě byla teorie regresní hloubky vybudována (Rousseeuw a Hubert, 1999).

Definice 23. *Nechť $\boldsymbol{\beta} \in \mathbb{R}^p$. Řekneme, že $\boldsymbol{\beta}$ je **nonfit**, pokud existují $\boldsymbol{u} \in \mathbb{R}^{p-1}$, $v \in \mathbb{R}$ taková, že pro všechna $i \in \{1, \ldots, n\}$ platí $\boldsymbol{u}^\top \boldsymbol{X}_i - v \neq 0$, a zároveň je pro všechna $i \in \{1, \ldots, n\}$ splněna jedna z následujících dvojic podmínek:*

$$\begin{aligned} &\boldsymbol{u}^\top \boldsymbol{X}_i - v < 0 \implies U_i(\boldsymbol{\beta}) < 0, \quad \text{a zároveň} \\ &\boldsymbol{u}^\top \boldsymbol{X}_i - v > 0 \implies U_i(\boldsymbol{\beta}) > 0, \end{aligned} \tag{5.3}$$

nebo

$$\begin{aligned} &\boldsymbol{u}^\top \boldsymbol{X}_i - v < 0 \implies U_i(\boldsymbol{\beta}) > 0, \quad \text{a zároveň} \\ &\boldsymbol{u}^\top \boldsymbol{X}_i - v > 0 \implies U_i(\boldsymbol{\beta}) < 0. \end{aligned} \tag{5.4}$$

Poznámka I. Nechť $\boldsymbol{a} \in \mathbb{R}^{p-1}, b \in \mathbb{R}, c \in \mathbb{R}$ taková, že $\sqrt{\boldsymbol{a}^\top \boldsymbol{a} + b^2} > 0$ a uvažujme obecnou rovnici nadroviny v prostoru $\mathbb{R}^p$ ve tvaru

$$\rho : \boldsymbol{a}^\top \boldsymbol{x} + by + c = 0. \tag{5.5}$$

Omezme se nyní pouze na nadroviny, které jsou rovnoběžné s osou y. Rovnice (5.5) musí být v takovém případě pro pevné $\boldsymbol{x} \in \mathbb{R}^{p-1}$ splněna pro libovolné $y \in \mathbb{R}$, z čehož vyplývá, že $b = 0$. Pokud ale trochu poupravíme značení a zvolíme si $\boldsymbol{u} := \boldsymbol{a}$, $c := -v$, pak přímo dostáváme rovnici nadroviny používané v definici 23. Definice 23 tedy říká, že $\boldsymbol{\beta}$ je nonfit právě tehdy, pokud existuje dělící nadrovina rovnoběžná s osou y taková, že všechna pozorování, která leží v jednom z jí určených poloprostorů, mají stejná znaménka reziduí.

Pozorování I. Vraťme se zpět k obrázku 5.1 vlevo a zvolme si jako dělící nadrovinu $x = -0.2$, pak je zřejmě $\boldsymbol{\gamma}$ v tomto případě nonfit.

Pozorování II. Pokud bychom v definici nonfitu náhodou uvažovali situaci kdy $\boldsymbol{u} = \boldsymbol{0}_{p-1}, v = 0$, tj. byl by porušen předpoklad $\sqrt{\boldsymbol{a}^\top \boldsymbol{a} + b^2} > 0$, pak by byl $\boldsymbol{\beta}$ nonfit právě tehdy, pokud by *všechna* pozorování měla stejná znaménka reziduí.

Definice nonfitu na základě geometrické interpretace

Výše uvedená definice nonfitu pomocí znamének reziduí se ukáže později jako stěžejní pro zavedení regresní hloubky (viz definice 25 níže). Nonfit lze ale definovat i jiným způsobem, a to na základě geometrické interpretace (která se ukáže později jako užitečná pro některé z důkazů). Než si geometrickou definici představíme, vraťme se ještě naposledy k obrázku 5.1:

Pozorování. Na obrázku 5.1 proveďme „rotaci“ nadrovinou určenou vektorem $\boldsymbol{\gamma}$ po směru hodinových ručiček (okolo průsečíku s nadrovinou $x = -0.2$). V takovém případě není těžké nahlédnout, že na rozdíl od „rotace“ nadrovinou určenou vektorem $\boldsymbol{\delta}$, který není nonfit, při „rotaci“ nonfitem „neprojdeme“ žádným z pozorování.

Geometrická definice nonfitu formalizuje předchozí pozorování spolu s pojmy „rotace nadrovinou“/„projít pozorováním“ a zobecňuje je do p-rozměrného prostoru:

Definice 24 (Geometrická)**.** *Vektor koeficientů* $\boldsymbol{\beta}$ *je* ***nonfit*** *právě tehdy, pokud existují* $\boldsymbol{u} \in \mathbb{R}^{p-1} \backslash \{\mathbf{0}_{p-1}\}$, $v \in \mathbb{R}$ *taková, že při označení nadrovin* $\rho_1 \colon \boldsymbol{u}^\top \boldsymbol{x} - v = 0$, $\rho_2 \colon \boldsymbol{\beta}_X^\top \boldsymbol{x} - y + \beta_0 = 0$ *a jejich příslušných normálových vektorů* $\boldsymbol{n}_1 = (\boldsymbol{u}^\top, 0)^\top$, $\boldsymbol{n}_2 = (\boldsymbol{\beta}_X^\top, -1)^\top$ *pro všechna* $i \in \{1, \ldots, n\}$, $\lambda \in [0, 1]$ *platí*

$$\boldsymbol{Z}_i \notin \rho(P, \boldsymbol{n}(\lambda)),$$

kde

1. $P \in \mathbb{R}^p$ *je libovolný bod takový, že* $P \in \rho_1 \cap \rho_2$,
2. $\boldsymbol{n}(\lambda) = \lambda \boldsymbol{n}_1 + (1 - \lambda)\boldsymbol{n}_2$, *a*
3. $\rho(P, \boldsymbol{n}(\lambda))$ *je nadrovina procházející bodem* P *s normálovým vektorem* $\boldsymbol{n}(\lambda)$.

Poznámka. Vektor koeficientů $\boldsymbol{\beta}$ také představuje nadrovinu dle rovnice (5.5) při volbě $b = -1, \boldsymbol{a} = \boldsymbol{\beta}_X, c = \ \beta_0$.

Poznámka I. Na rozdíl od definice 23 nebereme do úvahy situace kdy $\boldsymbol{u} = \mathbf{0}_{p-1}$, v takovém případě totiž není možné dělící nadrovinu zkonstruovat.

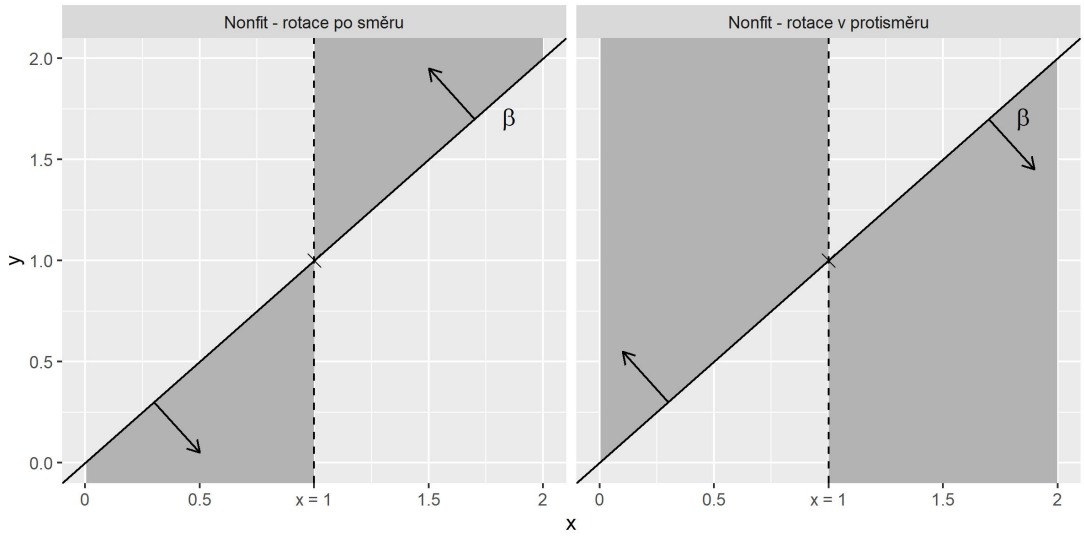
Pozorování I. Protože nadrovina určená vektorem koeficientů $\boldsymbol{\beta}$ nikdy nebude rovnoběžná s osou y, platí, že $\rho_1 \cap \rho_2 \neq \emptyset$, a bod P tedy vždy existuje.

Příklad 2. Pro ilustraci se podívejme, jakou množinu tvoří nadroviny $\rho(P, \boldsymbol{n}(\lambda))$ v prostoru $\mathbb{R}^2$. Nadroviny ρ_1, ρ_2 se v $\mathbb{R}^2$ redukují pouze na přímky, a platí tedy, že $\rho_1 \cap \rho_2 =: P \in \mathbb{R}^2$ (tj. průnik tvoří *právě jeden* bod). Množina přímek $\{\rho(P, \boldsymbol{n}(\lambda)), \lambda \in [0, 1]\}$ pak vždy tvoří jednu ze dvou oblastí (viz obrázek 5.2), a to v závislosti na směru normálového vektoru $\boldsymbol{n}_1$ (resp. na tom, jestli uvažujeme nadrovinu určenou rovnicí ρ_1 nebo $-\rho_1$, podle toho totiž dostaneme normálový vektor $\boldsymbol{n}_1$ nebo $-\boldsymbol{n}_1$). Volba normálového vektoru $\boldsymbol{n}_1$ se dá zároveň interpretovat i jako volba směru „rotace“ vektoru koeficientů $\boldsymbol{\beta}$.

Pozorování II. Protože uvažujeme všechny možné volby $\boldsymbol{u} \in \mathbb{R}^{p-1} \backslash \{\mathbf{0}_{p-1}\}, v \in \mathbb{R}$, je vždy možné zvolit $\boldsymbol{u}^* = -\boldsymbol{u}, v^* = -v$, a „rotovat“ tak vektorem koeficientů $\boldsymbol{\beta}$ i v opačném směru. Z definice nonfitu 24 pak vyplývá, že pokud má být vektor koeficientů $\boldsymbol{\beta}$ nonfit, pak při rotaci v některém ze směrů nesmí „projít“ žádným z pozorování.

Poznámka II. Pro zjednodušení budeme v práci později používat dva neformální pojmy, a to *rotovat nadrovinou* a *procházet pozorováním*. Rotací nadrovinou budeme chápat přechod od regresní nadroviny k nonfitu (nebo jiné nadrovině v $\mathbb{R}^p$) ve smyslu definice 24 přes všechny možné volby $\lambda \in [0, 1]$. Projít bodem $\boldsymbol{Z}_i$ pak bude znamenat, že existuje $\lambda \in [0, 1]$ takové, že platí $\boldsymbol{Z}_i \in \rho(P, \boldsymbol{n}(\lambda))$.

Poznámka III. V prostoru $\mathbb{R}^2$ budeme vždy rotovat přímku kolem bodu. Pokud ale budeme uvažovat prostor o obecné dimenzi p, nebude se již dále jednat pouze o rotaci kolem bodu, nýbrž kolem $(p-2)$-rozměrného podprostoru $\rho_1 \cap \rho_2$. Platí totiž, že pokud existuje bod $P \in \mathbb{R}^p$ takový, že platí $P \in \rho_1 \cap \rho_2$, pak již nutně $(\rho_1 \cap \rho_2) \subset \rho(P, \boldsymbol{n}(\lambda))$.



Obrázek 5.2: Geometrická definice nonfitu. V obou případech uvažujeme stejnou volbu regresní přímky $\boldsymbol{\beta}$ a dělící nadroviny $x = 1$, které se protínají v bodě $P = [1, 1]$. Na obrázku vlevo je pak vyznačena oblast všech přímek jejichž normálový vektor je konvexní kombinací vektorů $\boldsymbol{n}_1 = (1, 0)^\top$ a $\boldsymbol{n}_2 = (1, -1)^\top$. Na obrázku vpravo je naopak vyznačena oblast všech přímek jejichž normálový vektor je konvexní kombinací vektorů $-\boldsymbol{n}_1 = (-1, 0)^\top$ a $\boldsymbol{n}_2 = (1, -1)^\top$ (tj. místo nadroviny ρ_1 jsme uvažovali nadrovinu $-\rho_1$, které jsou sice shodné, ale v našem zápisu dávají opačné normálové vektory).

5.3 Regresní hloubka vzhledem k náhodnému výběru

Nyní můžeme konečně přejít k definici regresní hloubky. Její definice je spojena s definicí nonfitu, a byla tak společně s ní poprvé představena v práci Rousseeuw a Hubert (1999) v následujícím znění:

> „*Regresní hloubkou vektoru koeficientů $\boldsymbol{\beta}$ budeme rozumět nejmenší počet pozorování, který musí být z datového souboru odebrán, aby se z $\boldsymbol{\beta}$ stal nonfit. Ekvivalentně, regresní hloubka určuje nejmenší počet pozorování, která musí v takovém případě změnit znaménko rezidua.*“

Poznámka. I když se autoři pravděpodobně snažili o co nejvíce intuitivní pojetí, ukázala se bohužel původní definice jako příliš vágní. Definici regresní hloubky si totiž někteří autoři mírně upravili, např. opomenutím situací, kdy některá z pozorování leží přímo na regresní přímce (Bai a He, 1999).

Definice 25. ***Regresní hloubkou vektoru koeficientů $\boldsymbol{\beta} \in \mathbb{R}^p$ vzhledem k náhodnému výběru*** $\mathbb{Z}$ *budeme rozumět funkci* $rdepth(\cdot, \cdot) \colon \mathbb{R}^p \times \mathbb{R}^{(n \times p)} \longrightarrow \{0, \ldots, n\}$*, která splňuje následující definici*

$$rdepth(\boldsymbol{\beta}, \mathbb{Z}) = \inf_{\boldsymbol{u}, v} \sum_{i=1}^{n} I\Big(U_i(\boldsymbol{\beta})(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big), \tag{5.6}$$

kde infimum je uvažované přes $\boldsymbol{u} \in \mathbb{R}^{p-1}, v \in \mathbb{R}$ taková, že pro všechna $i \in \{1, \ldots, n\}$ je splněna podmínka $\boldsymbol{u}^\top \boldsymbol{X}_i - v \neq 0$.

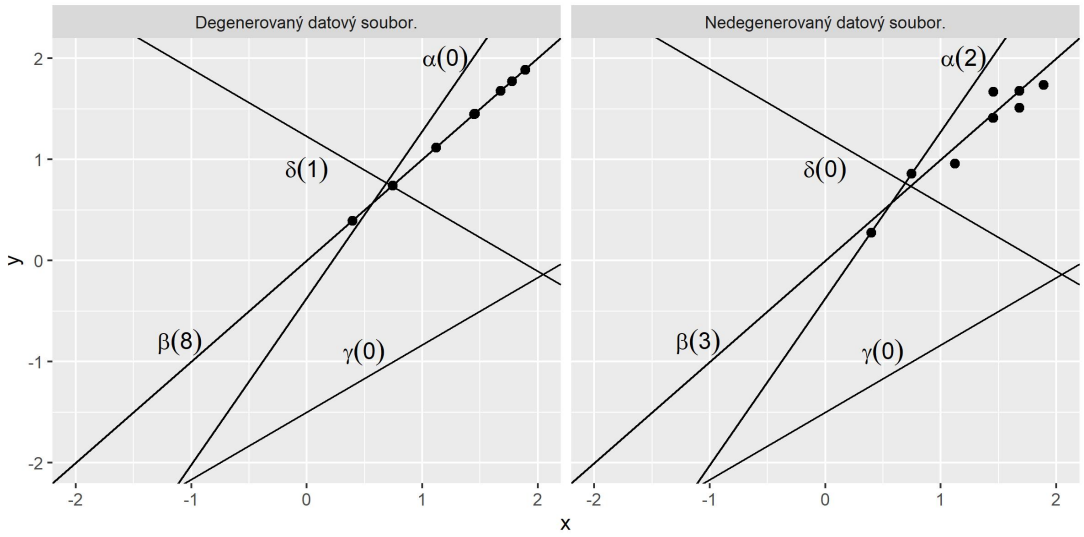
Poznámka k úpravě. Původní definice regresní hloubky v článku Bai a He (1999) byla ve tvaru

$$rdepth(\boldsymbol{\beta}, \mathbb{Z}) := \inf_{\boldsymbol{u}, v} \min \Big\{ \sum_{i=1}^{n} I\Big(U_i(\boldsymbol{\beta})(\boldsymbol{u}^\top \boldsymbol{X}_i - v) > 0\Big), \sum_{i=1}^{n} I\Big(U_i(\boldsymbol{\beta})(\boldsymbol{u}^\top \boldsymbol{X}_i - v) < 0\Big)\Big\},$$

a ve srovnání s naší definicí tak nezohledňovala případy, kdy pozorování přímo leží na regresní přímce, tj. $U_i(\boldsymbol{\beta}) = 0$. Navíc, pokud zvolíme $\boldsymbol{u}^* := -\boldsymbol{u}, v^* := -v$ je možné nerovnost v první sumě obrátit, a dostat tak vzorec pro druhou sumu s opačnou nerovností. Zápis je tak možné zjednodušit pouze na jednu sumu, protože druhou můžeme vždy získat volbou $\boldsymbol{u}, v$ s opačným znaménkem.

Pozorování. Z definice regresní hloubky vyplývá, že hodnota regresní hloubky se mění pouze v případě, kdy nadrovinou $\boldsymbol{\beta}$ *rotujeme* takovým způsobem, aby došlo ke změně znaménka alespoň u jednoho rezidua. Ke změně znaménka ale může dojít pouze v momentě, kdy regresní nadrovina daným pozorováním *prochází*.

Příklad 3. Na obrázku 5.3 můžeme vidět dva různé datové soubory a pro každý z nich 4 různé regresní přímky $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ a $\boldsymbol{\delta}$. Na obrázku vlevo je pak znázorněn degenerovaný datový soubor, kdy všechna pozorování leží na přímce $\boldsymbol{\beta}$, a regresní hloubka zde tak nabývá svého maxima n. Všechny ostatní přímky mají pak nutně hodnotu regresní hloubky 1 nebo 0, a to v závislosti na tom, zda některým z pozorování prochází. Na obrázku vpravo pak můžeme vidět nedegenerovaný datový soubor, kde na první pohled existuje lineární závislost mezi regresorem a odezvou. Přímka $\boldsymbol{\gamma}$ je zřejmě nonfit, stejně tak ale i přímka $\boldsymbol{\delta}$ (jako dělící nadrovinu zvolme $x = 0.5$). Přímka $\boldsymbol{\alpha}$ má v tomto případě hodnotu regresní hloubky 2, protože dvěma z pozorování přímo prochází, zatímco regresní hloubka přímky $\boldsymbol{\beta}$ je dokonce 3 (jako dělící nadrovinu zvolme $x = 0$).



