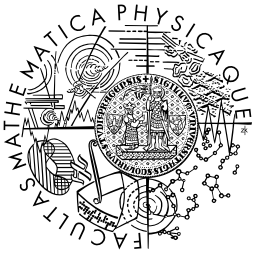


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Klára Čelikovská

L_1 regrese

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Matúš Maciak, Ph.D.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2020

Prohlašuji, že jsem tuto diplomovou práci vypracovala 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

Zde bych ráda poděkovala RNDr. Matúšovi Maciakovi, Ph.D. za zajímavé téma a také za rady a připomínky při vedení práce.

Název práce: L_1 regrese

Autor: Klára Čelikovská

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Matúš Maciak, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá L_1 regresí, která je možnou alternativou klasické lineární regrese. Odhad metodou nejmenších čtverců je u L_1 regrese nahrazen odhadem metodou nejmenších absolutních odchylek, což vede k zobecnění výběrového mediánu v lineárním regresním modelu. Oproti klasické lineární regresi umožňuje L_1 regrese uvolnit některé předpoklady a je robustnější vůči odlehlým pozorováním. Je dokázána základní teorie včetně asymptotického rozdělení odhadů regresních koeficientů, testů hypotéz, konfidenčních intervalů a konfidenčních množin. Metoda je následně porovnána s klasickou lineární regresí v simulační studii, která je zaměřená na data z rozdělení s těžkými chvosty a na data znečištěná odlehlými pozorováními.

Klíčová slova: L_1 norma, mediánová regrese, robustní odhad, odlehlá pozorování

Title: L_1 Regression

Author: Klára Čelikovská

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis is focused on the L_1 regression, a possible alternative to the ordinary least squares regression. L_1 regression replaces the least squares estimation with the least absolute deviations estimation, thus generalizing the sample median in the linear regression model. Unlike the ordinary least squares regression, L_1 regression enables loosening of certain assumptions and leads to more robust estimates. Fundamental theoretical results, including the asymptotic distribution of regression coefficient estimates, hypothesis testing, confidence intervals and confidence regions, are derived. This method is then compared to the ordinary least squares regression in a simulation study, with a focus on heavy-tailed distributions and the possible presence of outlying observations.

Keywords: L_1 norm, median regression, robust estimation, outlying observations

Obsah

Úvod

V práci se budeme zabývat regresní analýzou, kde je naším cílem vysvětlit vztah mezi závislou proměnnou (odezvou) a jednou či několika nezávislými proměnnými (vysvětlujícími proměnnými, prediktory, regresory). Nejběžnějším modelem pro tento problém je lineární model, který předpokládá lineární závislost na nezávislých proměnných vyjádřenou pomocí neznámých parametrů. Parametry jsou v modelu obvykle odhadovány metodou nejmenších čtverců, alternativním způsobem odhadu může být metoda nejmenších absolutních odchylek, která umožňuje některé předpoklady uvolnit.

Je dobře známé, že výběrový průměr minimalizuje součet čtverců odchylek (L_2 ztrátová funkce), zatímco výběrový medián minimalizuje součet absolutních odchylek (L_1 ztrátová funkce). V případě symetrického rozdělení jsou si střední hodnota a medián rovny a je tak odhadována stejná veličina, nicméně odhady mají odlišné vlastnosti. Zatímco výběrový průměr je nejlepším nestranným odhadem střední hodnoty, výběrový medián není tak citlivý na odlehlá pozorování a dává smysl, i pokud střední hodnota neexistuje, protože medián existuje vždy. Na rozdíl od výběrového průměru však není eficientní v případech, kdy jsou všechny předpoklady splněny. Asymptotické rozdělení pro výběrový medián lze odvodit i v případě, že rozdělení nemá konečný rozptyl, zatímco u výběrového průměru je potřeba konečný rozptyl pro aplikaci centrální limitní věty.

Tyto poznatky lze zobecnit i na regresní úlohu, kdy klasická lineární regrese (L_2 regrese) je zobecněním výběrového průměru a mediánová regrese (L_1 regrese) je zobecněním výběrového mediánu. V případě klasické lineární regrese pak modelujeme vztah mezi závislou proměnnou a nezávislými proměnnými ve smyslu podmíněné střední hodnoty, u L_1 regrese ve smyslu podmíněného mediánu.

Cílem této práce je ukázat, že mediánová regrese dává lepší výsledky v případě porušení některých předpokladů klasické lineární regrese, a je tedy vhodné ji použít, pokud si nejsme jisti splněním požadovaných předpokladů. To se týká především rozdělení s těžkými chvosty (např. Cauchyho rozdělení) a dat znečištěných odlehlými pozorováními.

V případě nesymetrického rozdělení může být žádané modelovat medián místo střední hodnoty, nicméně v takovém případě nelze lineární regresi a mediánovou regresi přímo porovnat, protože modelují různé aspekty podmíněného rozdělení odezvy, a i jejich lineární závislost na regresorech se může lišit.

L_1 regrese je speciálním případem kvantilové regrese, která modeluje obecný podmíněný kvantil odezvy. Některá základní odvození v této práci budou provedena pro obecný kvantil, u hlavních výsledků se zaměříme pouze na medián. Většinu výsledků lze však zobecnit i pro obecný kvantil.

Kapitola 1 slouží k zavedení základního značení a terminologie lineárního modelu a připomenutí metody nejmenších čtverců. V Kapitole 2 jsou odvozeny teoretické poznatky L_1 regrese, zejména asymptotické rozdělení odhadů parametrů, pomocí kterého jsou sestrojeny nástroje pro inferenci (testy hypotéz, konfidenční intervaly a množiny). Kapitola 3 obsahuje simulační studii, která srovnává L_1 regresi s L_2 regresí ve smyslu přesnosti odhadů parametrů, chyby I. typu a síly testů a pokrytí konfidenčních intervalů a množin.

1. Lineární model

V této kapitole si nejprve zavedeme potřebné značení a terminologii související s lineárním regresním modelem, které budeme používat v průběhu celého textu. Následně si připomeneme odhad parametrů v lineárním modelu metodou nejmenších čtverců, který budeme v průběhu práce a zejména v simulační studii v Kapitole 3 srovnávat s L_1 odhadem.

1.1 Značení a terminologie

Budeme předpokládat, že máme k dispozici n pozorování nezávislých a stejně rozdělených náhodných vektorů

$$\begin{pmatrix} Y_i \\ \boldsymbol{X}_i \end{pmatrix} \overset{iid}{\sim} F_{Y,\boldsymbol{X}},\ i = 1, \dots, n,$$

kde $F_{Y,\boldsymbol{X}}$ je sdružená distribuční funkce generického náhodného vektoru $(Y, \boldsymbol{X}^\top)^\top$, $\boldsymbol{X} = (X_0, \dots, X_{k-1})^\top$ a $\boldsymbol{X}_i = (X_{i,0}, \dots, X_{i,k-1})^\top,\ i = 1, \dots, n$.