Obrázek 5.3: Grafické znázornění degenerovaného (vlevo) a nedegenerovaného (vpravo) datového souboru se čtyřmi různými typy regresních přímek (hodnota regresní hloubky každé z přímek vzhledem k danému datovému souboru je pak vždy uvedena za označením přímky v závorce).

Definice 26. ***Odhadem parametru β metodou regresní hloubky*** *budeme rozumět odhad získaný na základě následujícího maximalizačního kritéria:*

$$\hat{\boldsymbol{\beta}}^{RD} := \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} rdepth(\boldsymbol{\beta}, \mathbb{Z}). \tag{5.7}$$

Poznámka. Odhad dle definice 26 budeme nazývat **odhadem regresního mediánu**, zkráceně pouze **odhad RD**.

Pozorování. Odhad regresního mediánu nemusí být určen jednoznačně, tj. může existovat hned několik argumentů (*kandidátů*), ve kterých funkce regresní hloubky svého maxima nabývá.

Definice 27. ***Úrovňovou množinou hloubky $l \in \{0, \ldots, n\}$*** *budeme rozumět množinu*

$$D_l(\mathbb{Z}) := \{\boldsymbol{\beta} \in \mathbb{R}^p \colon rdepth(\boldsymbol{\beta}, \mathbb{Z}) \geq l\}.$$

Pozorování. Označme si $l^* := \max_{\boldsymbol{\beta}\in\mathbb{R}^p} rdepth(\boldsymbol{\beta}, \mathbb{Z})$, pak $D_{l^*}(\mathbb{Z})$ je množina právě těch argumentů, ve kterých funkce regresní hloubky svého maxima nabývá.

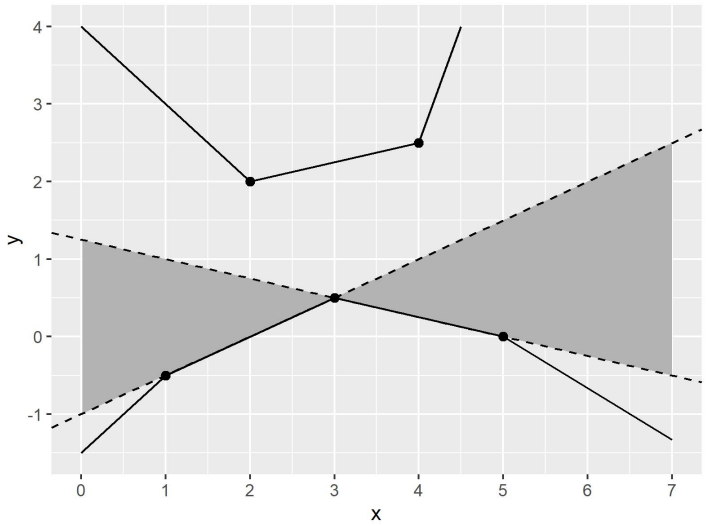
Otázkou tedy zůstává, jak v případě většího množství kandidátů odhad regresního mediánu definovat. První varianta je uvažovat jejich průměr. To lze ale pouze v případě, kdy je počet kandidátů konečný, což nemusí být vždy splněno (viz obrázek 5.4). Pokud je počet kandidátů nekonečný, můžeme buď některého z kandidátů náhodně zvolit, nebo jej zvolit na základě jedné z následujících dvou možností:

1. První možnost je uvažovat na kandidátech rovnoměrné rozdělení a definovat odhad jako jejich střední hodnotu, stejně jako tomu bude v případě regresní hloubky založené na distribuční funkci (viz sekce 5.4, pozorování u definice 31). Tento způsob se však příliš nepoužívá (pravděpodobně z výpočetních důvodů).

2. Druhá možnost je pak vybrat z množiny $D_{l^*}(\mathbb{Z})$ pouze konečný počet kandidátů, a pro ty opět spočítat jejich průměr. Předpokládejme nyní, že pozorování $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ jsou v obecné pozici[2] a uvažujme výběr kandidátů následujícím způsobem: Z předpokladu plné hodnosti regresní matice vyplývá, že existuje alespoň jedna kombinace p pozorování taková, že tato pozorování již jsou v obecné pozici. Protože se hodnota regresní hloubky může zvýšit pouze v případech, kdy regresní nadrovina při rotaci některým z pozorování *prochází* (viz pozorování za příkladem 3), a protože zároveň víme, že každá nadrovina je již v p-rozměrném prostoru p pozorováními v obecné pozici jednoznačně určena (z předpokladu obecné pozice regresorů navíc plyne, že tyto nadroviny nebudou rovnoběžné s osou y, a tedy se bude jednat o regresní nadroviny), stačí pro určení maximální hodnoty regresní hloubky uvažovat pouze ty regresní nadroviny, které prochází alespoň p pozorováními. Tyto nadroviny tak budou představovat naši vybranou skupinu kandidátů, a protože počet takových nadrovin je vždy maximálně $\binom{n}{p}$, bude konečný (Van Aelst a kol., 2002).

Pozorování. Konstrukce odhadu regresního mediánu v případě několika kandidátů nám zároveň dává návod, jak maximální hodnotu regresní hloubky spočítat.

[2]Pozorování v $\mathbb{R}^p$ jsou v obecné pozici právě tehdy, pokud žádných $p+1$ pozorování neleží v jedné nadrovině.



Obrázek 5.4: Příklad nespočetné množiny $D_{l^*}(\mathbb{Z})$. Uvažujme datový soubor o pěti pozorováních s hodnotami $[1, -0.5], [2, 2], [3, 0.5], [4, 2.5], [5, 0]$. Hodnota regresní hloubky všech přímek procházejících bodem $[3, 0.5]$ ležících ve vyznačené oblasti je pak 2, což je ale zároveň maximální možná hodnota regresní hloubky. Z toho vyplývá, že jsou součástí množiny kandidátů, která tím pádem ale nemůže být konečná.

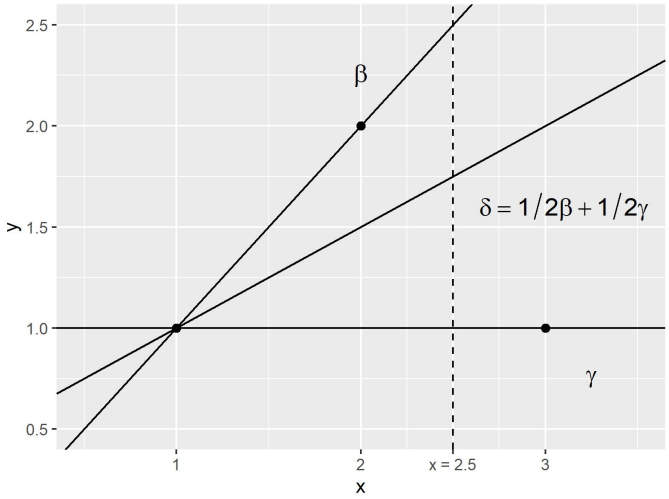
Odhad na základě průměru má oproti náhodnému zvolení některého z kandidátů několik výhod:

1. Jednoznačnost.
2. Vyšší eficience.
3. Dobré robustní vlastnosti, konkrétně hodnota BP, která je při použití tohoto typu odhadu stejná jako v případě, kdy by existoval pouze jeden kandidát (viz tvrzení 42).

Pozorování. Je tu však jedna nepříjemná skutečnost, a to ta, že množina $D_{l^*}(\mathbb{Z})$ není obecně konvexní, a není tedy nikde zaručeno, že průměr prvků z $D_{l^*}(\mathbb{Z})$ opět náleží do $D_{l^*}(\mathbb{Z})$ (viz obrázek 5.5). Odhad regresního mediánu definovaný tímto způsobem tak již nadále nemusí regresní hloubku maximalizovat.

5.4 Regresní hloubka vzhledem k distribuční funkci

Zatím byly všechny definice regresní hloubky konstruovány pouze na základě náhodného výběru $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, a jednalo se tedy pouze o odhad. Nyní se podíváme, co ve skutečnosti tento odhad vlastně odhaduje. Uvažujme distribuční funkci H generického vektoru $\boldsymbol{Z} = (Y, \boldsymbol{X}^\top)^\top$ a definujme si regresní hloubku vzhledem k distribuční funkci jakožto na funkcionál (Van Aelst a Rousseeuw, 2000):



Obrázek 5.5: Nekonvexita množiny $D_l^*(\mathbb{Z})$. Uvažujme datový soubor o třech pozorováních, $[1,1], [3,1], [2,2]$ a k němu tři regresní přímky $\boldsymbol{\beta} = (0,1)^\top$, $\boldsymbol{\gamma} = (1,0)^\top$ a $\boldsymbol{\delta} = (0.5, 0.5)^\top$. Regresní hloubka přímek $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ je v takovém případě maximální, tj. 2, v obou případech totiž procházíme právě dvěma ze tří pozorování, které jsou v obecné pozici. Regresní hloubka přímky $\boldsymbol{\delta}$ je ale pouze 1 (zvolme jako dělící nadrovinu $x = 2.5$), a to i přesto, že se jedná o jejich konvexní kombinaci.

Definice 28. ***Regresní hloubkou vektoru koeficientů*** $\boldsymbol{\beta} \in \mathbb{R}^p$ ***vzhledem k distribuční funkci*** $\boldsymbol{H}$ *s pravděpodobnostní mírou* P_H *budeme rozumět funkcionál* $rdepth(\cdot,\cdot)\colon \mathbb{R}^p \times \mathcal{H} \longrightarrow [0,1]$ *splňující následující definici*

$$rdepth(\boldsymbol{\beta}, H) := \inf_{\boldsymbol{u}, v} P_H\Big((Y - \beta_0 - \boldsymbol{X}^\top \boldsymbol{\beta}_X)(\boldsymbol{u}^\top \boldsymbol{X} - v) \geq 0\Big), \tag{5.8}$$

kde infimum je uvažované přes všechny $\boldsymbol{u} \in \mathbb{R}^{p-1}$, $v \in \mathbb{R}$ *splňující podmínku* $P_H(\boldsymbol{u}^\top \boldsymbol{X} = v) = 0$.

Poznámka. Stejně jako v případě definice regresní hloubky založené na náhodném výběru byla i definice 28 v článku Bai a He (1999) pouze s ostrou nerovností.

Pozorování I. Pokud budeme předpokládat, že $H \in \mathcal{H}^C$, kde $\mathcal{H}^C$ značí množinu všech distribučních funkcí s absolutně spojitým rozdělením s kladnou hustotou, pak je podmínka $P_H(\boldsymbol{u}^\top \boldsymbol{X} = v) = 0$ splněna automaticky.

Pozorování II. Pokud budeme naopak předpokládat, že rozdělení H je diskrétní a rovnoměrné na množině $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, pak platí $rdepth(\boldsymbol{\beta}, \mathbb{Z}) = n\, rdepth(\boldsymbol{\beta}, H)$. Regresní hloubka založená na náhodném výběru tedy představuje pouze speciální typ regresní hloubky založené na distribuční funkci.

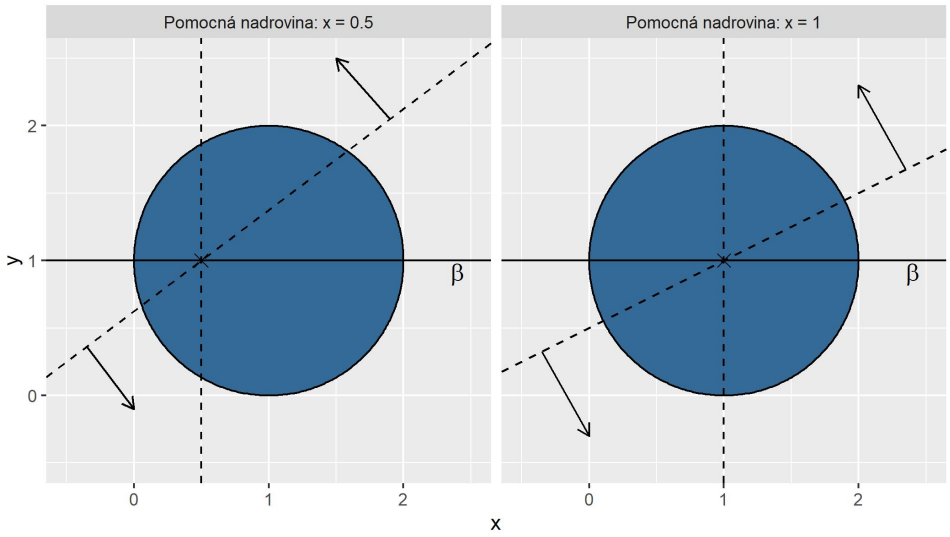
Lemma 29. *Nechť* $H \in \mathcal{H}^C$, *pak je hodnota regresní hloubky maximálně* $1/2$.

Důkaz (vlastní). Zvolme $\boldsymbol{u} \in \mathbb{R}^{p-1}$, $v \in \mathbb{R}$ tak, že platí

$$P_H\Big((Y - \beta_0 - \boldsymbol{X}^\top \boldsymbol{\beta}_X)(\boldsymbol{u}^\top \boldsymbol{X} - v) \geq 0\Big) > 1/2.$$

Protože $H \in \mathcal{H}^C$ dostáváme, že

$$P_H\Big((Y - \beta_0 - \boldsymbol{X}^\top \boldsymbol{\beta}_X)(\boldsymbol{u}^\top \boldsymbol{X} - v) \geq 0\Big) + P_H\Big((Y - \beta_0 - \boldsymbol{X}^\top \boldsymbol{\beta}_X)(\boldsymbol{u}^\top \boldsymbol{X} - v) \leq 0\Big) = 1.$$



Obrázek 5.6: Rovnoměrné rozdělení na kruhu. Ať již zvolíme dělící nadrovinu libovolným způsobem, např. $x = 0.5$ (vlevo) nebo $x = 1$ (vpravo), pravděpodobnostní masa, kterou „projdeme“ bude vždy $1/2$.

Zvolme tedy $\boldsymbol{u}^* := -\boldsymbol{u}$, $v^* := -v$, pak

$$\begin{aligned} & P_H\Big((Y - \beta_0 - \boldsymbol{X}^\top \boldsymbol{\beta}_X)(\boldsymbol{u}^{*\top} \boldsymbol{X} - v^*) \geq 0\Big) \\ & \quad = P_H\Big((Y - \beta_0 - \boldsymbol{X}^\top \boldsymbol{\beta}_X)(\boldsymbol{u}^\top \boldsymbol{X} - v) \leq 0\Big) \\ & \quad < 1/2. \end{aligned}$$

Protože regresní hloubka je definována jako infimum přes všechny možné volby $\boldsymbol{u} \in \mathbb{R}^{p-1}$, $v \in \mathbb{R}$, lze je vždy zvolit takovým způsobem, aby její hodnota byla menší nebo rovna $1/2$.

□

Pozorování. Z předchozího důkazu vyplývá, že pokud $H \in \mathcal{H}^C$ a pro nějakou volbu $\boldsymbol{u}, v$ je hodnota regresní hloubky vektorového parametru $1/2$, pak na volbě dělící nadroviny nezáleží (viz obrázek 5.6).

Definice 30. ***Regresním mediánem*** *budeme rozumět*

$$\boldsymbol{\beta}^{RD} := \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} rdepth(\boldsymbol{\beta}, H).$$

Stejně jako v případě regresní hloubky založené na náhodném výběru, je potřeba definovat regresní medián v případě, kdy existuje více argumentů, které regresní hloubku maximalizují. Pokud je rozdělení distribuční funkce H diskrétní a rovnoměrné, můžeme regresní medián definovat stejným způsobem, jako v případě jeho odhadu.

Definice 31. ***Úrovňovou množinou hloubky*** $\boldsymbol{l} \in [0, 1]$ *budeme rozumět množinu*

$$D_l(H) := \{\boldsymbol{\beta} \in \mathbb{R}^p \colon rdepth(\boldsymbol{\beta}, H) \geq l\}.$$

Pozorování. Označme si $l^* := \max_{\boldsymbol{\beta}\in\mathbb{R}^p} rdepth(\boldsymbol{\beta}, H)$, pak $D_{l^*}(H)$ je množina právě těch argumentů, ve kterých funkce regresní hloubky svého maxima nabývá.

Poznámka. Předpokládejme, že $H \in \mathcal{H}^C$ s hustotou $h > 0$ skoro všude, a navíc je úrovňová množina $D_{l^*}(H)$ **omezená** a vzhledem k prostoru $\mathbb{R}^p$ má **neprázdný vnitřek**, pak na ní můžeme uvažovat rovnoměrné rozdělení a regresní medián definujeme na základě rovnice

$$\boldsymbol{\beta}^{RD} := \Big(\int\limits_{D_{l^*}(H)} d\boldsymbol{\beta} \Big)^{-1} \int\limits_{D_{l^*}(H)} \boldsymbol{\beta}\, d\boldsymbol{\beta}. \tag{5.9}$$

5.5 Teoretické vlastnosti odhadu RD

V následující sekci se omezíme pouze na $H \in \mathcal{H}^C$.

5.5.1 Vztah mezi β^{RD} a β^{MED}

Stejně jako v případě metody LAD předpokládejme i nyní, že je splněn předpoklad linearity v mediánu, tj. předpokládáme, že existuje $\boldsymbol{\beta}^{MED} \in \mathbb{R}^p$ takové, že

$$\mathsf{med}\,(Y \mid \widetilde{\boldsymbol{X}}) = \widetilde{\boldsymbol{X}}^\top \boldsymbol{\beta}^{MED}. \tag{5.10}$$

Z lemmatu 29 pak víme, že pro $H \in \mathcal{H}^C$ nabývá regresní hloubka pouze hodnot z intervalu $[0, 1/2]$. Pokud je však navíc splněn i předpoklad linearity v mediánu, pak je možné ukázat, že jedním z argumentů maximalizujících regresní hloubku je právě parametr zájmu $\boldsymbol{\beta}^{MED}$.

Lemma 32. *Nechť $H \in \mathcal{H}^C$, a zároveň je splněn předpoklad linearity v mediánu, pak platí*

$$rdepth(\boldsymbol{\beta}^{MED}, H) = rdepth(\boldsymbol{\beta}^{RD}, H) = 1/2.$$

Důkaz (vlastní). Buď libovolné $\boldsymbol{u} \in \mathbb{R}^{p-1}$, $v \in \mathbb{R}$ a využijme skutečnosti, že platí

$$P_H(\dots) = \mathsf{E}_{\boldsymbol{Z}}\, I(\dots),$$

kde $\mathsf{E}_{\boldsymbol{Z}}(\cdot)$ značí střední hodnotu vzhledem k rozdělení náhodného vektoru $\boldsymbol{Z}$. Z předpokladu $H \in \mathcal{H}^C$ a z vlastností podmíněné střední hodnoty dostáváme:

$$\begin{aligned}
&P_H\Big((Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED})(\boldsymbol{u}^\top \boldsymbol{X} - v) \geq 0\Big) \\
&\quad = \mathsf{E}_{\boldsymbol{Z}}\, I\Big((Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED})(\boldsymbol{u}^\top \boldsymbol{X} - v) \geq 0\Big) \\
&\quad = \mathsf{E}_{\boldsymbol{Z}}\, \mathsf{E}_{Y|\boldsymbol{X}} \left(I\Big((Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED})(\boldsymbol{u}^\top \boldsymbol{X} - v) \geq 0\Big)\right) \\
&\quad = \mathsf{E}_{\boldsymbol{Z}} \Big(\mathsf{E}_{Y|\boldsymbol{X}} \left(I(Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED} \geq 0) I(\boldsymbol{u}^\top \boldsymbol{X} - v \geq 0)\right) \\
&\qquad + \mathsf{E}_{Y|\boldsymbol{X}} \left(I(Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED} \leq 0) I(\boldsymbol{u}^\top \boldsymbol{X} - v \leq 0)\right)\Big).
\end{aligned} \tag{5.11}$$

Protože náhodné veličiny $I(\boldsymbol{u}^\top \boldsymbol{X} - v \geq 0)$ a $I(\boldsymbol{u}^\top \boldsymbol{X} - v \leq 0)$ jsou $\sigma(\boldsymbol{X})$-měřitelné, je možné rovnici (5.11) psát ve tvaru

$$\begin{aligned} \mathsf{E}_{\boldsymbol{Z}} \Big(& I(\boldsymbol{u}^\top \boldsymbol{X} - v \geq 0)\, \mathsf{E}_{Y|\boldsymbol{X}}\, I(Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED} \geq 0) \\ & + I(\boldsymbol{u}^\top \boldsymbol{X} - v \leq 0)\, \mathsf{E}_{Y|\boldsymbol{X}}\, I(Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED} \leq 0) \Big). \end{aligned} \tag{5.12}$$

Navíc předpokládáme, že $H \in \mathcal{H}^C$, což nám zároveň s předpokladem linearity v mediánu dává rovnost

$$\mathsf{E}_{Y|\boldsymbol{X}}\, I(Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED} \geq 0) = P_F\Big(Y \geq \beta_0^{MED} + \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED}\Big) = 1/2,$$

a stejně tak

$$\mathsf{E}_{Y|\boldsymbol{X}}\, I(Y - \beta_0^{MED} - \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED} \leq 0) = P_F\Big(Y \leq \beta_0^{MED} + \boldsymbol{X}^\top \boldsymbol{\beta}_X^{MED}\Big) = 1/2,$$

z čehož získáme rovnici (5.12) ve tvaru

$$\mathsf{E}_{\boldsymbol{Z}} \Big(I(\boldsymbol{u}^\top \boldsymbol{X} - v \geq 0)\, 1/2 + I(\boldsymbol{u}^\top \boldsymbol{X} - v \leq 0)\, 1/2 \Big).$$

Předpoklad $H \in \mathcal{H}^C$ ale zároveň i zaručuje, že $\mathsf{E}_{\boldsymbol{Z}}(I(\boldsymbol{u}^\top \boldsymbol{X} - v = 0)) = 0$, a regresní hloubka vektoru $\boldsymbol{\beta}^{MED}$ tak skutečně bude vždy nabývat hodnoty $1/2$.

□

V předchozím tvrzení jsme tedy ukázali, že $\boldsymbol{\beta}^{MED}$ je jedním z argumentů maximalizujících regresní hloubku. Nyní budeme chtít ukázat, že je zároveň i argumentem **jediným**.

Následující lemmata a tvrzení jsou z velké části převzata (i s důkazy) z článku Rousseeuw a Hubert (1999), jednotlivé kroky jsou však podrobněji rozebrány a doplněny o grafické znázornění.

Definice 33. *Uvažujme distribuční funkci H s pravděpodobnostní mírou P_H a libovolné vektory koeficientů $\boldsymbol{\beta}^1,\, \boldsymbol{\beta}^2 \in \mathbb{R}^p$, pak definujme funkci*

$$d_H(\boldsymbol{\beta}^1,\, \boldsymbol{\beta}^2) := P_H(A(\boldsymbol{\beta}^1,\, \boldsymbol{\beta}^2)),$$

kde $A(\boldsymbol{\beta}^1,\, \boldsymbol{\beta}^2)$ je oblast definována jako

$$A(\boldsymbol{\beta}^1,\, \boldsymbol{\beta}^2) := \Big\{(\boldsymbol{x}^\top, y) : \boldsymbol{x} \in \mathbb{R}^{p-1}, y \in [a, b]\Big\},$$

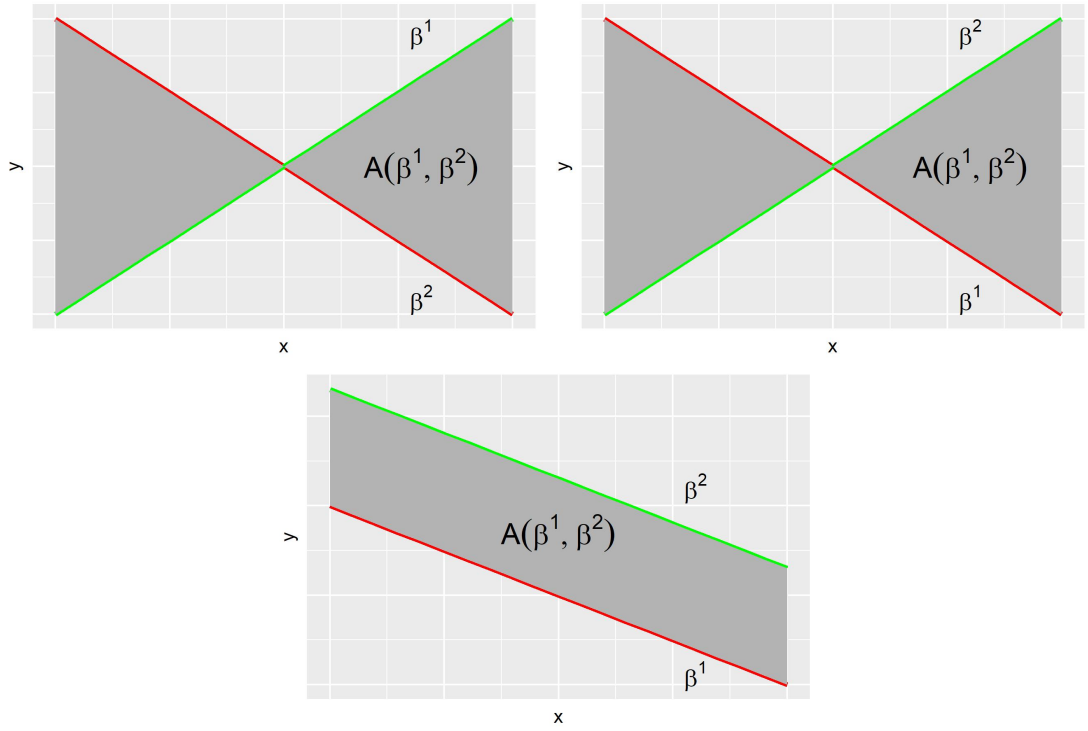
a

- $a := a(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{x}) = \min\{\beta_0^1 + \boldsymbol{x}^\top \boldsymbol{\beta}_X^1, \beta_0^2 + \boldsymbol{x}^\top \boldsymbol{\beta}_X^2\}$,
- $b := b(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{x}) = \max\{\beta_0^1 + \boldsymbol{x}^\top \boldsymbol{\beta}_X^1, \beta_0^2 + \boldsymbol{x}^\top \boldsymbol{\beta}_X^2\}$.

Poznámka. Vybrané typy oblastí $A(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2)$ jsou znázorněny v obrázku 5.8.

Lemma 34. *Pro libovolnou distribuční funkci $H \in \mathcal{H}^C$ je d_H metrika na $\mathbb{R}^p$.*

Důkaz. Z definice metriky potřebujeme ukázat, že pro libovolné vektory koeficientů $\boldsymbol{\beta}^1,\, \boldsymbol{\beta}^2, \boldsymbol{\beta}^3 \in \mathbb{R}^p$ jsou splněny následující podmínky:



Obrázek 5.7: Grafické znázornění oblasti $A(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2)$, resp. $A(\boldsymbol{\beta}^2, \boldsymbol{\beta}^1)$, pro tři různé volby $\boldsymbol{\beta}^1$, $\boldsymbol{\beta}^2$ v prostoru o dimenzi $p = 2$.

1. $d_H(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2) \geq 0$,
2. $d_H(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2) = d_H(\boldsymbol{\beta}^2, \boldsymbol{\beta}^1)$,
3. $d_H(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2) = 0 \iff \boldsymbol{\beta}^1 = \boldsymbol{\beta}^2$ a
4. $d_H(\boldsymbol{\beta}^1, \boldsymbol{\beta}^3) \leq d_H(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2) + d_H(\boldsymbol{\beta}^2, \boldsymbol{\beta}^3)$.

Bod 1 plyne přímo z definice funkce d_H na základě distribuční funkce, bod 2 ze symetrie množiny A, bod 3 z předpokladu, že $H \in \mathcal{H}^C$, a bod 4 nakonec ze skutečnosti, že pro každé $\boldsymbol{x} \in \mathbb{R}^{p-1}$ platí inkluze

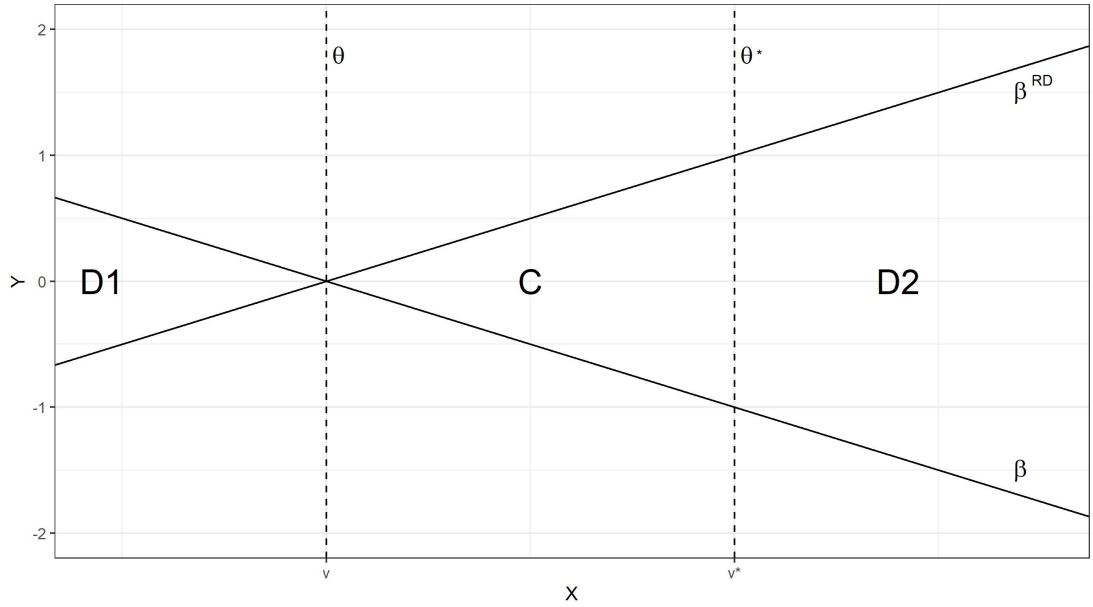
$$[a(\boldsymbol{\beta}^1, \boldsymbol{\beta}^3, \boldsymbol{x}),\, b(\boldsymbol{\beta}^1, \boldsymbol{\beta}^3, \boldsymbol{x})] \subset [a(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{x}),\, b(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{x})] \cup [a(\boldsymbol{\beta}^2, \boldsymbol{\beta}^3, \boldsymbol{x}),\, b(\boldsymbol{\beta}^2, \boldsymbol{\beta}^3, \boldsymbol{x})].$$

□

Lemma 35. *Nechť $H \in \mathcal{H}^C$, $p = 2$, a zároveň je splněn předpoklad linearity v mediánu, pak pro libovolný vektor koeficientů $\boldsymbol{\beta} \in \mathbb{R}^p$ platí, že*

$$rdepth(\boldsymbol{\beta},\, H) = 1/2 - d_H(\boldsymbol{\beta}^{RD},\, \boldsymbol{\beta}). \tag{5.13}$$

Důkaz (podrobně rozepsáno). Rovnost (5.13) budeme dokazovat pomocí dvou nerovností. Nejprve si dokážeme nerovnost $rdepth(\boldsymbol{\beta},\, H) \leq 1/2 - d_H(\boldsymbol{\beta}^{RD},\, \boldsymbol{\beta})$:



Obrázek 5.8: Grafické znázornění rozdělení oblastí v lemmatu 35.

Uvažujme libovolný vektor koeficientů $\boldsymbol{\beta} \in \mathbb{R}^2$. Z pozorování za lemmatem 29 víme, že při výpočtu regresní hloubky $\boldsymbol{\beta}^{RD}$ nezáleží na volbě v, tj. nezáleží, okolo kterého bodu regresní přímkou *rotujeme*. Zvolme tedy za v hodnotu x-ové složky průsečíku přímek určených vektory koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\beta}^{RD}$, a proveďme *rotaci* přímkou určenou vektorem koeficientů $\boldsymbol{\beta}^{RD}$ okolo tohoto průsečíku takovým způsobem, abychom získali přímku $\boldsymbol{\theta} := \{(v, y)^\top : y \in \mathbb{R}\}$ (ta je rovnoběžná s osou y, viz obrázek 5.8). Z lemmatu 32 víme, že $rdepth(\boldsymbol{\beta}^{RD}, H) = 1/2$, tj. pravděpodobnostní masa, kterou jsme při rotaci prošli je rovna $1/2$ (a to jak při rotaci po směru, tak v protisměru hodinových ručiček). Při jedné z rotací jsme tedy museli „projít“ přímkou totožnou s přímkou $\boldsymbol{\beta}$, z čehož již vyplývá nerovnost

$$1/2 \geq d_H(\boldsymbol{\beta}^{RD}, \boldsymbol{\beta}) + rdepth(\boldsymbol{\beta}, H). \tag{5.14}$$

Nyní si dokážeme opačnou nerovnost. Předpokládejme pro spor, že vektor koeficientů $\boldsymbol{\beta}$ nabývá regresní hloubky v libovolném bodě $v^* \in \mathbb{R}$ takovém, že $v^* \neq v$, a stejně jako v předchozí části si označme $\boldsymbol{\theta}^* := \{(v^*, y)^\top : y \in \mathbb{R}\}$ přímku rovnoběžnou s osou y procházející bodem v^*. Proveďme opět rotaci přímkou určenou vektorem koeficientů $\boldsymbol{\beta}$, tentokrát okolo průsečíku $\boldsymbol{\beta}$ a $\boldsymbol{\theta}^*$, a uvažujme přitom následující značení (na obrázku 5.8 platí $D = D_1 \cup D_2$):

$$C = \{(x, y) \colon x \in [\min(v, v^*), \max(v, v^*)],\ y \in [\min(a, b), \max(a, b)]\},$$
$$D = \{(x, y) \colon x \notin [\min(v, v^*), \max(v, v^*)],\ y \in [\min(a, b), \max(a, b)]\}.$$

kde $a := a(\boldsymbol{\beta}, x) = \beta_0 + x\beta_1$, $b := (\boldsymbol{\beta}^{RD}, x) = \beta_0^{RD} + x\beta_1^{RD}$. V závislosti na směru rotace pak vždy nastane jedna ze dvou možností:

1. Při rotaci cestou „projdeme“[3] přímkou určenou vektorem koeficientů $\boldsymbol{\beta}^{RD}$ (na obrázku 5.8 odpovídá rotaci v protisměru).

[3]Tj. během rotace dostaneme přímku, která je s ní rovnoběžná.

2. Při rotaci cestou „neprojdeme“ přímkou určenou vektorem koeficientů $\boldsymbol{\beta}^{RD}$ (na obrázku 5.8 odpovídá rotaci po směru).

V obou případech však víme, že k výpočtu regresní hloubky $\boldsymbol{\beta}^{RD}$ je možné využít rotaci okolo bodu v^*, a platí $rdepth(\boldsymbol{\beta}^{RD}, H) = 1/2$. Pokud tedy nastane situace 1, pak jsme během rotace „prošli“ oblast M^1, kde

$$P_H(M^1) = 1/2 + P_H(D) - P_H(C) = 1/2 - (P_H(C) - P_H(D)).$$

Z předpokladů, že $v^* \neq v$ a $H \in \mathcal{H}^C$ pak vyplývá $P_H(D) > 0$, a tedy

$$P_H(M^1) > 1/2 - (P_H(C) + P_H(D)) = 1/2 - P_H(C \cup D). \tag{5.15}$$

Pokud naopak nastane situace 2, pak jsme během rotace prošli oblastí M^2, kde

$$P_H(M^2) = 1/2 + P_H(C) - P_H(D) = 1/2 - (P_H(D) - P_H(C)).$$

Stejně jako v předchozím případě tak z předpokladů $v^* \neq v$ a $H \in \mathcal{H}^C$ vyplývá, že $P_H(C) > 0$, a tedy

$$P_H(M^2) > 1/2 - (P_H(D) + P_H(C)) = 1/2 - P_H(D \cup C). \tag{5.16}$$

Z rovnic (5.15) a (5.16) nakonec dostáváme, že

$$\min\{P_H(M^1), P_H(M^2)\} > 1/2 - P_H(C \cup D),$$

a protože $A(\boldsymbol{\beta}, \boldsymbol{\beta}^{RD}) = C \cup D$, také

$$\min\{P_H(M^1), P_H(M^2)\} > 1/2 - d_H(\boldsymbol{\beta}, \boldsymbol{\beta}^{RD}).$$

To je ale ve sporu s nerovností (5.14), z čehož vyplývá, regresní hloubky musí $\boldsymbol{\beta}$ nabývat v bodě v. V takovém případě však $P_H(C) = 0$, z čehož vyplývá rovnost $d_H(\boldsymbol{\beta}, H) = 1/2 - d_H(\boldsymbol{\beta}^{RD}, \boldsymbol{\beta})$.

□

Poznámka. Tato kapitola ilustruje hlavní problém regresní hloubky, a tím je formalizace některých pojmů pro vyšší dimenze. Protože se koncept regresní hloubky ve velké míře opírá o grafické znázornění a geometrii, množství důkazů je vedeno tímto způsobem, a stejně tak je tomu i v případě lemmatu 35. V originálním článku Van Aelst a Rousseeuw (2000) je důkaz sice uveden i pro $p \geq 3$, autoři však pracují s pojmem rotace v intuitivním smyslu bez jakékoliv formální definice. Myšlenka důkazu tak sice zůstává stejná, zatímco ale pro $p = 2$ je možné využít relativně přesnou geometrickou představu, pro $p \geq 3$ je již situace složitější. Formální důkaz by tak mohl spočívat ve využití normálových vektorů a nonfitu dle definice 24, bohužel toto zobecnění přesahuje rozsah této práce.