Dále budeme značit

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{X} = \begin{pmatrix} X_{1,0} & \dots & X_{1,k-1} \\ \vdots & & \vdots \\ X_{n,0} & \dots & X_{n,k-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = (\boldsymbol{X}^0, \dots, \boldsymbol{X}^{k-1}),$$

kde $\boldsymbol{Y}$ nazýváme vektorem odezvy a $n \times k$ matici $\mathbb{X}$ nazýváme regresní maticí či maticí modelu. Vektory $\boldsymbol{X}_i = (X_{i,0}, \dots, X_{i,k-1})^\top, i = 1, \dots, n$ reprezentují hodnoty vysvětlujících proměnných pro pozorování i.

Lineární regresní model definujeme jako

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{k-1})^\top \in \mathbb{R}^k$ je vektor neznámých parametrů (regresních koeficientů) a náhodný vektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top, \varepsilon_i \overset{iid}{\sim} F_\varepsilon$ nazýváme chybovým vektorem, kde F_ε je distribuční funkce generické náhodné veličiny $\varepsilon = Y - \boldsymbol{X}^\top\boldsymbol{\beta}$. Složky chybového vektoru $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}$, $i = 1, \dots, n$ nazýváme chybové členy či náhodné chyby modelu.

Obvykle $\boldsymbol{X}_i = (1, X_{i,1}, \dots, X_{i,k-1})^\top,\ i = 1, \dots, n$, tedy regresor X_0 je konstantně roven jedné. V takovém případě model nazýváme modelem s absolutním členem a parametr β_0 nazýváme absolutním členem. Platí pak

$$\mathbb{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,k-1} \\ \vdots & & & \vdots \\ 1 & X_{n,1} & \dots & X_{n,k-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = (\mathbf{1}_n, \boldsymbol{X}^1, \dots, \boldsymbol{X}^{k-1}).$$

V této práci se budeme zabývat pouze spojitou odezvou Y, vysvětlující proměnné mohou být libovolné (spojité, případně kategoriální). Základní metodou odhadu pro lineární model je metoda nejmenších čtverců.

1.2 Metoda nejmenších čtverců

V lineární regresi je obvykle modelována podmíněná střední hodnota Y při daných hodnotách vysvětlujících proměnných. Předpokládáme, že rozdělení náhodného vektoru $(Y, \boldsymbol{X}^\top)^\top$ splňuje

$$\mathsf{E}\,(Y|\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{\beta}, \qquad \mathrm{var}(Y|\boldsymbol{X}) = \sigma^2,$$

kde $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{k-1})^\top \in \mathbb{R}^k$ a $0 < \sigma^2 < \infty$ jsou neznámé parametry. Parametr σ^2 nazýváme reziduálním rozptylem.

Pro náhodný výběr $(Y_i, \boldsymbol{X}_i^\top)^\top, i = 1, \ldots, n$ splňující tyto předpoklady zřejmě platí

$$\mathsf{E}\,(\,\boldsymbol{Y} \mid \mathbb{X}) = \mathbb{X}\boldsymbol{\beta}, \qquad \mathrm{var}(\,\boldsymbol{Y} \mid \mathbb{X}) = \sigma^2\mathbf{I}_n.$$

Pro chybový vektor pak platí

$$\mathsf{E}\,(\boldsymbol{\varepsilon}|\mathbb{X}) = \mathsf{E}\,(\boldsymbol{\varepsilon}) = \mathbf{0}_n, \qquad \mathrm{var}(\boldsymbol{\varepsilon}|\mathbb{X}) = \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n.$$

Odhad regresních koeficientů $\boldsymbol{\beta} \in \mathbb{R}^k$ je získán minimalizací L_2 ztrátové funkce (metoda nejmenších čtverců):

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^k} ||\,\boldsymbol{Y} - \mathbb{X}\mathbf{b}||^2 = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^k} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\top\mathbf{b})^2,$$

kde $||\cdot||$ je L_2 norma. Podle Gaussovy–Markovovy věty je vektor vyrovnaných hodnot $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$ nejlepším lineárním nestranným odhadem $\boldsymbol{Y}$.

Pro řadu výsledků je dále potřeba předpokládat normální lineární model

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

což je ekvivalentní předpokladu normality $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Pro normální lineární model je pak odhad regresních koeficientů $\widehat{\boldsymbol{\beta}}$ také maximálně věrohodným odhadem.

V případě normálního lineárního modelu lze odvodit přesné rozdělení odhadu $\widehat{\boldsymbol{\beta}}$ a na jeho základě zkonstruovat i přesné testy, konfidenční intervaly a predikční intervaly.

V případě obecného lineárního modelu lze za platnosti určitých předpokladů ukázat, že odhady regresních parametrů jsou konzistentní a mají asymptoticky normální rozdělení. Tyto vlastnosti lze dokázat, i pokud není splněn předpoklad na rovnost rozptylů (homoskedasticita) $\mathrm{var}(Y|\boldsymbol{X}) = \sigma^2$ představený výše, ale platí $\mathrm{var}(Y|\boldsymbol{X}) = \sigma^2(\boldsymbol{X})$, tedy rozptyl je nějakou funkcí regresorů (heteroskedasticita). V obou případech je potřeba existence druhého momentu ε.

Kromě předpokladů existence momentů je další nevýhodou lineární regrese citlivost L_2 ztrátové funkce na odlehlá pozorování. Metoda tak není příliš robustní, protože i jedno odlehlé pozorování může způsobit velké vychýlení odhadu.

2. L_1 regrese

2.1 Kvantily a ztrátová funkce

Nejprve si připomeneme definici kvantilu a shrneme si související základní poznatky, o které se následně opírá kvantilová regrese.

Definice 1. *Nechť X je náhodná veličina s distribuční funkcí $F(x) = P(X \leq x)$, $x \in \mathbb{R}$. Pro $0 < \alpha < 1$ definujeme kvantilovou funkci*

$$F^{-1}(\alpha) = inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

Pro pevné $\alpha \in (0,1)$ pak nazýváme $F^{-1}(\alpha)$ α-kvantilem náhodné veličiny X. Speciálně pro $\alpha = \frac{1}{2}$ nazýváme $F^{-1}(\frac{1}{2})$ mediánem náhodné veličiny X.