Tvrzení 36. *Nechť $H \in \mathcal{H}^C$, a zároveň je splněn předpoklad linearity v mediánu, pak platí*

$$\boldsymbol{\beta}^{RD} = \boldsymbol{\beta}^{MED}.$$

Důkaz. Nechť $p = 2$. Z lemmatu 32 víme, že platí $rdepth(\boldsymbol{\beta}^{MED}, H) = 1/2$. Pokud však $\boldsymbol{\beta}^{MED}$ dosadíme do rovnice (5.13), pak z lemmatu 35 dostáváme, že zároveň

$$rdepth(\boldsymbol{\beta}^{MED}, H) = 1/2 - d_H(\boldsymbol{\beta}^{RD}, \boldsymbol{\beta}^{MED}).$$

Protože dle lemmatu 34 je d_H metrika, platí $d_H(\boldsymbol{\beta}^{RD}, \boldsymbol{\beta}^{MED}) = 0$ právě tehdy, když $\boldsymbol{\beta}^{RD} = \boldsymbol{\beta}^{MED}$. Důkaz pro $p \geq 3$ by pak byl analogický, pokud se spoleh-neme na platnost lemmatu 35 pro vyšší dimenze z původního článku Van Aelst a Rousseeuw (2000).

□

Důsledek 1. *Pokud tedy $H \in \mathcal{H}^C$, a zároveň je splněn předpoklad linearity v me-diánu, pak je parametr $\boldsymbol{\beta}^{RD}$ určen jednoznačně a je roven parametru zájmu $\boldsymbol{\beta}^{MED}$.*

5.5.2 Konzistence

Důkaz konzistence můžeme nalézt například v článcích Bai a He (1999) nebo Zuo (2020), v obou případech se však důkaz opírá o teorii stochastických procesů, která značně přesahuje úroveň této práce. V našem případě jsme se proto rozhodli vycházet z důkazu z článku Bai a He (1999), na základě kterého alespoň:

1. Dokážeme ekvivalentní tvar regresní hloubky (5.19) (který byl v původním článku pouze uveden bez důkazu) a rozšíříme jej i pro pozorování, která přímo náleží regresní přímce.

2. Dokážeme ekvivalentní tvar regresní hloubky (5.24) (který v původním článku nebyl ani uveden).

3. Na základě předchozích dvou odvození nahlédneme tvrzení 39 o konzistenci z původního článku.

Naším cílem bude tedy nyní ukázat, že odhad $\hat{\boldsymbol{\beta}}^{RD}$ je pro $H \in \mathcal{H}^C$ za předpokladu linearity v mediánu konzistentním odhadem parametru $\boldsymbol{\beta}^{RD}$, tj. že platí

$$\arg\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} rdepth(\boldsymbol{\beta}, \mathbb{Z}) \xrightarrow{s.j.} \arg\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} rdepth(\boldsymbol{\beta}, H). \tag{5.17}$$

V prvním kroku bude tedy potřeba ukázat, že platí

$$\frac{rdepth(\boldsymbol{\beta}, \mathbb{Z})}{n} \overset{s.j.}{\rightrightarrows} rdepth(\boldsymbol{\beta}, H),$$

tj.

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \left| \frac{rdepth(\boldsymbol{\beta}, \mathbb{Z})}{n} - rdepth(\boldsymbol{\beta}, H) \right| \xrightarrow{s.j.} 0. \tag{5.18}$$

a ve druhém kroku pak dokázat konvergenci argumentů. K tomu si však budeme muset vyjádřit jak regresní hloubku založenou na náhodném výběru, tak regresní hloubku založenou na distribuční funkci, trochu jiným způsobem:

Lemma 37. *Regresní hloubku založenou na náhodném výběru z definice 25 lze psát ve tvaru:*

$$rdepth(\boldsymbol{\beta},\, \mathbb{Z}) = \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} I\Big(U_i(\boldsymbol{\beta}) = 0\Big) + \frac{1}{2}\inf_{\boldsymbol{\gamma}\in S^p}\sum_{i=1}^{n}\operatorname{sgn}(U_i(\boldsymbol{\beta}))\operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma}), \quad (5.19)$$

kde infimum je uvažované přes $\boldsymbol{\gamma} \in S^p$ takové, že pro všechna $i \in \{1,\dots,n\}$ je splněna podmínka $\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma} \neq 0$.

Důkaz (vlastní). Z definice 25 předpokládáme, že pro všechna $i \in \{1,\dots,n\}$ je splněna podmínka $\boldsymbol{u}^\top\boldsymbol{X}_i - v \neq 0$, regresní hloubku je tedy možné přepsat následujícím způsobem:

$$\inf_{\boldsymbol{u},\,v}\sum_{i=1}^{n} I\Big(U_i(\boldsymbol{\beta})(\boldsymbol{u}^\top\boldsymbol{X}_i - v) \geq 0\Big) = \inf_{\boldsymbol{u},\,v}\sum_{i=1}^{n} I\Big(\operatorname{sgn}(U_i(\boldsymbol{\beta})) = \operatorname{sgn}(\boldsymbol{u}^\top\boldsymbol{X}_i - v)\Big) + \sum_{i=1}^{n} I\Big(\operatorname{sgn}(U_i(\boldsymbol{\beta})) = 0\Big).$$

Zaveďme si dále následující značení:

- $d(\boldsymbol{\beta}, \boldsymbol{u}, v) := \sum_{i=1}^{n} I\Big(\operatorname{sgn}(U_i(\boldsymbol{\beta})) = \operatorname{sgn}(\boldsymbol{u}^\top\boldsymbol{X}_i - v)\Big)$,
- $c(\boldsymbol{\beta}) := \sum_{i=1}^{n} I\Big(\operatorname{sgn}(U_i(\boldsymbol{\beta})) = 0\Big) = \sum_{i=1}^{n} I\Big(U_i(\boldsymbol{\beta}) = 0\Big)$,

a přepišme regresní hloubku do zkráceného tvaru

$$rdepth(\boldsymbol{\beta},\, \mathbb{Z}) = c(\boldsymbol{\beta}) + \inf_{\boldsymbol{u},\,v} d(\boldsymbol{\beta}, \boldsymbol{u}, v). \quad (5.20)$$

Nyní se podíváme na pravou stranu rovnice (5.19) a ukažme, že odpovídá pravé straně rovnice (5.20). Za předpokladu, že pro všechna $i \in \{1,\dots,n\}$ je splněna podmínka $\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma} \neq 0$, z vlastností funkce signum vyplývá, že

$$\begin{aligned} \operatorname{sgn}(U_i(\boldsymbol{\beta}))\operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma}) = 1 \quad &\Longleftrightarrow \quad \operatorname{sgn}(U_i(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma}), \\ \operatorname{sgn}(U_i(\boldsymbol{\beta}))\operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma}) = -1 \quad &\Longleftrightarrow \quad \operatorname{sgn}(U_i(\boldsymbol{\beta})) = -\operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma}). \end{aligned} \quad (5.21)$$

Pokud si tedy opět zavedeme zkrácené značení

$$\begin{aligned} a(\boldsymbol{\beta},\boldsymbol{\gamma}) &:= \sum_{i=1}^{n} I(\operatorname{sgn}(U_i(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma})), \\ b(\boldsymbol{\beta},\boldsymbol{\gamma}) &:= \sum_{i=1}^{n} I(\operatorname{sgn}(U_i(\boldsymbol{\beta})) = -\operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top\boldsymbol{\gamma})), \end{aligned} \quad (5.22)$$

je možné pravou stranu rovnice (5.19) psát ve tvaru:

$$\frac{n}{2} + \frac{1}{2}\,c(\boldsymbol{\beta}) + \frac{1}{2}\inf_{\boldsymbol{\gamma}}\Big\{a(\boldsymbol{\beta},\boldsymbol{\gamma}) - b(\boldsymbol{\beta},\boldsymbol{\gamma})\Big\}. \quad (5.23)$$

Při vyjádření $b(\boldsymbol{\beta},\boldsymbol{\gamma})$ z rovnosti $n = a(\boldsymbol{\beta},\boldsymbol{\gamma}) + b(\boldsymbol{\beta},\boldsymbol{\gamma}) + c(\boldsymbol{\beta})$ a dosazení do rovnice (5.23) dostáváme pravou stranu rovnice (5.19) ve zjednodušeném tvaru

$$c(\boldsymbol{\beta}) + \inf_{\boldsymbol{\gamma}} a(\boldsymbol{\beta},\boldsymbol{\gamma}).$$

Abychom konečně získali ekvivalenci s rovnicí (5.20), je ještě potřeba ukázat, že platí:

$$\forall\, \boldsymbol{\gamma} \in S^p\, \exists \boldsymbol{u} \in \mathbb{R}^{p-1}\, \exists v \in \mathbb{R} : a(\boldsymbol{\beta}, \boldsymbol{\gamma}) = d(\boldsymbol{\beta}, \boldsymbol{u}, v),$$
$$\forall\, \boldsymbol{u} \in \mathbb{R}^{p-1}\, \forall v \in \mathbb{R}\, \exists \boldsymbol{\gamma} \in S^p : d(\boldsymbol{\beta}, \boldsymbol{u}, v) = a(\boldsymbol{\beta}, \boldsymbol{\gamma}).$$

Buď $\boldsymbol{u} \in \mathbb{R}^{p-1}, v \in \mathbb{R}$ pevné. Z předpokladu, že $\boldsymbol{u}^\top \boldsymbol{X}_i - v \neq 0$ dostáváme, že $\boldsymbol{u}^\top \boldsymbol{u} + v^2 > 0$, a můžeme tedy definovat $\boldsymbol{\gamma} \in S^p$ jako

$$\boldsymbol{\gamma} := \frac{(-v, \boldsymbol{u}^\top)^\top}{\sqrt{\boldsymbol{u}^\top \boldsymbol{u} + v^2}}.$$

Naopak, buď $\boldsymbol{\gamma} \in S^p$ pevné, pak definujme $\boldsymbol{u} \in \mathbb{R}^{p-1}, v \in \mathbb{R}$ jako

$$\boldsymbol{u} := (\gamma_2, \ldots, \gamma_p)^\top,\ v := -\gamma_1.$$

Tím jsme dokázali, že

$$\inf_{\boldsymbol{u},\, v} \min\{d(\boldsymbol{\beta}, \boldsymbol{u}, v),\, c(\boldsymbol{u}, v)\} = \inf_{\boldsymbol{\gamma}} a(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

což zároveň dokazuje ekvivalenci rovnic (5.8) a (5.19).

□

Lemma 38. *Regresní hloubku založenou na distribuční funkci z definice 28 lze za předpokladu $H \in \mathcal{H}^C$ psát ve tvaru:*

$$\mathit{rdepth}(\boldsymbol{\beta}, H) = \frac{1}{2} + \frac{1}{2} \inf_{\boldsymbol{\gamma} \in S^p} \mathsf{E}_{\boldsymbol{Z}} \operatorname{sgn}(U(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma}), \tag{5.24}$$

kde $U(\boldsymbol{\beta}) := Y - \widetilde{\boldsymbol{X}}^\top \boldsymbol{\beta}$.

Důkaz (vlastní).

$$\begin{aligned}
&\mathit{rdepth}(\boldsymbol{\beta}, H) \\
&= \inf_{\boldsymbol{u},v} P_H(\operatorname{sgn}(U(\boldsymbol{\beta})) = \operatorname{sgn}(\boldsymbol{X}^\top \boldsymbol{u} - v)) \\
&= \inf_{\boldsymbol{\gamma} \in S^p} P_H(\operatorname{sgn}(U(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma})) \\
&= \frac{1}{2} + \frac{1}{2} \inf_{\boldsymbol{\gamma} \in S^p} (P_H(\operatorname{sgn}(U(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma})) - 1 + P_H(\operatorname{sgn}(U(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma}))) \\
&= \frac{1}{2} + \frac{1}{2} \inf_{\boldsymbol{\gamma} \in S^p} (P_H(\operatorname{sgn}(U(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma})) - P_H(\operatorname{sgn}(U(\boldsymbol{\beta})) = -\operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma}))) \\
&= \frac{1}{2} + \frac{1}{2} \inf_{\boldsymbol{\gamma} \in S^p} (\mathsf{E}_{\boldsymbol{Z}}\, I(\operatorname{sgn}(U(\boldsymbol{\beta})) = \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma})) - \mathsf{E}_{\boldsymbol{Z}}\, I(\operatorname{sgn}(U(\boldsymbol{\beta})) = -\operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma})) \\
&= \frac{1}{2} + \frac{1}{2} \inf_{\boldsymbol{\gamma} \in S^p} \mathsf{E}_{\boldsymbol{Z}} \operatorname{sgn}(U(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma}),
\end{aligned}$$

kde přechod od parametrů $\boldsymbol{u} \in \mathbb{R}^{p-1}, v \in \mathbb{R}$ k $\boldsymbol{\gamma} \in S^p$ proběhne analogicky jako u lemmatu 37.

□

Protože z předpokladu $H \in \mathcal{H}^C$ také vyplývá, že pozorování $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ jsou *s.j.* v obecné pozici, dostáváme, že

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \left| \frac{1}{n} \sum_{i=1}^{n} I(U_i(\boldsymbol{\beta}) = 0) \right| \overset{s.j.}{\leq} \frac{p}{n} \overset{n \to \infty}{\longrightarrow} 0,$$

z čehož vyplývá, že pomocí lemmat 37 a 38 lze rovnici (5.18) přepsat do tvaru

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \left| \inf_{\boldsymbol{\gamma} \in S^p} \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(U_i(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\gamma}) - \inf_{\boldsymbol{\gamma} \in S^p} \mathsf{E}_{\,Z} \operatorname{sgn}(U(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma}) \right| \overset{s.j.}{\longrightarrow} 0.$$

K důkazu konvergence (5.18) tedy bude stačit ukázat, že platí

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in S^p} \left| \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(U_i(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\gamma}) - \mathsf{E}_{\,Z} \operatorname{sgn}(U(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{\gamma}) \right| \overset{s.j.}{\longrightarrow} 0. \quad (5.25)$$

Poznámka. Následující tvrzení bylo i s podmínkami převzato z článku Bai a He (1999). Je zde však jeden problém, a to ten, že tvrzení 39 bylo v původním článku dokázáno **pouze** pro fixní hodnoty regresorů. Autoři článku pouze poznamenali, že „*vše by proběhlo stejně ve smyslu s.j., pokud bychom uvažovali regresory náhodné*“. Bohužel, jak již bylo dříve zmíněno, náročnost tohoto důkazu značně přesahuje úroveň této práce, a proto jej nebylo možné ověřit. Určitě však doporučujeme důkaz tvrzení 39 podrobněji prozkoumat a ověřit, zda z něho skutečně vyplývá rovnice (5.25) i při náhodných regresorech.

Předpoklad 4. *BÚNO předpokládejme, že $\boldsymbol{\beta}^{MED} := \mathbf{0}_p$ (viz regresní invariance tvrzení 40).* ***Předpoklady pro konzistenci*** *odhadu $\hat{\boldsymbol{\beta}}^{RD}$ budeme pak rozumět následující podmínky:*

(1) $\exists\, 0 < b < \infty \colon \max_{1 \leq i \leq n} ||\widetilde{\boldsymbol{X}}_i||_2 = \mathcal{O}(n^b)$ *s.j.,*

(2) $\forall\, \{a_n\}_{n=1}^{\infty} \colon a_n \longrightarrow 0 \implies Q_n(a_n) \overset{s.j.}{\longrightarrow} 1,$

(3) $\exists A < \infty \colon \begin{cases} n^{-1} \sum_{i=1}^{n} \{1 - F_i(n^A) + F_i(-n^A)\} \overset{s.j.}{\longrightarrow} 0 \\ \max_{1 \leq i \leq n} \sup_{y \in \mathbb{R}} (F_i(y + n^{-A}) - F_i(y - n^{-A})) \overset{s.j.}{\longrightarrow} 0, \end{cases}$

(4) $\forall\, r > 0 \colon \eta(r) = \inf_{1 \leq i \leq n} \min\{|1 - 2F_i(-r)|, |1 - 2F_i(r)|\} \overset{s.j.}{>} 0$

kde F_i značí distribuční funkci podmíněné náhodné veličiny $Y|\boldsymbol{X} = \boldsymbol{X}_i$ a

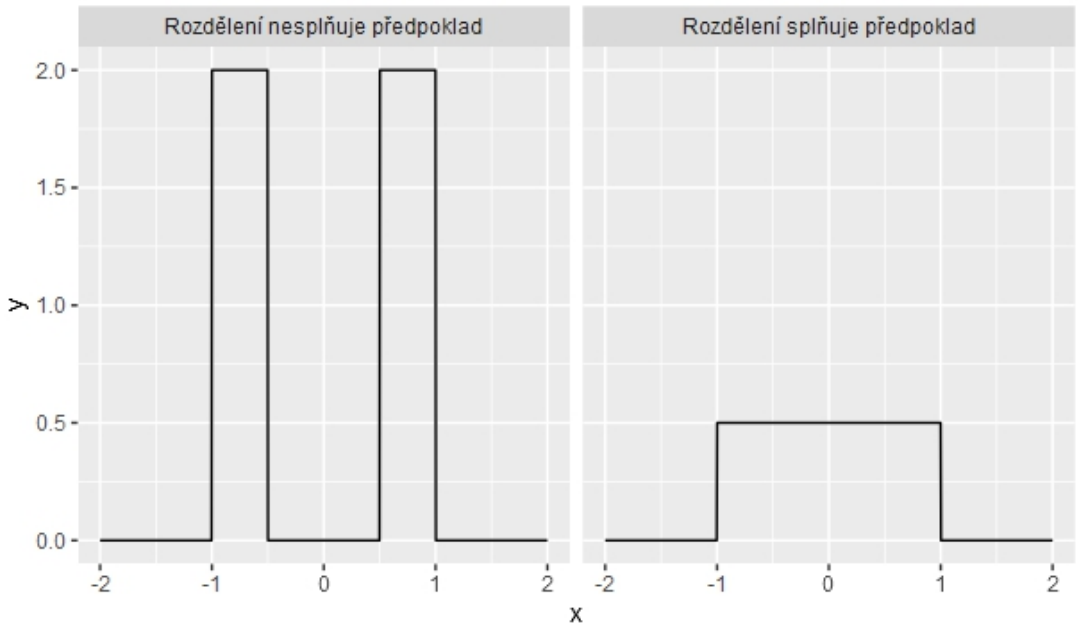
$$Q_n(c) := \inf_{\boldsymbol{l} \in S^p} n^{-1} \sum_{i=1}^{n} I(|\boldsymbol{w}_i^\top \boldsymbol{l}| > c).$$

Poznámka. Podmínka *(4)* nám říká, že všechna podmíněná rozdělení F_i mají okolo 0 (tj. okolo mediánu) „dostatečný“ nosič (viz obrázek 5.9).