Poznámka. Pokud je distribuční funkce F rostoucí a spojitá, existuje právě jedno $x \in \mathbb{R}$ takové, že $F(x) = \alpha$, a kvantilová funkce je tedy inverzní funkcí F.

Definice 2. *Pro $\alpha \in (0,1)$ definujeme ztrátovou funkci*

$$\rho_\alpha(u) = u(\alpha - \mathbb{I}[u < 0]),\ u \in \mathbb{R}.$$

Poznámka. Pro $\alpha = \frac{1}{2}$ je

$$\rho_{\frac{1}{2}}(u) = \frac{1}{2}|u|$$

při minimalizaci ekvivalentní L_1 ztrátové funkci, která je obvykle definována jako absolutní hodnota, tedy

$$\arg\min_{u \in \mathbb{R}} \rho_{\frac{1}{2}}(u) = \arg\min_{u \in \mathbb{R}} |u|.$$

Kvantil lze identifikovat také minimalizací ztrátové funkce ρ_α, což si dokážeme v následujícím lemmatu.

Lemma 1. *Nechť X je náhodná veličina s distribuční funkcí $F(x) = P(X \leq x)$. Pak*

$$F^{-1}(\alpha) = \arg\min_{\theta \in \mathbb{R}} E[\rho_\alpha(X - \theta)].$$

Důkaz. Protože $\rho_\alpha(X)$ nezávisí na θ, platí

$$\arg\min_{\theta \in \mathbb{R}} \mathsf{E}\left[\rho_\alpha(X - \theta)\right] = \arg\min_{\theta \in \mathbb{R}} \mathsf{E}\left[\rho_\alpha(X - \theta) - \rho_\alpha(X)\right].$$

Ztrátová funkce ρ_α je diferencovatelná skoro všude, a lze tak její přírůstek $\rho_\alpha(X - \theta) - \rho_\alpha(X)$ vyjádřit jako integrál ze zobecněné derivace

$$\psi_\alpha(x) = \begin{cases} \rho'_\alpha(x) = \alpha\mathbb{I}[x > 0] + (1 - \alpha)\mathbb{I}[x < 0], & \text{pokud } x \neq 0, \\ 0, & \text{pokud } x = 0. \end{cases}$$

Tvrzení z tohoto důvodu budeme dokazovat pro $M(\theta) = \mathsf{E}\,[\rho_\alpha(X-\theta) - \rho_\alpha(X)]$ a můžeme psát

$$\begin{aligned} M(\theta) &= -\,\mathsf{E}\int_0^\theta \psi_\alpha(X-t)\mathrm{d}t = -\int_0^\theta \mathsf{E}\;\psi_\alpha(X-t)\mathrm{d}t \\ &= -\int_0^\theta \alpha\,\mathsf{P}(X>t) - (1-\alpha)\,\mathsf{P}(X<t)\mathrm{d}t \\ &= -\int_0^\theta \alpha - \alpha F(t) - (1-\alpha)F(t)\mathrm{d}t = -\alpha\theta + \int_0^\theta F(t)\mathrm{d}t. \end{aligned}$$

Funkce $M(\theta)$ je spojitá, protože ρ_α je spojitá funkce. Pro $\theta_1 < F^{-1}(\alpha)$ máme

$$M'(\theta_{1-}) = -\alpha + F(\theta_{1-}) \le -\alpha + F(\theta_1) < 0 \;\text{ a }\; M'(\theta_{1+}) = -\alpha + F(\theta_{1+}) < 0.$$

Dostáváme tedy, že $M(\theta)$ je klesající na intervalu $(-\infty, F^{-1}(\alpha))$. Analogicky pro $\theta_2 > F^{-1}(\alpha)$ máme

$$M'(\theta_{2-}) = -\alpha + F(\theta_{2-}) \ge -\alpha + F(F^{-1}(\alpha)) = 0 \;\text{ a }\; M'(\theta_{2+}) \ge 0,$$

tedy funkce $M(\theta)$ je neklesající na intervalu $(F^{-1}(\alpha), +\infty)$.

Dohromady dostáváme, že $F^{-1}(\alpha)$ je bodem lokálního minima $M(\theta)$, a tedy i $\mathsf{E}\,[\rho_\alpha(X-\theta)]$. Díky konvexitě ρ_α je pak $F^{-1}(\alpha)$ také bodem globálního minima. □

Pokud máme náhodný výběr $X_1, \ldots, X_n \sim F$, distribuční funkci F můžeme odhadnout pomocí empirické distribuční funkce

$$\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \le x],\; x \in \mathbb{R}$$

a výběrový kvantil je obvykle definován jako

$$\widehat{F_n^{-1}}(\alpha) = inf\{x \in \mathbb{R} : \widehat{F_n}(x) \ge \alpha\} \text{ pro } \alpha \in (0{,}1).$$

Potom $\widehat{F_n^{-1}}(\alpha) = X_{(k_\alpha)}$, kde

$$k_\alpha = \begin{cases} [\alpha n] + 1, & \alpha n \notin \mathbb{N}, \\ \alpha n, & \alpha n \in \mathbb{N}. \end{cases}$$

Díky Lemmatu 1 získáme výběrový kvantil také jako

$$\widehat{F_n^{-1}}(\alpha) = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(X_i - \theta).$$

Minimum není jednoznačné právě tehdy, pokud $n\alpha \in \mathbb{N}$ a zároveň platí $X_{(n\alpha)} < X_{(n\alpha+1)}$. V takovém případě funkce $\sum_{i=1}^n \rho_\alpha(X_i-\theta)$ nabývá minima na celém intervalu $[X_{(n\alpha)}, X_{(n\alpha+1)}]$ a jako $\widehat{F_n^{-1}}(\alpha)$ uvažujeme levý krajní bod tohoto intervalu. Speciálně pro případ mediánu to znamená, že řešení není jednoznačné, pokud máme sudý počet pozorování a $X_{\frac{n}{2}} < X_{\frac{n}{2}+1}$, tedy prostřední dvě pozorování v seřazeném náhodném výběru si nejsou rovna.