Tvrzení 39. *Nechť $H \in \mathcal{H}^C$, a zároveň je splněn předpoklad linearity v mediánu spolu s podmínkami 1 až 3 z předpokladů konzistence 4, pak platí, že*

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in S^p} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\operatorname{sgn}(U_i(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\gamma}) - \mathsf{E}_{\,\boldsymbol{Z}} \operatorname{sgn}(U_i(\boldsymbol{\beta})) \operatorname{sgn}(\widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\gamma}) \right) \right| \overset{s.j.}{\longrightarrow} 0,$$

a pokud je navíc splněna i podmínka (4), pak i $\hat{\boldsymbol{\beta}}^{RD} \overset{s.j.}{\longrightarrow} \boldsymbol{\beta}^{RD}$:



Obrázek 5.9: Grafická ilustrace podmínky *(4)*. Na obrázku vlevo můžeme vidět hustotu podmíněného rozdělení, které předpoklad nesplňuje, na obrázku vpravo pak rozdělení, které předpoklad splňuje vždy.

5.5.3 Eficience

Odhad RD nemá bohužel na rozdíl od většiny používaných odhadů (M-odhady, Z-odhady, atd.) asymptoticky normální rozdělení (viz Bai a He (1999)). Odvození skutečného asymptotické rozdělení je sice možné nalézt v již zmíněném článku, odvození však relativně komplikované, a opět tak značně přesahuje úroveň této práce (a nebude zde proto uvedeno).

Kromě toho bylo provedeno několik simulací porovnávajících hodnoty rozptylu odhadů regresních koeficientů metodou RD a LAD pro dvourozměrné normální rozdělení (Van Aelst a Rousseeuw, 2000). Jak pro odhad absolutního členu, tak pro odhad směrnice regresní přímky dává metoda RD horší výsledky. Relativní eficience jí získaných odhadů vychází v porovnání s metodou LAD cca 82 % pro absolutní člen a cca 87 % pro směrnici regresní přímky. I přesto jsou však tyto hodnoty mnohem lepší, než se z počátku usuzovalo. Ze předpokladu, že by totiž rozdělení odhadu RD bylo stejně jako rozdělení LAD odhadu asymptoticky normální, by hodnota relativní eficience vycházela ještě menší, a to konkrétně pouze cca 75 % pro směrnici regresní přímky (hodnoty pro absolutní člen nebyly v článku uvedeny).

Pozorování. Rozdílné výsledky jsou způsobeny tím, že skutečné limitní rozdělení RD odhadu se zdá mít kratší chvosty, což vysvětluje jeho menší rozptyl (Van Aelst a Rousseeuw, 2000).

Poznámka. Je nutné podotknout, že výpočty eficience se provádějí na *správných* datech. V případě, kdy pozorování pochází ze správného rozdělení, bude tedy metoda LAD vždy lepší volbou. Hlavní cíl regresních metod je ale konstruovat odhady, které dávají smysluplné výsledky i při *kontaminovaných* datech (viz sekce 2.2). Na rozdíl od metody LAD metoda RD disponuje BP v hodnotě 1/3

(viz tvrzení 42 níže), pokud bychom tedy pracovali s *kontaminovaným* datovým souborem, pak nás relativně malý pokles v eficienci může ochránit před celkovým znehodnocením našeho odhadu.

5.6 Praktické vlastnosti odhadu RD

Následující tvrzení 40 dokazuje vlastnosti invariance *pouze* pro odhad RD, a to z důvodu zachování konzistentního značení v celé práci. Důkaz invariančních vlastností pro regresní hloubku jakožto funkci a parametr $\boldsymbol{\beta}^{RD}$ by byl obdobný.

5.6.1 Transformace

Tvrzení 40. *Odhad $\hat{\boldsymbol{\beta}}^{RD}$ je regresně, škálově i afinně invariantní.*

Důkaz (vlastní). Buď $\boldsymbol{h} \in \mathbb{R}^p$ libovolný p-rozměrný vektor, pak

$$\begin{aligned} rdepth(\boldsymbol{\beta}, (\boldsymbol{Y} + \widetilde{\mathbb{X}}\boldsymbol{h}, \widetilde{\mathbb{X}})) &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big((Y_i + \widetilde{\boldsymbol{X}}_i^\top h - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta})(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big) \\ &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big((Y_i - \widetilde{\boldsymbol{X}}_i^\top (\boldsymbol{\beta} - \boldsymbol{h}))(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big) \\ &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big((Y_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}^*)(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big) \end{aligned}$$

kde $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{h}$, tj. odhad $\hat{\boldsymbol{\beta}}^{RD}$ je *regresně* invariantní dle definice 5.

Buď $c \in \mathbb{R}$, $c \neq 0$ libovolná konstanta, pak

$$\begin{aligned} rdepth(\boldsymbol{\beta}, (c\boldsymbol{Y}, \widetilde{\mathbb{X}})) &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big((cY_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta})(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big) \\ &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big(\operatorname{sgn}(c)(Y_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}/c)(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big) \\ &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big(\operatorname{sgn}(c)(Y_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}^*)(\boldsymbol{u}^\top \boldsymbol{X}_i - v) \geq 0\Big) \end{aligned}$$

kde $\boldsymbol{\beta} = c\boldsymbol{\beta}^*$. Protože volbou $\boldsymbol{u}^* := -\boldsymbol{u}, v^* := -v$ můžeme vždy získat členy s opačnou nerovností, na znaménku c regresní hloubka nezávisí, a platí tedy, že předchozí rovnici lze psát ve tvaru

$$rdepth(\boldsymbol{\beta}, (c\boldsymbol{Y}, \widetilde{\mathbb{X}})) = \inf_{\boldsymbol{u}^*,\, v^*} \sum_{i=1}^{n} I\Big((Y_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}^*)(\boldsymbol{u}^{*\top} \boldsymbol{X}_i - v^*) \geq 0\Big),$$

tj. odhad $\hat{\boldsymbol{\beta}}^{RD}$ je *škálově* invariantní dle definice 6.

Buď $\mathbb{A} \in \mathbb{R}^{p \times p}$ je libovolná regulární matice, z vlastností skalárního součinu vyplývá, že můžeme psát

$$\boldsymbol{u}^\top \boldsymbol{X}_i - v = -v + \boldsymbol{X}_i^\top \boldsymbol{u} = (1, \boldsymbol{X}_i^\top)(-v, \boldsymbol{u}^\top)^\top,$$

a tedy

$$\begin{aligned} rdepth(\boldsymbol{\beta},\, (\boldsymbol{Y}, \tilde{\mathbb{X}}\mathbb{A})) &= \inf_{\boldsymbol{u},\, v} \sum_{i=1}^{n} I\Big((Y_i - \widetilde{\boldsymbol{X}}_i^\top \mathbb{A}\boldsymbol{\beta})(1, \boldsymbol{X}_i^\top)\mathbb{A}(-v, \boldsymbol{u}^\top)^\top \geq 0\Big) \\ &= \inf_{\boldsymbol{u}^*,\, v^*} \sum_{i=1}^{n} I\Big((Y_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}^*)(1, \boldsymbol{X}_i^\top)(-v^*, \boldsymbol{u}^{*\top})^\top \geq 0\Big) \\ &= \inf_{\boldsymbol{u}^*,\, v^*} \sum_{i=1}^{n} I\Big((Y_i - \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}^*)(\boldsymbol{u}^{*\top}\boldsymbol{X}_i - v^*) \geq 0\Big) \end{aligned}$$

kde $\boldsymbol{\beta} = \mathbb{A}^{-1}\boldsymbol{\beta}^*, \boldsymbol{u} = \mathbb{A}^{-1}\boldsymbol{u}^*, v = \mathbb{A}^{-1}v^*$, a protože $\boldsymbol{u}^*$ (resp. v^*) nabývá stejně jako $\boldsymbol{u}$ (resp. v) libovolných hodnot z $\mathbb{R}^{p-1}$ (resp. $\mathbb{R}$), je možné substituci zanedbat. Tj. odhad $\hat{\boldsymbol{\beta}}^{RD}$ je *afinně* invariantní dle definice 7.

□

5.6.2 Robustnost

Jedním z hlavních důvodů pro zavedení odhadu RD jsou jeho dobré robustní vlastnosti, proto se na ně nyní blíže podíváme. Následující tvrzení je i s důkazem převzato z článku Van Aelst a kol. (2002):

Breakdown-point

Tvrzení 41. *Předpokládejme, že $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ jsou v obecné pozici, pak hodnota breakdown-pointu definovaného na základě přidávání nových pozorování (viz rovnice* (2.5)*) splňuje následující nerovnost:*

$$\varepsilon_n^*(\hat{\boldsymbol{\beta}}^{RD}, \mathbb{Z}) \geq \frac{n - p^2 + 1}{n(p+1) - p^2 + 1} \overset{n\to\infty}{\approx} \frac{1}{p+1}. \tag{5.26}$$

Důkaz. Odhad regresního mediánu jsme si definovali jako průměr těch nadrovin s maximální hloubkou, které procházejí alespoň p pozorováními a jsou těmito pozorováními již jednoznačně určeny. (Z předpokladu obecné pozice $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ víme, že tyto nadroviny nejsou rovnoběžné s osou y, a je tedy možné je jako kandidáty uvažovat). Jejich počet je tedy vždy ohraničen hodnotou $\binom{n}{p}$, a pokud chceme, aby se odhad stal „zcela bezcenným“, je nutné, aby se do této množiny dostal alespoň jeden „nesmyslný“ kandidát (viz sekce 2.2).

Uvažujme definici breakdown-pointu pomocí přidávání nových pozorování (viz rovnice (2.5)) a přidejme k našemu náhodnému nejmenší možný počet m pozorování tak, aby existoval vektor $\boldsymbol{\eta} \in \mathbb{R}^p$, který

1. prochází alespoň p pozorováními (a je jimi tedy jednoznačně určený),
2. $rdepth(\boldsymbol{\eta}, \mathbb{Z}_{n+m}) = l^*$, kde $l^* = \max_{\boldsymbol{\beta}\in\mathbb{R}^p} rdepth(\boldsymbol{\beta}, \mathbb{Z}_{n+m})$, a zároveň
3. $rdepth(\boldsymbol{\eta}, \mathbb{Z}) \leq p - 1$.

V takovém případě již máme totiž zaručeno, že v původním výběru $\mathbb{Z}$ vektor $\boldsymbol{\eta}$ mezi kandidáty nepatřil, a bude se tak jednat o našeho „nesmyslného“ kandidáta

v novém výběru. Na základě článku Amenta a kol. (2000) pak víme, že je splněna nerovnost $rdepth(\boldsymbol{\eta}, \mathbb{Z}_{n+m}) \geq \lceil (n+m)/(p+1) \rceil$, což dohromady s podmínkou $rdepth(\boldsymbol{\eta}, \mathbb{Z}) \leq p-1$ dává nerovnost

$$\left\lceil \frac{n+m}{p+1} \right\rceil \leq rdepth(\boldsymbol{\eta}, \mathbb{Z}_{n+m}) \leq p-1+m.$$

Po úpravě této nerovnosti dostáváme nerovnost $m \geq (n-p^2+1)/p$, z čehož po dosazení do rovnice (2.5) již plyne dokazovaná dolní mez

$$\varepsilon_n^*(\hat{\boldsymbol{\beta}}^{RD}, \mathbb{Z}) \geq \frac{m}{n+m} \geq \frac{n-p^2+1}{n(p+1)-p^2+1}.$$

□

Poznámka. Tvrzení 41 nemá na rozdíl od předchozích tvrzení žádné předpoklady, nevyžadujeme absolutně spojité rozdělení pozorování, ani nedegenerovaný datový soubor. Jedná se tedy o dolní mez při „nejhorší možné variantě". Pokud je ale v takovém případě počet regresorů blízký počtu pozorování, odhad RD se příliš neliší od nerobustních odhadů.

Pokud ale budeme opět předpokládat, že $H \in \mathcal{H}^C$ s hustotou $h > 0$, pak je již odhad RD robustní. Důkaz tohoto tvrzení je možné nalézt např. v článku Van Aelst a kol. (2002):

Tvrzení 42. *Nechť $H \in \mathcal{H}^C$ s hustotou $h > 0$, a zároveň je splněn předpoklad linearity v mediánu, pak platí*

$$\varepsilon_n^*(\hat{\boldsymbol{\beta}}^{RD}, \mathbb{Z}) \xrightarrow{s.j.} \frac{1}{3}.$$

kde $\varepsilon_n^*(\hat{\boldsymbol{\beta}}^{RD}, \mathbb{Z})$ je hodnota breakdown-pointu získaného na základě přidávání nových pozorování (viz rovnice (2.5))

Influenční funkce

Uvažujme libovolný funkcionál $\boldsymbol{T} = \boldsymbol{T}(H) = \boldsymbol{T}(\boldsymbol{Z}) = \boldsymbol{T}((Y, \boldsymbol{X}^\top)^\top)$. Z konstrukce influenční funkce z definice 10 pak vyplývá, že pro libovolné $\boldsymbol{T} = (T_1, \boldsymbol{T}_2^\top)^\top$, kde $T_1 \in \mathbb{R}$, $\boldsymbol{T}_2 \in \mathbb{R}^{p-1}$ platí

$$\mathsf{IF}(\boldsymbol{z}, \boldsymbol{T}, H) = (\mathsf{IF}(\boldsymbol{z}, T_1, H), \mathsf{IF}(\boldsymbol{z}, \boldsymbol{T}_2, H)^\top)^\top.$$

Nyní předpokládejme, že $H = H_{\boldsymbol{\mu},\Sigma} \in \mathcal{H}^C$ je distribuční funkce s *elipticky symetrickým* rozdělení s hustotou $h_{\boldsymbol{\mu},\Sigma}$, kterou je možné vyjádřit ve tvaru

$$h_{\boldsymbol{\mu},\Sigma}(y, \boldsymbol{x}) = \frac{g(((y, \boldsymbol{x}^\top)^\top - \boldsymbol{\mu})^\top \Sigma^{-1} ((y, \boldsymbol{x}^\top)^\top - \boldsymbol{\mu}))}{\sqrt{\det(\Sigma)}}, \tag{5.27}$$

kde $\boldsymbol{\mu} \in \mathbb{R}^p$ je vektor polohy, $\Sigma \in \mathbb{R}^{p \times p}$ je pozitivně definitní matice a funkce g má striktně negativní derivaci (z čehož vyplývá, že $H_{\boldsymbol{\mu},\Sigma}$ je unimodální). Po malé

úpravě Choleského dekompozice (viz Příloha 1) dostáváme, že matici Σ lze po řadě rozložit na horní a dolní trojúhelníkovou matici, tj. $\Sigma = UU^\top$, kde

$$U = \begin{pmatrix} c & \boldsymbol{v}^\top \\ \mathbf{0}_p & A \end{pmatrix}, \tag{5.28}$$

$A \in \mathbb{R}^{(p-1)\times(p-1)}$, $\det(A) \neq 0$ je horní trojúhelníková matice, $c \in \mathbb{R}\backslash\{0\}$, $\boldsymbol{v} \in \mathbb{R}^{p-1}$. Dále uvažujme transformovaný náhodný vektor

$$(\widetilde{Y}, \widetilde{\boldsymbol{X}}^\top)^\top := U^{-1}((Y, \boldsymbol{X}^\top)^\top - \boldsymbol{\mu}),$$

kde rozdělení tohoto vektoru je dle rovnice (5.27) určeno distribuční funkcí $H_{\mathbf{0},I}$, tj. $\boldsymbol{\mu} = \mathbf{0}_p$ a $\Sigma = I_p$ je jednotková matice. Rovněž zřejmě platí, že

$$(Y, \boldsymbol{X}^\top)^\top = U(\widetilde{Y}, \widetilde{\boldsymbol{X}}^\top)^\top + \boldsymbol{\mu},$$

a pokud si označíme $\boldsymbol{\mu} = (\mu_Y, \boldsymbol{\mu}_X^\top)^\top$, můžeme psát

$$\begin{aligned} Y &= c\widetilde{Y} + \boldsymbol{v}^\top \widetilde{\boldsymbol{X}} + \mu_Y \\ \boldsymbol{X} &= A\widetilde{\boldsymbol{X}} + \boldsymbol{\mu}_X. \end{aligned} \tag{5.29}$$

Pozorování. Vektor $\boldsymbol{X}$ v rovnici (5.29) vznikl pomocí *afinní* transformace $\widetilde{\boldsymbol{X}}$, zatímco Y vznikl pomocí *škálování* $\widetilde{Y}$ a následné aplikace *regresní* transformace.

Poznámka. Pokud bychom navíc předpokládali, že $(Y, \boldsymbol{X}^\top) \in \mathcal{L}_2$,[4] pak $\boldsymbol{\mu}$ je vektor střední hodnoty a Σ je kladným násobkem rozptylové matice.

Tvrzení 43. *Nechť $\boldsymbol{T}$ je regresně, afinně a škálově invariantní funkcionál, pak za předpokladu, že $H_{\boldsymbol{\mu},\Sigma} \in \mathcal{H}^C$ je distribuční funkce eliptického rozdělení, je influenční funkce $\boldsymbol{T}$ v $H_{\boldsymbol{\mu},\Sigma}$ již jednoznačně určena jeho influenční funkcí v $H_{\mathbf{0},I}$, a to pomocí následujících rovností:*

$$\begin{aligned} \mathsf{IF}((y, \boldsymbol{x}^\top)^\top, \boldsymbol{T}_2, H_{\boldsymbol{\mu},\Sigma}) &= cA^{-\top}\, \mathsf{IF}((\tilde{y},\, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{T}_2, H_{\mathbf{0},I}) + A^{-\top}\boldsymbol{v} \\ \mathsf{IF}((y, \boldsymbol{x}^\top)^\top, T_1, H_{\boldsymbol{\mu},\Sigma}) &= c\, \mathsf{IF}((\tilde{y},\, \tilde{\boldsymbol{x}}^\top)^\top, T_1, H_{\mathbf{0},I}) - c\boldsymbol{\mu}_X^\top A^{-\top}\, \mathsf{IF}((\tilde{y},\, \tilde{\boldsymbol{x}}^\top), \boldsymbol{T}_2, H_{\mathbf{0},I}) \\ &\quad - \boldsymbol{\mu}_X^\top A^{-\top}\boldsymbol{v} + \mu_Y, \end{aligned}$$

kde $A^{-\top} := (A^{-1})^\top$ a $A \in \mathbb{R}^{(p-1)\times(p-1)}, \boldsymbol{v} \in \mathbb{R}^{p-1}, c \in \mathbb{R}$ pochází z upravené Choleského dekompozice matice Σ z rovnice (5.28).

Důkaz. Z rovnice (5.29) vyplývá, že

$$\boldsymbol{T}((Y, \boldsymbol{X}^\top)) = \boldsymbol{T}((c\widetilde{Y} + \boldsymbol{v}^\top \widetilde{\boldsymbol{X}} + \mu_Y, (A\widetilde{\boldsymbol{X}} + \boldsymbol{\mu}_X)^\top)),$$

a stejně tak tedy

$$\boldsymbol{T}((Y, (1, \boldsymbol{X}^\top))) = \boldsymbol{T}((c\widetilde{Y} + \boldsymbol{v}^\top \widetilde{\boldsymbol{X}} + \mu_Y), (1, (A\widetilde{\boldsymbol{X}} + \boldsymbol{\mu}_X)^\top))).$$

V takovém případě však můžeme psát $(1, (A\widetilde{\boldsymbol{X}} + \boldsymbol{\mu}_X)^\top) = (1, \widetilde{\boldsymbol{X}}^\top)B$, kde

$$B = \begin{pmatrix} 1 & \boldsymbol{\mu}_X^\top \\ \mathbf{0}_p & A^\top \end{pmatrix}.$$

[4] kde $\mathcal{L}_2$ značí množinu náhodných vektorů jejichž rozdělení má konečné druhé momenty

Z afinní invariance dle definice 7 tedy vyplývá, že platí

$$\boldsymbol{T}((Y,(1,\widetilde{\boldsymbol{X}}^\top)B)) = B^{-1}\,\boldsymbol{T}((Y,(1,\widetilde{\boldsymbol{X}}^\top))),$$

a z regresní a škálové invariance dále vyplývá (dle definic 5 a 6), že můžeme psát

$$\begin{aligned}
B^{-1}\boldsymbol{T}(((1,\widetilde{\boldsymbol{X}}^\top),Y)) &= B^{-1}\,\boldsymbol{T}((c\widetilde{Y}+\boldsymbol{v}^\top\widetilde{\boldsymbol{X}}+\mu_Y,(1,\widetilde{\boldsymbol{X}}^\top))) \\
&= B^{-1}\,\boldsymbol{T}((c\widetilde{Y}+(1,\widetilde{\boldsymbol{X}}^\top)(\mu_Y,\boldsymbol{v}^\top)^\top,(1,\widetilde{\boldsymbol{X}}^\top))) \\
&= B^{-1}\boldsymbol{T}((c\widetilde{Y},(1,\widetilde{\boldsymbol{X}}^\top))) + B^{-1}(\mu_Y,\boldsymbol{v}^\top)^\top \\
&= B^{-1}c\,\boldsymbol{T}((\widetilde{Y},(1,\widetilde{\boldsymbol{X}}^\top))) + B^{-1}(\mu_Y,\boldsymbol{v}^\top)^\top.
\end{aligned}$$

Konečně tedy, pokud si předchozí rovnici rozepíšeme pomocí inverzní matice

$$B^{-1} = \begin{pmatrix} 1 & -\boldsymbol{\mu}_X^\top A^{-\top} \\ \mathbf{0}_p & A^{-\top} \end{pmatrix}$$

a samostatně si vyjádříme T_1 a $\boldsymbol{T}_2$, dostáváme, že

$$\begin{aligned}
\boldsymbol{T}_2((Y,\boldsymbol{X}^\top)) &= cA^{-\top}\boldsymbol{T}_2((\widetilde{Y},\widetilde{\boldsymbol{X}}^\top)) + A^{-\top}\boldsymbol{v} \\
T_1((Y,\boldsymbol{X}^\top)) &= c\,T_1((\widetilde{Y},\widetilde{\boldsymbol{X}}^\top)) - c\boldsymbol{\mu}_X^\top A^{-\top}\boldsymbol{T}_2((\widetilde{Y},\widetilde{\boldsymbol{X}}^\top)) - \boldsymbol{\mu}_X^\top A^{-\top}\boldsymbol{v} + \mu_Y.
\end{aligned}$$

Dohromady s definicí influenční funkce z rovnice (2.8) již dostáváme požadované tvrzení.

□

Protože regresní medián je dle tvrzení 40 regresně, škálově i afinně invariantní, je možné využít předchozího tvrzení k určení jeho influenční funkce pro všechna eliptická rozdělení pouze na základě odvození influenční funkce pro $H_{\mathbf{0},I}$. Důkaz následujícího tvrzení nalezneme v práci Van Aelst a Rousseeuw (2000):

Věta 44. *Nechť $p = 2$ a $H \in \mathcal{H}^C$ je elipticky symetrické rozdělení, pak influenční funkci regresního mediánu $\boldsymbol{\beta}^{RD}$ pro $H = H_{\mathbf{0},I}$ lze po složkách zapsat následujícím způsobem:*

- *pro absolutní člen:*

$$\begin{aligned}
\mathsf{IF}((y,x),\beta_0^{RD},H) = {} & 1/2\,\frac{\operatorname{sgn}(y)}{h_Y(0)} \\
& \left(\frac{I(H_{X|Y=0}(|x|)\le 2/3)}{H_{X|Y=0}(|x|)} + \frac{I(H_{X|Y=0}(|x|)\ge 2/3)}{2(2H_{X|Y=0}(|x|)-1)}\right),
\end{aligned} \tag{5.30}$$

- *pro směrnici regresní přímky:*

$$\begin{aligned}
\mathsf{IF}((y,x),\beta_1^{RD},H) = {} & \operatorname{sgn}(x)\operatorname{sgn}(y) \\
& \left(\frac{I(G(|x|)\le 2G(\infty)/3)}{4(G(\infty)-G(|x|))} + \frac{I(G(|x|)\ge 2G(\infty)/3)}{2G(\infty)-G(|x|)}\right)
\end{aligned} \tag{5.31}$$

kde právě jeden z indikátorů v závorce je vždy nenulový, $G(t) := \int_0^{t^2} g(u)du$, $G(\infty) := \lim_{t\to\infty} G(t)$, $h_Y(0)$ *je marginální hustota* Y *v bodě 0 a* $H_{X|Y=0}(|x|)$ *je podmíněná distribuční funkce* X *za podmínky* $Y = 0$ *v bodě* $|x|$.

Důsledek 2. *Buď* $H = \mathcal{N}_2(\mathbf{0}_2, I_2)$, *pak influenční funkci regresního mediánu* $\boldsymbol{\beta}^{RD}$ *lze po částech zapsat následujícím způsobem:*

- *pro absolutní člen:*

$$\mathsf{IF}((x,y), \beta_0^{RD}, H) = 1/2\, \frac{\operatorname{sgn}(y)}{\phi(0)} \left(\frac{I(|x| \le \Phi^{-1}(2/3))}{\Phi(|x|)} + \frac{I(|x| > \Phi^{-1}(2/3))}{2(2\Phi(|x|) - 1)} \right),$$

- *pro směrnici regresní přímky:*

$$\mathsf{IF}((x,y), \beta_1^{RD}, H) = 1/2\, \frac{\operatorname{sgn}(x)\operatorname{sgn}(y)}{\phi(0)} \left(\frac{I(\phi(x) \ge \phi(0)/3)}{4\phi(x)} + \frac{I(\phi(x) < \phi(0)/3)}{\phi(0) + \phi(x)} \right),$$

kde ϕ *je hustota a* Φ *je distribuční funkce normálního rozdělení* $\mathcal{N}(0,1)$.

Důkaz. Za předpokladu $(Y, X)^\top \sim \mathcal{N}_2(\mathbf{0}_2, I_2)$ jsou Y a X nezávislé náhodné veličiny, a tedy $H_{X|Y=0} = H_X$. Z předpokladu normality pak dostáváme, že $H_X = \Phi$ a $h_Y(0) = \phi(0)$. V neposlední řadě po dosazení za $\boldsymbol{\mu} = (0,0)^\top, \Sigma = I_2$ dostáváme i vyjádření funkce g ve tvaru

$$g(u) := g(x^2 + y^2) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2} \right),$$

a tedy po substituci $u := x^2 + y^2$ získáváme vyjádření

$$G(t) = \frac{1}{2\pi} \int_0^{t^2} \exp(-u/2) du = 2\phi(0)(\phi(0) - \phi(t)).$$

Ze symetrie hustoty normálního rozdělení dostáváme, že $\phi(t) = \phi(-t)$, a tedy

$$G(|t|) = 2\phi(0)(\phi(0) - \phi(t)).$$

Požadované výrazy pak dostaneme po dosazení do rovnic (5.30) a (5.31).

□

Pozorování. Influenční funkce regresního mediánu $\boldsymbol{\beta}^{RD}$ je pro $H = \mathcal{N}_2(\mathbf{0}_2, I_2)$ **omezená**, a to jak v x, tak v y, následujícím způsobem:

- pro absolutní člen:

$$|\,\mathsf{IF}\,((y,x), \beta_0^{RD}, H)| \le 1/2\, \frac{1}{\phi(0)} \max\{2, 3/2\} = \frac{1}{\phi(0)} \approx 2.51,$$

- pro směrnici regresní přímky:

$$|\,\mathsf{IF}\,((y,x),\beta_1^{RD},H)| \le 1/2\,\frac{1}{\phi(0)}\max\Big\{\frac{3}{4\phi(0)},\frac{1}{\phi(0)}\Big\} = 1/2\,\frac{1}{\phi(0)^2} \approx 3.14.$$

Pro obecné, elipticky symetrické rozdělení H pak vycházíme ze symetrie rozdělení $H_{X|Y=0}$ a z nerovnosti $H_{X|Y=0}(|x|) \ge H_{X|Y=0}(0) = 1/2$, z čehož dostáváme pro influenční funkci následující omezení:

- pro absolutní člen:

$$|\,\mathsf{IF}\,((y,x),\beta_0^{RD},H)| \le 1/2\,\frac{1}{h_Y(0)}\max\{2,3/2\} = \frac{1}{h_Y(0)},$$

- pro směrnici regresní přímky (protože funkce G je kladná a rostoucí):

$$|\,\mathsf{IF}\,((y,x),\beta_1^{RD},H)| \le \max\Big\{\frac{3}{4G(\infty)},\frac{1}{G(\infty)}\Big\} = \frac{1}{G(\infty)}.$$

5.7 Výpočetní aspekty*

Algoritmus výpočtu regresní hloubky

Uvažujme situaci v dimenzi $p = 2$. Pak pro daný vektor koeficientů $\boldsymbol{\beta} \in \mathbb{R}^2$ spočteme regresní hloubku následujícím způsobem:

1. Z předpokladu ordinálního rozdělení regresorů víme, že je můžeme uspořádat tak, aby $X_{[1]} \le \cdots \le X_{[n]}$. V případě, že mezi regresory existují shody, budeme tyto shody ignorovat, tj. budeme uvažujme jejich podvýběr $X^*_{[1]} < \cdots < X^*_{[n^*]}$, kde $2 \le n^* \le n$.

2. Označme si nyní

$$v_1 := X^*_{[1]} - 1, \quad v_i := \frac{X^*_{[i]} - X^*_{[i-1]}}{2} \quad \text{pro} \quad i \in \{2,\dots,n^*\},$$

kde $x = v_i$ budou představovat pomocné dělící nadroviny. Regresní hloubku parametru $\boldsymbol{\beta}$ vzhledem k náhodnému výběru $\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n$ pak spočteme jako

$$\mathit{rdepth}(\boldsymbol{\beta},\mathbb{Z}) = \min_{1\le i\le n^*}(\min\{L^-(v_i)+R^+(v_i),\, L^+(v_i)+R^-(v_i)\}),$$

kde

$$L^-(v) := \sum_{i=1}^n I(U_i(\boldsymbol{\beta}) \le 0)I(X_i < v),\; R^+(v) := \sum_{i=1}^n I(U_i(\boldsymbol{\beta}) \ge 0)I(X_i > v)$$

a

$$L^+(v) := \sum_{i=1}^n I(U_i(\boldsymbol{\beta}) \ge 0)I(X_i < v),\; R^-(v) := \sum_{i=1}^n I(U_i(\boldsymbol{\beta}) \le 0)I(X_i > v).$$

Maximální výpočetní složitost části 1 je $\mathcal{O}(n\log n)$ a výpočetní složitost části 2 je lineární, tj. $\mathcal{O}(n)$, celková výpočetní složitost regresní hloubky pro konkrétní $\boldsymbol{\beta}$ je tedy $\mathcal{O}(n\log n)$.

Pozorování. Bohužel se ukazuje, že výpočetní složitost algoritmu roste n-násobně s každým dalším (netriviálním) regresorem, tj. v případě $p-1$ regresorů je $O(n^{p-1}\log n)$ (Rousseeuw a Hubert, 1999). Aproximativní algoritmus s výpočetní složitostí $\mathcal{O}(np + n\log n)$ je však možné nalézt např. v článku Rousseeuew a Struyf (1998).

Výpočetní složitost odhadu regresního mediánu

Pokud chceme spočítat maximální hodnotu regresní hloubky vzhledem k náhodnému výběru (tj. odhadnout regresní medián) je potřeba uvažovat všechny přímky procházející alespoň p pozorováními (tedy alespoň zatím nebyl objeven jiný, rychlejší způsob). Počet takových přímek ale vždy může být až $\binom{n}{p}$, z čehož vyplývá, že výpočetní složitost maximální hodnoty regresní hloubky je v případě $p-1$ regresorů $\mathcal{O}(n^{2p-1}\log n)$.

Pozorování. Jedním z hlavních důvodů, proč se regresní hloubka příliš nepoužívá, je právě její vysoká výpočetní složitost. Pro $p \leq 3$ si totiž často ještě vystačíme s regresní diagnostikou, a není tedy nezbytně nutné používat robustní metody. V případě většího počtu regresorů bychom naopak metodu regresní hloubky více využili, rychlost výpočtu je ale v takovém případě velice pomalá (viz kapitola 6).

Poznámka. Funkce regresní hloubky (*rdepth*) je naprogramována v programu R v balíčku **mrfDepth**, stejně jako odhad regresního mediánu (*rdepthmedian*). Přesný výpočet probíhá v případě regresní hloubky pouze pro dimenzi $p \leq 4$ (tj. maximálně pro 3 regresory), v případě odhadu regresního mediánu dokonce pouze pro dimenzi $p = 2$ (tj. pouze pro 1 regresor). V obou případech se pak pro vyšší hodnoty p používají aproximativní algoritmy.

Poznámka I. Pokud bychom chtěli pouze odhad regresního mediánu, pak je možné využít ještě funkci *deepReg2d* z balíčku **DepthProc**. V tomto případě nejsou uvedena žádná omezení a používá se pouze originální algoritmus, který byl uveden v této sekci.

6. Simulační studie

Tato simulační studie byla inspirována studiemi v práci Rábek (2021) a v knize Rousseeuw a Leroy (1987), str. 208-214 a provedena za pomocí analytického softwaru R verze 4.0.1 s využitím balíčků **quantreg** a **mrfDepth**.

Cílem této studie je porovnat odhad získaný metodou regresní hloubky s odhady získanými pomocí metod LAD a OLS. Abychom zajistili, že všechny tři metody budou odhadovat stejný parametr zájmu, budeme volit taková počáteční rozdělení (tj. rozdělení správných pozorování), aby platilo

$$\mathsf{E}\,(Y|\widetilde{\boldsymbol{X}}) = \mathsf{med}\,(Y|\widetilde{\boldsymbol{X}}) = \widetilde{\boldsymbol{X}}^\top \boldsymbol{\beta}. \tag{6.1}$$

To můžeme zajistit například tak, že pro všechna $i \in \{1, \ldots, n\}$ zvolíme $F_i := F$, kde distribuční funkce F má *symetrické* rozdělení. Kromě toho se také omezíme pouze na případ $p = 2$, a to z toho důvodu, že jsme se určitými vlastnostmi regresní hloubky zabývali výhradně v tomto případě (viz věta 44, resp. tvrzení 36).

Provedeme dvě nezávislé studie, první pro počáteční volbu *normálního* rozdělení a druhou pro *Studentovo* t-rozdělení. V obou případech budeme předpokládat platnost modelu

$$\mathsf{E}\,(Y|\widetilde{\boldsymbol{X}}) = \mathsf{med}\,(Y|\widetilde{\boldsymbol{X}}) = X, \tag{6.2}$$

tj. odhadovanými parametry zájmu jsou $\beta_0 := 0$ a $\beta_1 := 1$. Abychom byli schopni metody dostatečně mezi sebou porovnat, vygenerujeme si z každého počátečního rozdělení $n_{sim} = 200$ náhodných výběrů (*simulací*), a pro každý z nich následně spočteme všechny 3 typy odhadů. Celkovou kvalitu odhadu (resp. metody) pak budeme posuzovat na základě odhadu střední čtvercové chyby, zkráceně pouze MSE (z *a.j. mean squared error*), kterou si pro daný koeficient definujeme jako

$$MSE(\beta_j) := \frac{1}{n_{sim}} \sum_{k=1}^{n_{sim}} (\hat{\beta}_j^{(k)} - \beta_j)^2,$$

kde $j \in \{0, 1\}$ a $\hat{\beta}_j^{(k)}$ je odhad koeficientu β_j v k-té simulaci. Hlavním cílem je však porovnávat **robustní** vlastnosti vybraných odhadů, proto hned ve druhém kroku vygenerované náhodné výběry kontaminujeme, a porovnáme mezi sebou nově vzniklé odhady na základě *kontaminovaných* výběrů. V rámci každé simulace budeme pak dále uvažovat ještě 3 měnitelné parametry, a to rozsah náhodného výběru, **stupeň a typ kontaminace**.

Poznámka. Kvůli výpočetním aspektům regresní hloubky jsme se rozhodli volit relativně malé rozsahy výběrů, stejně jako relativně malý počet simulací. Již při rozsahu výběru $n = 500$ trvá výpočet odhadu regresního mediánu (při jedné simulaci) 11 sekund, při rozsahu $n = 700$ je to 26 sekund, a při rozsahu $n = 1000$ dokonce 54 sekund.

Typy kontaminace*

Jak již bylo naznačeno dříve, ke kontaminaci může docházet různými způsoby, kde každý způsob může mít za následek zanesení jiného typu chyby do datového souboru. Podívejme se pro ilustraci na soubor grafů v obrázku 6.1, kde na každém

z nich je znázorněn trochu odlišný typ kontaminace, a k němu příslušné odhady regresní přímky. Na základě tohoto obrázku můžeme usuzovat, že některé typy kontaminace budou ovlivňovat odhad regresní přímky více než jiné.

Poznámka. I když by kontaminace na první pohled přímo neovlivňovala odhady koeficientů regresní přímky, může kompromitovat některé jiné charakteristiky, jako je například rozptyl odhadu. Nás však bude zajímat především jeho vychýlení, a proto se tímto typem vlivu nebudeme v této práci zabývat.