Minimalizační úlohu

$$\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} \rho_\alpha(X_i - \theta)$$

lze formulovat jako úlohu lineárního programování zavedením $2n$ pomocných nezáporných proměnných $\{u_i, v_i \in \mathbb{R}_+, i = 1, \ldots, n\}$, které reprezentují kladnou a zápornou část vektoru chyb, tedy

$$u_i = (X_i - \theta)_+, \qquad v_i = (X_i - \theta)_-.$$

Úloha lineárního programování je formulována jako

$$\min_{(\theta,\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}\times\mathbb{R}_+^{2n}} \alpha \sum_{i=1}^{n} u_i + (1-\alpha) \sum_{i=1}^{n} v_i \tag{2.1}$$

za podmínek

$$\begin{aligned} \theta + u_i - v_i &= X_i, \quad i = 1, \ldots, n, \\ u_i \geq 0, v_i &\geq 0, \quad i = 1, \ldots, n, \\ \theta &\in \mathbb{R}. \end{aligned}$$

Úlohu (2.1) lze vyřešit například simplexovým algoritmem, jehož základy jsou shrnuty například v kapitole 3.5, Dupačová a Lachout (2011). Formulace úlohy lineárního programování pro výběrový kvantil bude hrát zásadní roli při odvození asymptotického rozdělení odhadů koeficientů v L_1 regresním modelu díky specifickým vlastnostem optimálního řešení.

Základní vlastnosti výběrových kvantilů jsou uvedené v následující větě.

Věta 2. *Nechť $\alpha \in (0,1)$ a $X_1, \ldots, X_n$ je náhodný výběr se spojitou distribuční funkcí F.*

(i) $\widehat{F_n^{-1}}(\alpha)$ je konzistentní odhad $F^{-1}(\alpha)$, tedy

$$\widehat{F_n^{-1}}(\alpha) \xrightarrow{p} F^{-1}(\alpha), \ n \to +\infty.$$

(ii) Nechť existuje hustota f, která je na okolí $F^{-1}(\alpha)$ nenulová a spojitá. Potom platí

$$\sqrt{n}(\widehat{F_n^{-1}}(\alpha) - F^{-1}(\alpha)) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\alpha(1-\alpha)}{[f(F^{-1}(\alpha))]^2}\Big), \ n \to +\infty.$$

Důkaz. Důkaz obou tvrzení lze nalézt například v Bhattacharya, Lin a Patrangenaru (2016), kde je část konzistence výběrového kvantilu dokázána v kapitole 6.4 a jeho asymptotická normalita je dokázána v kapitole 6.7.

□

Speciálně je tedy výběrový medián $\widehat{F_n^{-1}}(\frac{1}{2})$ konzistentním odhadem mediánu s asymptotickým rozptylem $\frac{1}{[2f(x)]^2}$, kde $x = F_n^{-1}(\frac{1}{2})$ je medián. Konzistence je vlastnost, kterou požadujeme od každého rozumného bodového odhadu, a znalost asymptotického rozdělení nám pak umožňuje sestrojení intervalových odhadů a testů hypotéz. Bude tedy naším cílem odvodit obdobné tvrzení také pro odhady v lineárním regresním modelu.

2.2 Kvantilová regrese

Nyní se dostáváme k regresní úloze, pro kterou budeme používat značení a terminologii zavedenou v podkapitole 1.1. Uvažujeme model ve tvaru

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

ekvivalentně

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n. \tag{2.3}$$

Pro účely kvantilové regrese budeme předpokládat, že pro distribuční funkci stejně rozdělených a nezávislých náhodných chyb ε_i, $i = 1, \ldots, n$ platí $F_\varepsilon^{-1}(\alpha) = 0$, což je v modelu s absolutním členem bez újmy na obecnosti. Připomínáme, že uvažujeme pouze odezvy se spojitým rozdělením, a tedy i náhodné chyby mají spojité rozdělení.

Pro nezávislé a stejně rozdělené náhodné vektory $(Y_i, \boldsymbol{X}_i^\top)^\top, i = 1, \ldots, n$ rozdělené jako generický náhodný vektor $(Y, \boldsymbol{X}^\top)^\top$ definujeme regresní α-kvantil jako

$$\widehat{\boldsymbol{\beta}}_n(\alpha) = \arg\min_{\mathbf{b}\in\mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i - \boldsymbol{X}_i^\top \mathbf{b}).$$

Odhad $\widehat{\boldsymbol{\beta}}_n(\alpha)$ odhaduje parametr

$$\boldsymbol{\beta}(\alpha) = \arg\min_{\mathbf{b}\in\mathbb{R}^k} \mathsf{E}\ \rho_\alpha(Y - \boldsymbol{X}^\top \mathbf{b})$$

a z Lemmatu 1 plyne

$$\begin{aligned} \mathsf{E}\ \rho_\alpha(Y - \boldsymbol{X}^\top\boldsymbol{\beta}) &= \mathsf{E}\left\{\mathsf{E}\left[\rho_\alpha(Y - \boldsymbol{X}^\top\boldsymbol{\beta})|\boldsymbol{X}\right]\right\} \\ &\geq \mathsf{E}\left\{\mathsf{E}\left[\rho_\alpha(Y - F^{-1}_{Y|\boldsymbol{X}}(\alpha))|\boldsymbol{X}\right]\right\} = \mathsf{E}\ \rho_\alpha(Y - F^{-1}_{Y|\boldsymbol{X}}(\alpha)), \end{aligned}$$

kde $F^{-1}_{Y|\boldsymbol{X}}(\alpha)$ je podmíněný α-kvantil Y za podmínky regresorů $\boldsymbol{X}$. Tedy pokud platí $F^{-1}_{Y|\boldsymbol{X}}(\alpha) = \boldsymbol{X}^\top\boldsymbol{\beta}$, pak $\boldsymbol{\beta}(\alpha) = \boldsymbol{\beta}$. Model (2.2), resp. (2.3), lze tedy také zapsat ve tvaru

$$Y = \boldsymbol{X}^\top\boldsymbol{\beta} + \varepsilon,$$

kde $F_\varepsilon^{-1}(\alpha) = 0$.

Analogicky jako v (2.1) lze minimalizační úlohu

$$\min_{\mathbf{b}\in\mathbb{R}^k} \sum_{i=1}^{n} \rho_\alpha(Y_i - \boldsymbol{X}_i^\top \mathbf{b})$$

přeformulovat na úlohu lineárního programování

$$\min_{(\mathbf{b},\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^k\times\mathbb{R}_+^{2n}} \alpha\sum_{i=1}^{n} u_i + (1-\alpha)\sum_{i=1}^{n} v_i \tag{2.4}$$

s omezeními

$$\begin{aligned} \boldsymbol{X}_i^\top\mathbf{b} + u_i - v_i = Y_i, &\quad i = 1, \ldots, n, \\ u_i \geq 0, v_i \geq 0, &\quad i = 1, \ldots, n, \\ b_i \in \mathbb{R}, &\quad i = 1, \ldots, n, \end{aligned}$$

kde

$$u_i = (Y_i - \boldsymbol{X}_i^\top\mathbf{b})_+, \qquad v_i = (Y_i - \boldsymbol{X}_i^\top\mathbf{b})_-.$$

Ve zbytku práce se již budeme zabývat pouze speciálním případem pro medián s předpokladem $F_\varepsilon^{-1}(\frac{1}{2}) = 0$. Odhad $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_n(\frac{1}{2})$ bude značit L_1 odhad regresních parametrů

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{b}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n} |Y_i - \boldsymbol{X}_i^\top \mathbf{b}|. \tag{2.5}$$

2.3 Značení a pomocná tvrzení

Nechť $N = \{1, \ldots, n\}$ a $\mathcal{J}$ značí množinu všech k-prvkových podmnožin N. Pro $H \in \mathcal{J}$ definujeme relativní doplněk $\bar{H} = N \setminus H$.

Jako $\boldsymbol{Y}(H)$ budeme značit k-rozměrný vektor s prvky $\{Y_i, i \in H\}$. Obdobně $\mathbb{X}(H)$ bude značit $k\times k$ matici s řádky $\{\boldsymbol{X}_i^\top, i \in H\}$ a $\mathbb{X}(\bar{H})$ bude značit $(n-k)\times k$ matici s řádky $\{\boldsymbol{X}_i^\top, i \in \bar{H}\}$.

Označme $\mathcal{H} = \{H \in \mathcal{J} \mid rank(\mathbb{X}(H)) = k\}$ a jako B budeme značit množinu řešení úlohy (2.5).

Výše zavedené značení závisí na regresní matici $\mathbb{X}$, případně na jejích rozměrech n a k. Tuto závislost ve značení pro přehlednost nezohledňujeme, regresní matice a její rozměry by vždy měly být jasně dané z kontextu.

Z vlastností řešení úlohy lineárního programování díky formulaci (2.4) plyne speciální tvar odhadu regresních koeficientů popsaný v Lemmatu 3.

Lemma 3. *Nechť $rank(\mathbb{X}) = k$. Pak existuje alespoň jedno řešení úlohy (2.5) $\widehat{\boldsymbol{\beta}}_n \in B$ ve tvaru*

$$\widehat{\boldsymbol{\beta}}_n = \mathbb{X}(H)^{-1}\boldsymbol{Y}(H)$$

pro nějaké $H \in \mathcal{H}$. Množina B je konvexní obal všech řešení tohoto tvaru.

Důkaz. Tvrzení plyne z formulace problému jako úlohy lineárního programování (2.4). Minimalizační úlohu

$$\min_{(\mathbf{b},\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^k\times\mathbb{R}_+^{2n}} \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i,$$

s omezeními

$$\begin{aligned} \boldsymbol{X}_i^\top \mathbf{b} + u_i - v_i = Y_i, &\quad i = 1, \ldots, n, \\ u_i \geq 0, v_i \geq 0, &\quad i = 1, \ldots, n, \\ b_i \in \mathbb{R}, &\quad i = 1, \ldots, n \end{aligned}$$

lze převést na duální úlohu

$$\max_{\boldsymbol{d}\in\mathbb{R}^n} \sum_{i=1}^{n} Y_i d_i,$$

s omezeními

$$\begin{aligned} \sum_{i=1}^{n} X_{ij} d_i = 0, &\quad j = 1, \ldots, k, \\ -1 \leq d_i \leq 1, &\quad i = 1, \ldots, n. \end{aligned}$$

Pokud zavedeme $f_i \equiv d_i + 1, i = 1, \ldots, n$, dostaneme reparametrizovanou úlohu

$$\max_{\boldsymbol{f} \in \mathbb{R}^n} \sum_{i=1}^{n} Y_i f_i + \sum_{i=1}^{n} Y_i,$$

s omezeními

$$\sum_{i=1}^{n} X_{ij} f_i = \sum_{i=1}^{n} X_{ij}, \quad j = 1, \ldots, k,$$
$$0 \leq f_i \leq 2, \quad i = 1, \ldots, n.$$

Z teorie lineárního programování víme, že existuje optimální bazické řešení, pro které je právě k proměnných f_i nenulových. Označíme $H = \{i \in \{1, \ldots, n\} : f_i \neq 0\}$. Pak z podmínek komplementarity dostaneme pro optimální řešení $\widehat{\boldsymbol{\beta}}_n$

$$\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n = Y_i, \quad i \in H,$$

z čehož plyne

$$\mathbb{X}(H)\widehat{\boldsymbol{\beta}}_n = \boldsymbol{Y}(H).$$

Optimální bazické řešení je ekvivalentní krajnímu řešení úlohy a další řešení lze u úlohy lineárního programování s omezenou množinou přípustných řešení získat jen jejich konvexní kombinací, čímž dostáváme, že B je konvexní obal všech řešení tvaru $\mathbb{X}(H)\widehat{\boldsymbol{\beta}}_n = \boldsymbol{Y}(H)$.

□

Poznámka. Odhad regresních koeficientů metodou nejmenších čtverců v lineárním modelu představeném v Kapitole 1 lze zapsat jako $\widehat{\boldsymbol{\beta}}_{LS} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{Y}$, je tedy lineární kombinací všech pozorování závislé veličiny. Oproti tomu z Lemmatu 3 plyne, že L_1 odhad regresních koeficientů je lineární kombinací pouze k různých pozorování závislé veličiny, nicméně všechna pozorování jsou potřeba k výběru těchto pozorování, nelze tedy říci, že by odhad na zbytku pozorování nezávisel.

Na tvrzení v Lemmatu 3 lze také pohlížet jako na analogii k výběrovému mediánu, který leží v prostředním pozorování výběru, pokud máme lichý počet pozorování. V případě lichého počtu pozorování leží mezi dvěma prostředními pozorováními, což odpovídá konvexnímu obalu těchto dvou pozorování. Regresní medián pak leží v nadrovině určené k pozorováními, které nejlépe reprezentují podmíněný medián.

Další poznatky k formulaci úlohy lineárního programování a jejímu řešení lze nalézt například ve Wagner (1959), odkud je převzat důkaz Lemmatu 3 a kde je také popsán způsob řešení úlohy simplexovým algoritmem.