Poznámka. Oblast statistiky, která se zabývá posuzováním vlivu konkrétního (nebo nějaké skupiny) pozorování na odhad regresní přímky se nazývá *regresní diagnostika* (viz např. skripta Komárek (2019)).

Pro účely naší studie jsme se rozhodli zvolit následující 4 typy kontaminací:

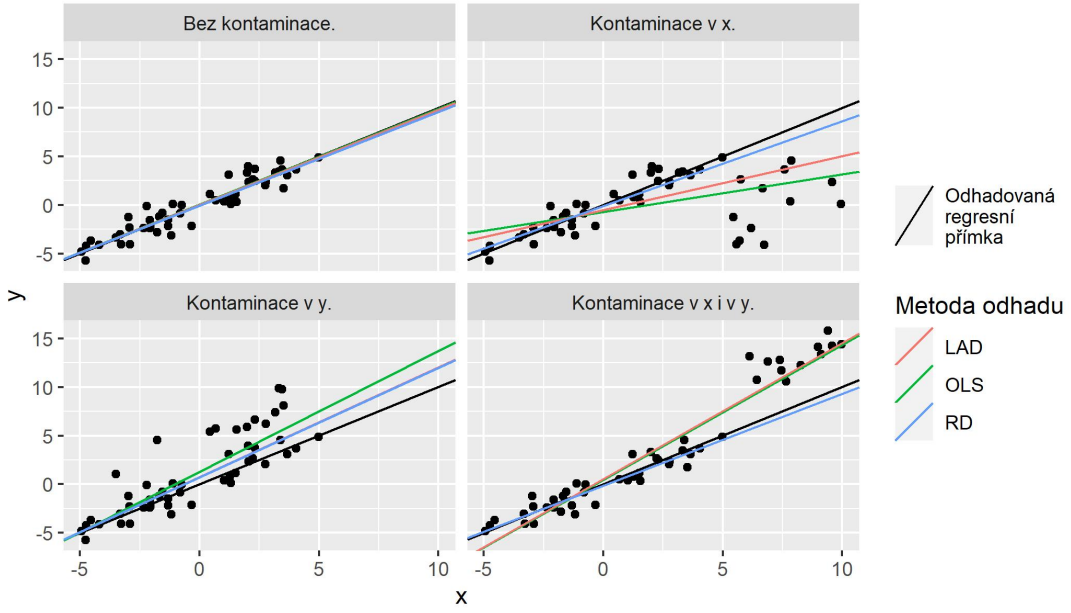
1. Kontaminaci chybových členů pozorováními, které pochází z jiného typu rozdělení, a to konkrétně z rozdělení s těžkými konci (tj. kontaminujeme pouze hodnoty y).

2. Kontaminaci hodnot regresorů pozorováními, které pochází ze stejného typu rozdělení, ale s jinými parametry (tj. kontaminujeme pouze hodnoty x).

3. Kontaminaci chybových členů pozorováními, které pochází ze stejného typu rozdělení, ale s jinými parametry (tj. kontaminujeme pouze hodnoty y).

4. Provedení kontaminace typu 2 a 3 současně (tj. kontaminujeme hodnoty x i y současně).

Pozorování. Kontaminace typu 2, 3 a 4 přesně odpovídají kontaminacím v grafech v obrázku 6.1.



Obrázek 6.1: Různé typy kontaminace datového souboru.

Vytvoření datového souboru

V následující části si popíšeme, jakým způsobem vznikla konkrétní sada pozorování pro **jednu** simulaci, kde každá simulace závisí na 3 volitelných parametrech:

- **Rozsah výběru:** $n = 20,\ 50,\ 100$,
- **Stupeň kontaminace:** $h = 0\,\%,\ 5\,\%,\ 10\,\%,\ 25\,\%$,
- **Typ kontaminace:** $t = 1,\ 2,\ 3,\ 4$ (viz dělení na předchozí stránce).

Poznámka. Protože je algoritmus pro obě volby počátečního rozdělení velmi podobný, budeme rozdílné kroky odlišovat pouze pomocí následujícího značení:

(i) Krok byl proveden při počáteční volbě *normálního rozdělení*,

(ii) Krok byl proveden při počáteční volbě *Studentova t-rozdělení*.

Algoritmus

1. Zvolíme si rozsah náhodného výběru $n = 20/50/100$.
2. Vygenerujeme náhodný výběr $X_1, \ldots, X_n$ z rovnoměrného rozdělení na intervalu $[-5, 5]$.
3. Vygenerujeme náhodný výběr $\varepsilon_1, \ldots, \varepsilon_n$ z
   (i) normálního rozdělení, a to konkrétně $\mathcal{N}(0, 1)$,
   (ii) Studentova t-rozdělení, a to konkrétně t_3.
4. Zvolíme si stupeň kontaminace $k = 0, 5, 10, 25$ na základě kterého náhodně vybereme $n_k := n \cdot k/100$ indexů $i_1, \ldots, i_{n_k}$ z množiny $\{1, \ldots, n\}$.
5. Zvolíme jeden z typů kontaminace a dle zvolené varianty postupujeme jedním z následujících způsobů:
   - Při kontaminaci typu 1 nahradíme výběr $\varepsilon_{i_1}, \ldots, \varepsilon_{i_{nk}}$ náhodným výběrem o rozsahu n_k z
     (i) Studentova t-rozdělení, a to konkrétně t_3,
     (ii) Cauchyho rozdělení.
   - Při kontaminaci typu 2 nahradíme výběr $X_{i_1}, \ldots, X_{n_k}$ náhodným výběrem o rozsahu n_k z rovnoměrného rozdělení na intervalu $[5, 10]$.
   - Při kontaminaci typu 3 nahradíme výběr $\varepsilon_{i_1}, \ldots, \varepsilon_{i_{nk}}$ náhodným výběrem o rozsahu n_k z
     (i) normálního rozdělení, a to konkrétně $\mathcal{N}(5, 1)$,
     (ii) posunutého Studentova t-rozdělení, a to konkrétně t_3 se střední hodnotou rovnou 5.

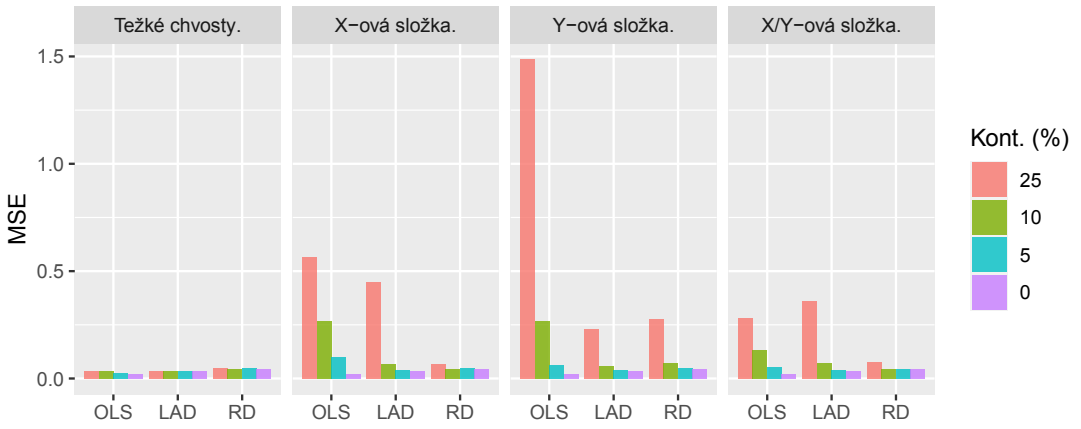
     (V obou případech tedy došlo oproti původnímu rozdělení pouze k posunu ve střední hodnotě, rozptyl zůstává zachován.)
   - Při kontaminaci typu 4 nejprve provedeme krok jako při kontaminaci typu 2, a následně pak krok jako při kontaminaci typu 3.
6. V neposlední řadě hodnoty *odezvy* zkonstruujeme jako $Y_i := X_i + \varepsilon_i$ (tj. předpokládáme $\beta_0 = 0$, $\beta_1 = 1$).

6.1 Shrnutí výsledků - $\mathcal{N}(0, 1)$

Výsledky jsou v následující části uvedeny pouze pro $n = 50$, a to jak v případě normálního rozdělení, tak Studentova t-rozdělení se 3 stupni volnosti. Je to z toho důvodu, že při měnícím se rozsahu výběru se sice mění MSE jednotlivých odhadů, pořadí metod však zůstává zachováno. Přesné hodnoty lze pak případně nalézt v tabulce v příloze.

Absolutní člen

V případě normálního rozdělení nemá kontaminace těžkými chvosty na odhad absolutního členu velký vliv (viz obrázek 6.2). Oproti tomu se zdá, že největší vliv na odhad absolutního členu má kontaminace v y-ové složce. Zde si nejlépe vede metoda LAD, za ní je metoda regresní hloubky (která dává o trochu horší výsledky), a nejhůře je na tom metoda OLS, která v tomto případě již značně selhává. K poněkud zvláštní situaci dochází v případě současné kontaminace v x-ové i y-ové složce, kdy i přes to, že metoda OLS v obou případech vychází výrazně horší než metoda LAD, se při vysoké kontaminaci (25%) pořadí metod otáčí. Nejlepší výsledky v tomto případě však jednoznačně dává metoda regresní hloubky, která i při vyšším stupni kontaminace dává relativně (v poměru k metodám LAD a OLS) malé hodnoty MSE. Jednoznačně nejlepší chování také regresní hloubka vykazuje v případě kontaminace v x-ové složce, kdy na rozdíl od metod OLS a LAD jsou hodnoty MSE malé při libovolném stupni kontaminace.

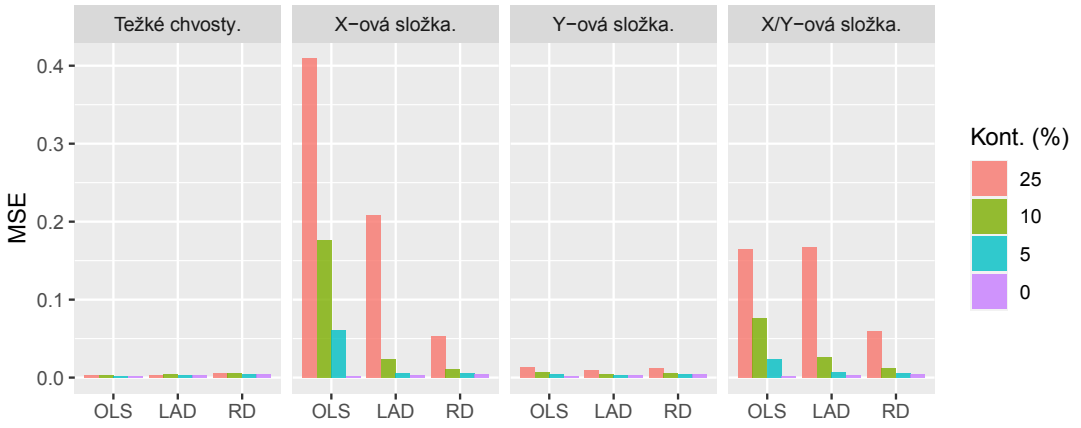


Obrázek 6.2: MSE absolutního členu pro normální rozdělení pro volbu $n = 50$ při různých typech a stupních kontaminace datového souboru.

Směrnice

Jak můžeme nahlédnout z obrázku 6.3, na vychýlení odhadu směrnice regresní přímky se ukazuje mít největší vliv kontaminace v x-ové složce, kde si opět vede nejlépe metoda regresní hloubky. Hodnota MSE je u 10% kontaminace cca poloviční oproti metodě LAD, a dokonce přibližně 15krát menší než v případě metody OLS. Stejně tak si opět metoda RD vede nejlépe i v případě současné kontaminace v x-ové i y-ové složce, což je opět nejspíše způsobeno tím, že je výrazně

robustnější v x-ové složce. Při kontaminaci v y-ové složce naopak opět zaostává za metodou LAD, a pro tento případ se tedy nejeví jako dobrá volba.

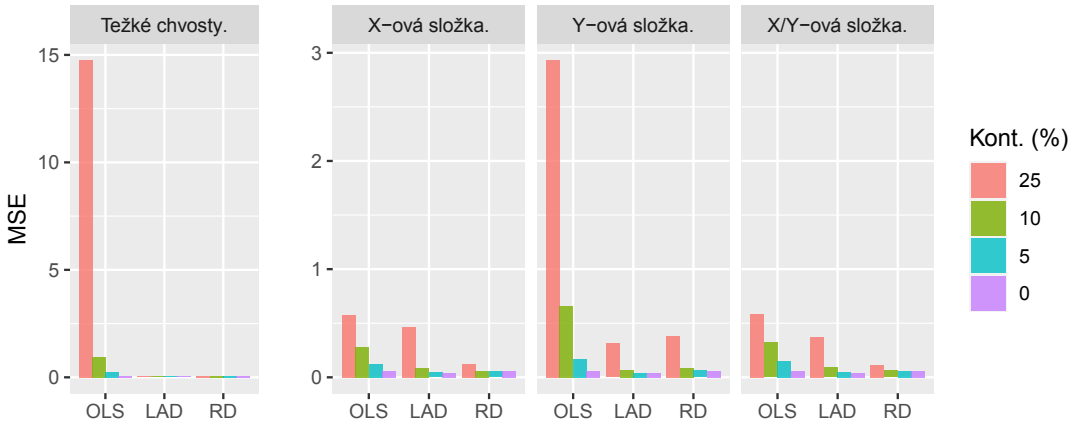


Obrázek 6.3: MSE směrnice pro normální rozdělení pro volbu $n = 50$ při různých typech a stupních kontaminace datového souboru.

6.2 Shrnutí výsledků - t_3

Absolutní člen

Na rozdíl od normálního rozdělení pozorujeme u Studentova t-rozdělení velké vychýlení odhadu OLS při kontaminaci rozdělením s těžkými chvosty, což je nejspíše způsobeno *výrazně* těžšími chvosty rozdělení chybných pozorování. Stejně jako v předchozím případě se však ukazuje, že metoda regresní hloubky si vede nejlépe v situaci, kdy dochází ke kontaminace v x-ové, resp. současné kontaminaci v x-ové i y-ové složce, zatímco metoda LAD si vede nejlépe při kontaminaci pouze v y-ové složce.

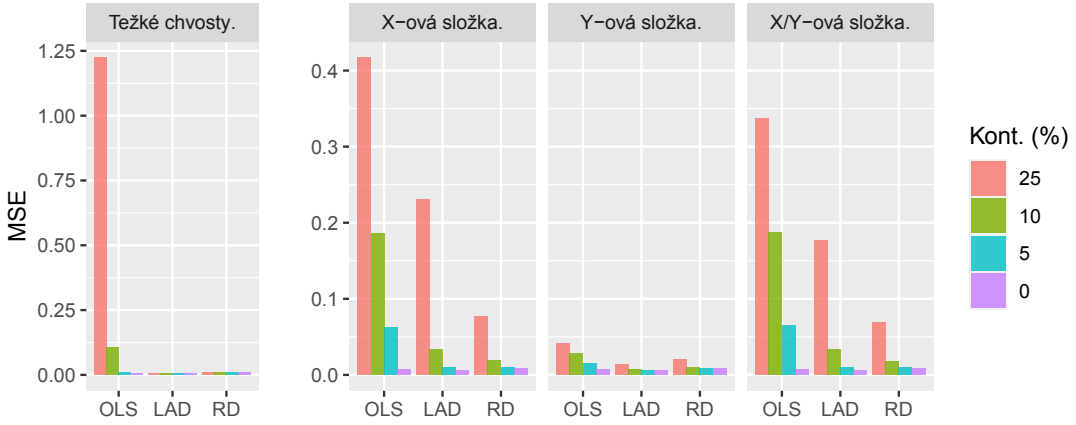


Obrázek 6.4: MSE absolutního členu pro Studentovo rozdělení o 3 stupních volnosti pro volbu $n = 50$ při různých typech a stupních kontaminace datového souboru.

Poznámka. Znovu pozorujeme, že vzhledem k odezvě se metoda LAD zdá robustnější než metoda regresní hloubky. Jak již však bylo dříve poznamenáno, robustnost metody LAD vzhledem k odezvě závisí na jejím vztahu k regresorům. Není tedy možné vyloučit, že při jiném způsobu kontaminace bychom dostali výrazně odlišné výsledky.

Směrnice

Stejně jako v případě absolutního členu, i u směrnice regresní přímky pozorujeme výrazný nárůst hodnoty MSE pří kontaminaci těžkými chvosty. Jinak je chování odhadů obdobné jako u normálního rozdělení, největší vliv se ukazuje mít kontaminace v x-ové složce, kde si nejlépe vede metoda regresní hloubky. Vliv tohoto typu kontaminace je opět velice výrazný i při současné kontaminaci v x-ové a y-ové složce, zatímco v případe kontaminace pouze v y-ové složce je vliv na vychýlení odhadů podstatně menší.



Obrázek 6.5: MSE směrnice pro Studentovo rozdělení o 3 stupních volnosti pro volbu $n = 50$ při různých typech a stupních kontaminace datového souboru.

Závěr

V rámci této práce byla představena metoda regresní hloubky jakožto jedna z novějších a méně známých robustních metod. Práce byla rozdělena na několik stěžejních částí: V první části práce (kapitoly 1 a 2) byly položeny základy v podobě zavedení používaného značení a představení několika základních pojmů (lineární model, robustní vlastnosti odhadu, atd.). Ve druhé části (kapitoly 3 a 4) byly pak ve stručnosti představeny vybrané metody, a to konkrétně metoda OLS a LAD, se kterými jsme později metodu regresní hloubky porovnávali.

Třetí část práce (kapitola 5) pak představuje nejobsáhlejší a nejdůležitější část, a byla celá věnována pouze metodě regresní hloubky. Na úvodní motivační část, kde bylo popsáno, jakým způsobem autoři při konstrukci regresní hloubky v minulosti postupovali, navazuje část ilustrační. Ukázalo se totiž, že kromě původní definice má regresní hloubka i jinou interpretaci, a to geometrickou. Ta se pro nás později ukázala jako stěžejní, když jsme chtěli dokázat, že *regresní medián* (tj. vektor koeficientů maximalizujícího regresní hloubku) za určitých předpokladů přímo odpovídá podmíněnému mediánu odezvy při znalosti hodnot regresorů. Další podstatná část této kapitoly se pak zaobírala důkazem konzistence odhadu regresního mediánu, který původně pochází z článku Bai a He (1999). Pojetí důkazu v původním článku je však velice stručné, a bylo tak potřeba jej doplnit o nemalé množství odvození a souvislostí, které dohromady tvoří jeden z podstatných přínosů této práce. Závěr kapitoly pak patřil robustním vlastnostem regresní hloubky, *breakdown-pointu* a *influenční funkci*, kde bylo ukázáno, že za určitých předpokladů množství chybných pozorování, která mohou být ve výběru obsažena aniž by se odhad stal „nepoužitelným“, je pro odhad metodou regresní hloubky 33 %, a v případě, kdy pracujeme pouze s jedním regresorem, je influenční funkce omezená jak vzhledem k odezvě, tak vzhledem k hodnotám regresorů.