Definice 3. *K-dimenzionální uzavřenou nadkrychli se středem $\boldsymbol{\delta} \in \mathbb{R}^k$ a rozměrem $\epsilon > 0$ definujeme jako*

$$C[\boldsymbol{\delta}, \epsilon] = \{\boldsymbol{c} \in \mathbb{R}^k : \max_{i=1,\ldots,k} |c_i - \delta_i| \leq \epsilon\}.$$

Analogicky definujeme k-dimenzionální otevřenou nadkrychli se středem $\boldsymbol{\delta} \in \mathbb{R}^k$ a rozměrem $\epsilon > 0$

$$C(\boldsymbol{\delta}, \epsilon) = \{\boldsymbol{c} \in \mathbb{R}^k : \max_{i=1,\ldots,k} |c_i - \delta_i| < \epsilon\}.$$

Dále definujeme vektorovou funkci $\boldsymbol{\zeta}(H, \boldsymbol{v}) = (\zeta_1, \ldots, \zeta_{|\bar{H}|}) : \mathcal{J} \times \mathbb{R}^k \to \mathbb{R}$ *takovou, že*

$$\boldsymbol{\zeta}(H, \boldsymbol{v}) = \sum_{i \in \bar{H}} \zeta_i(H, \boldsymbol{v}) = \sum_{i \in \bar{H}} \operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H); -\boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} \boldsymbol{v}) \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1},$$

kde

$$\operatorname{sgn}^*(u, w) = \begin{cases} \operatorname{sgn}(u), & \text{pokud } u \neq 0, \\ \operatorname{sgn}(w), & \text{pokud } u = 0. \end{cases}$$

Absolutní hodnota $|u|$ má derivaci $\operatorname{sgn}(u)$ pro $u \neq 0$, ale pro $u = 0$ derivace neexistuje, proto se nám bude hodit funkce $\operatorname{sgn}^*(u,w)$ pro její dodefinování. Funkce $\boldsymbol{\zeta}(H, \boldsymbol{v})$ pak slouží pro formulaci ekvivalentní podmínky existence a jednoznačnosti optimálního řešení v následujícím lemmatu.

Lemma 4. *Pro* $H \in \mathcal{H}$ *existuje* $\boldsymbol{b}(H) \equiv \mathbb{X}(H)^{-1}\boldsymbol{Y}(H) \in B$ *právě tehdy, když* $\boldsymbol{\zeta}(H, \boldsymbol{v}) \in C[\mathbf{0},1]$ *skoro jistě pro každé* $\boldsymbol{v} \neq \mathbf{0}$ *a* $\boldsymbol{b}(H) = B$*, tedy je jednoznačné, právě tehdy, když* $\boldsymbol{\zeta}(H, \boldsymbol{v}) \in C(\mathbf{0},1)$ *skoro jistě.*

Důkaz. Obecná funkce $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^k$ má v bodě $\boldsymbol{a} \in \mathbb{R}^k$ lokální minimum právě tehdy, když

$$\lim_{h \to 0^+} \frac{f(\boldsymbol{a} + h\boldsymbol{w}) - f(\boldsymbol{a})}{h} \geq 0, \ \forall \boldsymbol{w} \in \mathbb{R}^k, \ \boldsymbol{w} \neq \mathbf{0},$$

za předpokladu, že tato limita existuje. Znamená to, že přírůstky funkce f jsou kladné ve všech směrech $\boldsymbol{w}$ alespoň na nějakém okolí bodu $\boldsymbol{a}$, což je zřejmě ekvivalentní definici lokálního minima.

Pro funkci $\sum_{i=1}^n |Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}|$ nalezneme tuto limitu v $\boldsymbol{b}(H)$ pro jednotlivé sčítance $|Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}|$ a ve všech směrech $\boldsymbol{w} \in \mathbb{R}^k$. Pokud $Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H) = 0$, pak

$$\lim_{h \to 0^+} \frac{|Y_i - \boldsymbol{X}_i^\top(\boldsymbol{b}(H) + h\boldsymbol{w})| - |Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H)|}{h} = \lim_{h \to 0^+} \frac{|-\boldsymbol{X}_i^\top \boldsymbol{w}|h}{h} = |-\boldsymbol{X}_i^\top \boldsymbol{w}|.$$

Pokud $Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H) \neq 0$, dostáváme

$$-\boldsymbol{X}_i^\top \boldsymbol{w} \lim_{h \to 0^+} \frac{|Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H) - h\boldsymbol{X}_i^\top \boldsymbol{w}| - |Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H)|}{-h\boldsymbol{X}_i^\top \boldsymbol{w}} = -\operatorname{sgn}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H)) \boldsymbol{X}_i^\top \boldsymbol{w},$$

kde člen s limitou je derivace absolutní hodnoty v bodě $Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H)$. Protože $|-\boldsymbol{X}_i^\top \boldsymbol{w}| = -\operatorname{sgn}(-\boldsymbol{X}_i^\top \boldsymbol{w}) \boldsymbol{X}_i^\top \boldsymbol{w}$, můžeme díky zavedenému značení dohromady psát

$$\lim_{h \to 0^+} \frac{|Y_i - \boldsymbol{X}_i^\top(\boldsymbol{b}(H) + h\boldsymbol{w})| - |Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H)|}{h} = -\operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H); -\boldsymbol{X}_i^\top \boldsymbol{w}) \boldsymbol{X}_i^\top \boldsymbol{w}.$$

Označíme

$$\boldsymbol{\psi}(\boldsymbol{b}(H); \boldsymbol{w}) = -\sum_{i=1}^n \operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H); -\boldsymbol{X}_i^\top \boldsymbol{w}) \boldsymbol{X}_i^\top \boldsymbol{w},$$

pak díky konvexitě absolutní hodnoty (lokální minimum je i globální minimum) je existence $\boldsymbol{b}(H)$ ekvivalentní s podmínkou

$$\boldsymbol{\psi}(\boldsymbol{b}(H); \boldsymbol{w}) \geq 0, \, \forall \boldsymbol{w} \in \mathbb{R}^k, \, \boldsymbol{w} \neq \mathbf{0}.$$

Díky vlastnosti $Y_i = \boldsymbol{X}_i^\top \boldsymbol{b}(H)$ pro $i \in H$ platí

$$\boldsymbol{\psi}(\boldsymbol{b}(H); \boldsymbol{w}) = \sum_{i \in H} \operatorname{sgn}(\boldsymbol{X}_i^\top \boldsymbol{w}) \boldsymbol{X}_i^\top \boldsymbol{w} - \sum_{i \in \bar{H}} \operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H); -\boldsymbol{X}_i^\top \boldsymbol{w}) \boldsymbol{X}_i^\top \boldsymbol{w}.$$

Dále pokud položíme $\boldsymbol{v} = \mathbb{X}(H)\boldsymbol{w}$ (a tedy $\boldsymbol{w} = \mathbb{X}(H)^{-1}\boldsymbol{v}$), dostaneme, že

$$\boldsymbol{\psi}(\boldsymbol{b}(H); \boldsymbol{w}) \geq 0$$

je ekvivalentní s

$$\sum_{j=1}^{k} |v_j| - \boldsymbol{\zeta}(H, \boldsymbol{v}) \boldsymbol{v} \geq 0,$$

pro všechny $\boldsymbol{v} \neq \mathbf{0}$. Protože díky předpokladu spojitého rozdělení odezvy platí $Y_i \neq \boldsymbol{X}_i^\top \boldsymbol{b}(H)$, $i \in \bar{H}$ *s.j.*, dostáváme

$$\boldsymbol{\zeta}(H, \boldsymbol{v}) = \sum_{i \in \bar{H}} \operatorname{sgn}(Y_i - \boldsymbol{X}_i^\top \mathbf{b}(H)) \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} \; s.j.,$$

tedy $\boldsymbol{\zeta}(H) \equiv \boldsymbol{\zeta}(H, \boldsymbol{v})$ již nezávisí na $\boldsymbol{v}$ a vztah

$$\sum_{j=1}^{k} |v_j| - \boldsymbol{\zeta}(H) \boldsymbol{v} \geq 0$$

je dále ekvivalentní $\boldsymbol{\zeta}(H) \in C[\mathbf{0},1]$. Dohromady platí $\boldsymbol{\zeta}(H,\boldsymbol{v}) \in C[\mathbf{0},1]$ skoro jistě.

Pokud výše nahradíme neostré nerovnosti ostrými a uzavřenou nadkrychli $C[\mathbf{0},1]$ otevřenou nadkrychlí $C(\mathbf{0},1)$, dostaneme tvrzení pro jednoznačnost.

□

Důsledek. Uvažujme jednorozměrný případ, kde $X_i \equiv 1, i = 1, \ldots, n$, regresní koeficient $\boldsymbol{\beta} \equiv \beta \in \mathbb{R}$ je jednorozměrný a H je jednoprvková množina. Pak $\widehat{\beta} = \boldsymbol{Y}(H) = Y_j$ pro nějaké $j \in \{1, \ldots, n\}$ je výběrový medián právě tehdy, když

$$-1 \leq \sum_{i \in \bar{H}} \operatorname{sgn}^*(Y_i - \boldsymbol{Y}(H); -w) \leq 1$$

pro každé $w \neq 0$, což je ekvivalentní

$$-1 \leq \sum_{i \neq j} \operatorname{sgn}^*(Y_i - Y_j; -w) \leq 1,$$

tedy počet pozorování menších než Y_j a počet pozorování větších než Y_j se liší maximálně o 1. Pokud má být medián jednoznačný, platí

$$-1 < \sum_{i \neq j} \operatorname{sgn}^*(Y_i - Y_j; -w) < 1,$$

tedy počet pozorování větších a menších než Y_j je stejný.

Dále si dokážeme některé ekvivarianční vlastnosti L_1 odhadu, které jsou důležité, pokud uvažujeme transformovanou odezvu či regresní matici a chceme znát vztah mezi řešením původní úlohy a řešením transformované úlohy. Mezi základní možné transformace patří přeškálování odezvy či změna parametrizace regresního prostoru. Speciální vlastností mediánové regrese je možnost daným způsobem změnit některá pozorování, aniž by se změnilo řešení úlohy.

Lemma 5. *Pokud* $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y},\mathbb{X}) \in B(\boldsymbol{Y},\mathbb{X})$*, pak platí následující vztahy mezi řešeními transformovaných úloh:*

(i) $\widehat{\boldsymbol{\beta}}_n(\lambda\boldsymbol{Y},\mathbb{X}) = \lambda\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y},\mathbb{X}), \quad \lambda \in \mathbb{R},$

(ii) $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}+\mathbb{X}\boldsymbol{\gamma},\mathbb{X}) = \widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y},\mathbb{X}) + \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \in \mathbb{R}^k,$

(iii) $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y},\mathbb{X}\mathbb{A}) = \mathbb{A}^{-1}\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y},\mathbb{X})$*, kde* $\mathbb{A}_{k\times k}$ *je regulární matice,*

(iv) $\widehat{\boldsymbol{\beta}}_n(\mathbb{X}\widehat{\boldsymbol{\beta}}_n+\mathbb{D}\boldsymbol{u},\mathbb{X}) = \widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y},\mathbb{X})$*, kde* $\mathbb{D}_{n\times n}$ *je diagonální matice s nezápornými prvky a* $\boldsymbol{u} = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_n$*.*

Důkaz. Označíme

$$\xi(\boldsymbol{b};\boldsymbol{Y},\mathbb{X}) = \sum_{i=1}^{n} |Y_i - \boldsymbol{X}_i^\top \mathbf{b}|,$$

pak platí

(i) $|\lambda|\xi(\boldsymbol{b};\boldsymbol{Y},\mathbb{X}) = \xi(\lambda\boldsymbol{b};\lambda\boldsymbol{Y},\mathbb{X}),$

(ii) $\xi(\boldsymbol{b};\boldsymbol{Y},\mathbb{X}) = \xi(\boldsymbol{b}+\boldsymbol{\gamma};\boldsymbol{Y}+\mathbb{X}\boldsymbol{\gamma},\mathbb{X}),$

(iii) $\xi(\boldsymbol{b};\boldsymbol{Y},\mathbb{X}) = \xi(\mathbb{A}^{-1}\boldsymbol{b};\boldsymbol{Y},\mathbb{X}\mathbb{A}).$

Pro důkaz (iv) využijeme Lemma 4, podle kterého pro $\widehat{\boldsymbol{\beta}}_n \in B(\boldsymbol{Y},\mathbb{X})$ platí

$$-\sum_{i=1}^{n} \operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n; -\boldsymbol{X}_i^\top\boldsymbol{w})\boldsymbol{X}_i^\top\boldsymbol{w} \geq 0, \quad \boldsymbol{w} \in \mathbb{R}^k.$$

Dále platí

$$\begin{aligned}\operatorname{sgn}^*(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n + d_i(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n) - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n; -\boldsymbol{X}_i^\top\boldsymbol{w})\boldsymbol{X}_i^\top\boldsymbol{w} \\ = \operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n; -\boldsymbol{X}_i^\top\boldsymbol{w})\boldsymbol{X}_i^\top\boldsymbol{w}\end{aligned}$$

pro $d_i \geq 0$, z čehož již plyne požadované tvrzení.

□

Poznámka. Vlastnosti (i)–(iii) z Lemmatu 5 platí i pro odhad metodou nejmenších čtverců, ale jsou neobvyklé u robustních odhadů. Vlastnosti (i) a (ii) jsou vlastnostmi ekvivariance vůči afinní transformaci, vlastnost (iii) je ekvivariancí vůči reparametrizaci.