Na závěr práce byla vyhotovena malá simulační studie pro dvě různé volby počátečního rozdělení, a to konkrétně *normálního* rozdělení a *Studentova* t-rozdělení o 3 stupních volnosti. Na základě této studie se odhad metodou regresní hloubky skutečně ukázal být robustní, a to především při kontaminaci vzhledem k hodnotám regresorů, kde dával výrazně lepší výsledky než metody LAD i OLS. Pokud jde o kontaminaci vzhledem k hodnotám odezvy, ukázalo se, že chybná pozorování mohou mít na odhad metodou regresní hloubky větší vliv než jsme očekávali, a proto doporučujeme v tomto případě zůstat u standardních metod.

Metoda regresní hloubky se tedy celkově ukázala jako vhodná volba při výběru robustní metody, a to především při kontaminaci hodnot regresorů. Bohužel, výpočetní složitost této metody roste exponenciálně s počtem regresorů, a není tedy vhodná při jejich větším počtu. S pokročilejšími technologiemi se sice její použití stává čím dále více dostupnější, pořád však tato skutečnost představuje v praxi podstatný problém. V případě menšího množství regresorů si totiž často vystačíme pouze s regresní diagnostikou, a není tak potřeba využívat robustní metody. Aproximativní algoritmy jsou však neustále předmětem dalších studií, viz například článek Rousseeuew a Struyf (1998), a proto určitě nedoporučujeme metodu regresní hloubky do budoucna zavrhovat.

Seznam použité literatury

Amenta, N., Bern, M., Eppstein, D. a Teng, S. (2000). Regression depth and center points. *Discrete Comput. Geom.*, **23**, 305–323.

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha.

Bai, Z.-D. a He, X. (1999). Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *The Annals of Statistics*, **27**(5), 1616–1637.

Donoho, D. L. a Huber, P. J. (1983). The notion of breakdown point. *In A Festschrift for Erich L. Lehmann*, pages 157–184.

Dupačová, J. a Lachout, P. (2011). *Úvod do optimalizace*. MatfyzPress.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**(346).

Hampel, F. R. (1997). *Robust statistics*. J. Wiley.

Jurjen, D. T. E. (2012). *Analýzy Metod Pro Maticové výpočty: Základní Metody*. Matfyzpress, Praha.

Komárek, A. (2019). Course notes, nmsa407 linear regression. `https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2021_22/nmsa407/2021-NMSA407-notes.pdf`. Accessed: 2022–18-07.

Omelka (2020). Course notes, nmst 434 modern statistical methods. `https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmst434/nmst434_course-notes.pdf`. Accessed: 2022–18-07.

Prohorov, Y. (1956). Convergence of random processes and limit theorems in probability theory. *Theor. Prob. Appl.*, **1**(2), 157–214.

Rousseeuew, P. a Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, **8**, 193–203.

Rousseeuw, P. J. a Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, **94**(446), 388–402.

Rousseeuw, P. J. a Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. J. Wiley.

Rábek, J. (2021). Robust linear regression. *Master thesis*.

Tukey, J. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485.

Tukey, J. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, pages 523–531.

Van Aelst, S. a Rousseeuw, P. J. (2000). Robustness of deepest regression. *Journal of Multivariate Analysis*, **73**(1), 82–106.

Van Aelst, S., Rousseeuw, P. J., Hubert, M. a Struyf, A. (2002). The deepest regression method. *Journal of Multivariate Analysis*, **81**(1).

Zuo, Y. (2001). Some quantitative relationships between two types of finite sample breakdown point. *Statistics and Probability Letters*, **51**(4), 369–375.

Zuo, Y. (2020). Large sample properties of the regression depth induced median. *Statistics and Probability Letters*, **166**.

Seznam zkratek a symbolů

Zkratky

BÚNO .. bez újmy na obecnosti

MSE .... střední čtvercová chyba (z *a.j. mean squared error*)

OLS ..... metoda nejmenších čtverců (z *a.j. ordinary least squares*)

LAD .... metoda nejmenších absolutních odchylek
(z *a.j. least absolute deviation*)

RD ..... metoda regresní hloubky (z *a.j. regression depth*)

BP ..... breakdown-point

IF ..... influenční funkce

Symboly

$\mathbf{1}_n$ ..... n-rozměrný jednotkový vektor

$\mathbf{0}_n$ ..... n-rozměrný nulový vektor

I_n ..... jednotková matice typu $n \times n$

$s.v.$ ..... skoro všude

$s.j.$ ..... skoro jistě

$\xrightarrow{s.j.}$ ..... konvergence skoro jistě

$\xrightarrow{p}$ ..... konvergence v pravděpodobnosti

$\|\cdot\|_2$ ..... euklidovská norma

ϕ ..... hustota normálního rozdělení $\mathcal{N}(0,1)$

Φ ..... distribuční funkce normálního rozdělení $\mathcal{N}(0,1)$

$\mathcal{O}(\cdot)$ ..... notace velkého O

A. Přílohy

A.1 Pomocné lemma

Lemma 45. *Libovolnou pozitivně definitní matici $A \in \mathbb{R}^{p\times p}$, $p \in \mathbb{N}$ lze po řadě rozložit na horní a dolní trojúhelníkovou matici, tj. existuje vyjádření*

$$A = UU^\top,$$

kde $U \in \mathbb{R}^{p\times p}$ je horní trojúhelníková matice.

Důkaz. Buď $\widetilde{A} \in \mathbb{R}^{p\times p}$ libovolná pozitivně definitní matice, z Choleského dekompozice (viz například Jurjen (2012), str.101) vyplývá, že existuje dolní trojúhelníková matice $L \in \mathbb{R}^{p\times p}$ taková, že

$$\widetilde{A} = LL^\top.$$

Uvažujme nyní permutační matici $P \in \mathbb{R}^{p\times p}$ definovanou jako

$$P := \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix},$$

tj. matice P má jedničky na diagonále. Protože P je ortogonální a symetrická, můžeme psát

$$\widetilde{A} = LL^\top = (PU)(PU)^\top = PU(PP^\top)U^\top P^\top = PUU^\top P,$$

a tedy pro matici $A := P\widetilde{A}P = UU^\top$ existuje po řadě rozklad na horní a dolní trojúhelníkovou matici. Výsledek lemmatu nakonec vyplývá ze skutečnosti, že matice P je regulární a $\widetilde{A}$ je pozitivně definitní právě tehdy, když A je pozitivně definitní.

□

A.2 Výstupy ze simulační studie

Následující tabulka obsahuje výstupní hodnoty ze simulační studie, kde sloupec MSE 1 odpovídá průměrné hodnotě MSE *absolutního členu* a sloupec MSE 2 odpovídá průměrné hodnotě MSE *směrnice regresní přímky* (v obou případech se jedná o průměr z $n_{sim} = 200$ simulací).

Distribuce	Typ kontaminace	Kontaminace (%)	Metoda	MSE	
				Abs. člen	Směrnice
normální	Těžké chvosty.	25	OLS	0.082	0.009
			LAD	0.092	0.010
			RD	0.140	0.015
		10	OLS	0.066	0.010
			LAD	0.100	0.012
			RD	0.130	0.014
		5	OLS	0.053	0.009
			LAD	0.086	0.011
			RD	0.116	0.015
		0	OLS	0.050	0.009
			LAD	0.089	0.011
			RD	0.119	0.015
	X-ová složka.	25	OLS	0.692	0.455
			LAD	0.731	0.295
			RD	0.228	0.071
		10	OLS	0.338	0.217
			LAD	0.142	0.043
			RD	0.122	0.025
		5	OLS	0.189	0.103
			LAD	0.105	0.018
			RD	0.130	0.018
		0	OLS	0.050	0.009
			LAD	0.089	0.011
			RD	0.119	0.015
	Y-ová složka.	25	OLS	1.633	0.040
			LAD	0.360	0.030
			RD	0.530	0.056
		10	OLS	0.334	0.030
			LAD	0.137	0.018
			RD	0.161	0.022
		5	OLS	0.122	0.016
			LAD	0.090	0.013
			RD	0.138	0.018
		0	OLS	0.050	0.009
			LAD	0.089	0.011
			RD	0.119	0.015
	X/Y-ová složka.	25	OLS	0.363	0.182
			LAD	0.515	0.191
			RD	0.321	0.092
		10	OLS	0.195	0.084
			LAD	0.150	0.041
			RD	0.130	0.021
		5	OLS	0.107	0.040
			LAD	0.103	0.017
			RD	0.113	0.016
		0	OLS	0.050	0.009
			LAD	0.089	0.011
			RD	0.119	0.015

Tabulka A.1: Tabulka odhadů regresních koeficientů pro $p = 2$ při různých stupních a typech kontaminace pro počáteční volbu *normálního* rozdělení a rozsahu náhodného výběru $n = 20$.

				MSE	
Distribuce	Typ kontaminace	Kontaminace (%)	Metoda	Abs. člen	Směrnice
normální	Těžké chvosty.	25	OLS	0.035	0.004
			LAD	0.036	0.004
			RD	0.047	0.006
		10	OLS	0.033	0.003
			LAD	0.034	0.004
			RD	0.043	0.006
		5	OLS	0.024	0.002
			LAD	0.036	0.003
			RD	0.047	0.004
		0	OLS	0.022	0.002
			LAD	0.034	0.004
			RD	0.044	0.005
	X-ová složka.	25	OLS	0.564	0.409
			LAD	0.448	0.209
			RD	0.066	0.053
		10	OLS	0.266	0.176
			LAD	0.070	0.024
			RD	0.045	0.012
		5	OLS	0.098	0.061
			LAD	0.039	0.006
			RD	0.047	0.006
		0	OLS	0.022	0.002
			LAD	0.034	0.004
			RD	0.044	0.005
	Y-ová složka.	25	OLS	1.487	0.014
			LAD	0.230	0.009
			RD	0.278	0.012
		10	OLS	0.270	0.007
			LAD	0.058	0.004
			RD	0.073	0.006
		5	OLS	0.063	0.004
			LAD	0.039	0.004
			RD	0.050	0.005
		0	OLS	0.022	0.002
			LAD	0.034	0.004
			RD	0.044	0.005
	X/Y-ová složka.	25	OLS	0.280	0.165
			LAD	0.361	0.167
			RD	0.076	0.059
		10	OLS	0.132	0.076
			LAD	0.071	0.027
			RD	0.043	0.012
		5	OLS	0.054	0.024
			LAD	0.040	0.006
			RD	0.046	0.006
		0	OLS	0.022	0.002
			LAD	0.034	0.004
			RD	0.044	0.005

Tabulka A.2: Tabulka odhadů regresních koeficientů pro $p = 2$ při různých stupních a typech kontaminace pro počáteční volbu *normálního* rozdělení a rozsahu náhodného výběru $n = 50$.

				MSE	
Distribuce	Typ kontaminace	Kontaminace (%)	Metoda	Abs. člen	Směrnice
normální	Těžké chvosty.	25	OLS	0.016	0.002
			LAD	0.015	0.002
			RD	0.023	0.003
		10	OLS	0.012	0.002
			LAD	0.015	0.002
			RD	0.021	0.003
		5	OLS	0.012	0.001
			LAD	0.014	0.002
			RD	0.020	0.003
		0	OLS	0.011	0.001
			LAD	0.016	0.002
			RD	0.022	0.003
	X-ová složka.	25	OLS	0.468	0.413
			LAD	0.352	0.208
			RD	0.037	0.061
		10	OLS	0.204	0.173
			LAD	0.040	0.020
			RD	0.018	0.009
		5	OLS	0.089	0.075
			LAD	0.018	0.006
			RD	0.019	0.004
		0	OLS	0.011	0.001
			LAD	0.016	0.002
			RD	0.022	0.003
	Y-ová složka.	25	OLS	1.584	0.006
			LAD	0.212	0.004
			RD	0.243	0.005
		10	OLS	0.261	0.004
			LAD	0.038	0.003
			RD	0.042	0.004
		5	OLS	0.074	0.003
			LAD	0.020	0.003
			RD	0.029	0.003
		0	OLS	0.011	0.001
			LAD	0.016	0.002
			RD	0.022	0.003
	X/Y-ová složka.	25	OLS	0.255	0.169
			LAD	0.327	0.173
			RD	0.038	0.056
		10	OLS	0.105	0.072
			LAD	0.047	0.021
			RD	0.019	0.009
		5	OLS	0.048	0.030
			LAD	0.023	0.006
			RD	0.021	0.004
		0	OLS	0.011	0.001
			LAD	0.016	0.002
			RD	0.022	0.003

Tabulka A.3: Tabulka odhadů regresních koeficientů pro $p = 2$ při různých stupních a typech kontaminace pro počáteční volbu *normálního* rozdělení a rozsahu náhodného výběru $n = 100$.

				MSE	
Distribuce	Typ kontaminace	Kontaminace (%)	Metoda	Abs. člen	Směrnice
Studentovo	Těžké chvosty.	25	OLS	31.084	3.680
			LAD	0.122	0.022
			RD	0.164	0.027
		10	OLS	301.984	39.301
			LAD	0.112	0.016
			RD	0.165	0.022
		5	OLS	14.627	2.527
			LAD	0.111	0.016
			RD	0.166	0.020
		0	OLS	0.137	0.020
			LAD	0.113	0.018
			RD	0.148	0.020
	X-ová složka.	25	OLS	0.705	0.470
			LAD	0.635	0.329
			RD	0.359	0.138
		10	OLS	0.409	0.229
			LAD	0.188	0.058
			RD	0.146	0.031
		5	OLS	0.250	0.112
			LAD	0.122	0.027
			RD	0.154	0.021
		0	OLS	0.137	0.020
			LAD	0.113	0.018
			RD	0.148	0.020
	Y-ová složka.	25	OLS	3.216	0.115
			LAD	0.589	0.062
			RD	0.854	0.090
		10	OLS	0.794	0.075
			LAD	0.165	0.021
			RD	0.212	0.027
		5	OLS	0.372	0.044
			LAD	0.126	0.018
			RD	0.201	0.023
		0	OLS	0.137	0.020
			LAD	0.113	0.018
			RD	0.148	0.020
	X/Y-ová složka.	25	OLS	0.814	0.331
			LAD	0.610	0.196
			RD	0.333	0.107
		10	OLS	0.431	0.171
			LAD	0.181	0.048
			RD	0.164	0.034
		5	OLS	0.502	0.714
			LAD	0.132	0.021
			RD	0.147	0.025
		0	OLS	0.137	0.020
			LAD	0.113	0.018
			RD	0.148	0.020

Tabulka A.4: Tabulka odhadů regresních koeficientů pro $p = 2$ při různých stupních a typech kontaminace pro počáteční volbu *Studentova* t-rozdělení o 3 stupních volnosti a rozsahu náhodného výběru $n = 20$.

				MSE	
Distribuce	Typ kontaminace	Kontaminace (%)	Metoda	Abs. člen	Směrnice
Studentovo	Těžké chvosty.	25	OLS	14.766	1.224
			LAD	0.036	0.006
			RD	0.055	0.009
		10	OLS	0.950	0.105
			LAD	0.038	0.006
			RD	0.058	0.008
		5	OLS	0.234	0.010
			LAD	0.036	0.006
			RD	0.058	0.008
		0	OLS	0.059	0.008
			LAD	0.038	0.006
			RD	0.056	0.009
	X-ová složka.	25	OLS	0.576	0.417
			LAD	0.464	0.231
			RD	0.123	0.077
		10	OLS	0.281	0.186
			LAD	0.089	0.034
			RD	0.057	0.019
		5	OLS	0.122	0.062
			LAD	0.045	0.009
			RD	0.054	0.010
		0	OLS	0.059	0.008
			LAD	0.038	0.006
			RD	0.056	0.009
	Y-ová složka.	25	OLS	2.934	0.041
			LAD	0.311	0.014
			RD	0.381	0.020
		10	OLS	0.659	0.029
			LAD	0.067	0.008
			RD	0.086	0.010
		5	OLS	0.172	0.015
			LAD	0.041	0.006
			RD	0.065	0.009
		0	OLS	0.059	0.008
			LAD	0.038	0.006
			RD	0.056	0.009
	X/Y-ová složka.	25	OLS	0.586	0.338
			LAD	0.371	0.177
			RD	0.113	0.069
		10	OLS	0.320	0.188
			LAD	0.091	0.034
			RD	0.062	0.017
		5	OLS	0.153	0.065
			LAD	0.050	0.010
			RD	0.057	0.010
		0	OLS	0.059	0.008
			LAD	0.038	0.006
			RD	0.056	0.009

Tabulka A.5: Tabulka odhadů regresních koeficientů pro $p = 2$ při různých stupních a typech kontaminace pro počáteční volbu *Studentova* t-rozdělení o 3 stupních volnosti a rozsahu náhodného výběru $n = 50$.

				MSE	
Distribuce	Typ kontaminace	Kontaminace (%)	Metoda	Abs. člen	Směrnice
Studentovo	Těžké chvosty.	25	OLS	134.850	21.420
			LAD	0.019	0.003
			RD	0.027	0.004
		10	OLS	2.444	0.112
			LAD	0.021	0.003
			RD	0.025	0.003
		5	OLS	0.832	0.143
			LAD	0.022	0.002
			RD	0.030	0.003
		0	OLS	0.037	0.005
			LAD	0.020	0.002
			RD	0.028	0.003
	X-ová složka.	25	OLS	0.491	0.414
			LAD	0.358	0.225
			RD	0.053	0.084
		10	OLS	0.234	0.178
			LAD	0.052	0.024
			RD	0.028	0.010
		5	OLS	0.114	0.070
			LAD	0.027	0.006
			RD	0.026	0.005
		0	OLS	0.037	0.005
			LAD	0.020	0.002
			RD	0.028	0.003
	Y-ová složka.	25	OLS	2.948	0.027
			LAD	0.285	0.007
			RD	0.309	0.010
		10	OLS	0.588	0.013
			LAD	0.048	0.003
			RD	0.059	0.004
		5	OLS	0.155	0.008
			LAD	0.027	0.003
			RD	0.036	0.003
		0	OLS	0.037	0.005
			LAD	0.020	0.002
			RD	0.028	0.003
	X/Y-ová složka.	25	OLS	0.541	0.319
			LAD	0.343	0.166
			RD	0.057	0.085
		10	OLS	0.225	0.145
			LAD	0.058	0.027
			RD	0.031	0.011
		5	OLS	0.115	0.062
			LAD	0.029	0.007
			RD	0.027	0.005
		0	OLS	0.037	0.005
			LAD	0.020	0.002
			RD	0.028	0.003

Tabulka A.6: Tabulka odhadů regresních koeficientů pro $p = 2$ při různých stupních a typech kontaminace pro počáteční volbu *Studentova* t-rozdělení o 3 stupních volnosti a rozsahu náhodného výběru $n = 100$.