Vlastnost (iv) je zobecněním invariance mediánu. Tato vlastnost říká, že L_1 odhad se nezmění při změně některých pozorování, pokud upravená pozorování zůstávají na „správné“ straně nadroviny. Z této vlastnosti vyplývá robustnost odhadu vůči odlehlým pozorováním, což bude demonstrováno na příkladu níže.

Pro odhad metodou nejmenších čtverců tato vlastnost neplatí, což plyne z tvaru odhadu $\widehat{\boldsymbol{\beta}}_{LS} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}$, který závisí na všech hodnotách odezvy. Oproti tomu L_1 odhad $\widehat{\boldsymbol{\beta}}_n = \mathbb{X}(H)^{-1}\boldsymbol{Y}(H)$ přímo závisí pouze na k hodnotách odezvy. Zbylé hodnoty pouze určují tvar H a lze je tedy změnit takovým způsobem, aby odhad v transformované úloze vycházel ze stejné množiny H, viz Obrázek 2.1.

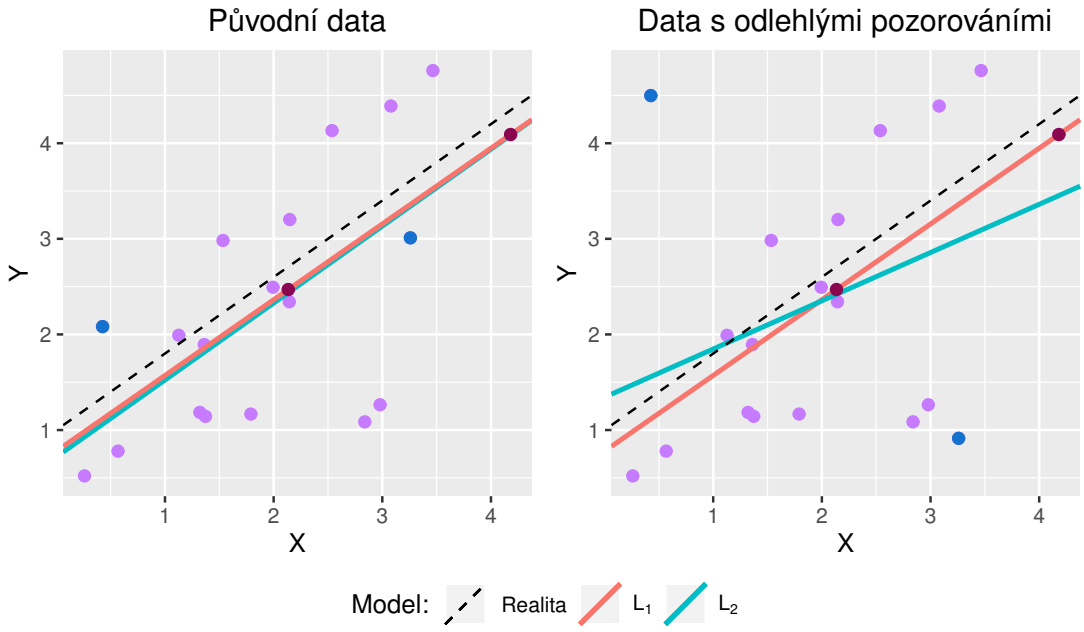
Příklad. Uvažujeme náhodný výběr $(Y_i, X_i)^\top, i = 1, \ldots, n$, pro který odhadneme regresní model s absolutním členem a jedním regresorem (tedy $k = 2$):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$$

pomocí L_1 a L_2 ztrátových funkcí. V prostředí R (R Core Team, 2019) vygenerujeme náhodný výběr pro $\beta_0 = 1$, $\beta_1 = 0.8$, $n = 20$ a $\varepsilon_i \sim \mathcal{N}(0,1)$. Následně upravíme dvojici pozorování dle části (iv) Lemmatu 5. Diagonální matice $\mathbb{D}$ má v tomto případě na diagonále samé jedničky kromě dvou pozic, kde jsou kladná čísla větší než jedna. Pozorování odpovídající jednotkovým pozicím zůstanou nezměněna, zatímco zbylá dvě pozorování posuneme dále od L_1 regresní přímky. Vytvořili jsme tak dvě odlehlá pozorování.

V Tabulce 2.1 jsou shrnuty L_1 a L_2 odhady regresních koeficientů na původních i upravených datech a na Obrázku 2.1 jsou vykreslena původní i upravená data s L_1 a L_2 regresními přímkami. Upravená pozorování jsou v obou grafech zvýrazněna modře.

Z grafů i tabulky je zřejmé, že L_1 a L_2 odhady jsou na původních datech téměř identické, od skutečnosti (černá přerušovaná čára) se výrazněji liší odhady interceptu, ale mějme na paměti, že máme poměrně malou velikost výběru. Pokud však vychýlíme dvě pozorování, L_2 odhady se výrazně změní. Oproti tomu L_1 odhady zůstávají stejné díky tomu, že množina H se nezměnila. Pozorování odpovídající množině H jsou tmavě červeně zvýrazněné body na L_1 regresní přímce na Obrázku 2.1.



Obrázek 2.1: Srovnání regresních přímek na původních a upravených datech.

Tabulka 2.1: Odhady regresních koeficientů na původních a upravených datech.

		β_0	β_1
Skutečná hodnota parametru		1	0.8
Odhad z původních dat	L_1	0.78	0.79
	L_2	0.72	0.80
Odhad z upravených dat	L_1	0.78	0.79
	L_2	1.34	0.50

Kromě upravených pozorování jsou na Obrázku 2.1 zvýrazněny také dva body ležící na L_1 regresní přímce. Zde je demonstrována vlastnost z Lemmatu 3 pro $k = 2$.

Při dalších výpočtech se nám přijde vhod následující identita pro determinant matice $\mathbb{X}^\top\mathbb{X}$.

Lemma 6. *Pro regresní matici $\mathbb{X}$ platí*

$$|\mathbb{X}^\top\mathbb{X}| = \sum_{H\in\mathcal{H}} |\mathbb{X}(H)|^2.$$

Důkaz. Dle Rao (1973, str. 32) pro determinant $n \times k, n \geq k$ matice platí

$$|\mathbb{X}^\top\mathbb{X}| = \sum_{H\in\mathcal{J}} |\mathbb{X}(H)|^2,$$

tedy determinant $k \times k$ matice $\mathbb{X}^\top\mathbb{X}$ je součtem čtverců determinantů všech $k \times k$ podmatic matice $\mathbb{X}$. Tvrzení pak plyne z toho, že pro $H \in \mathcal{J} \setminus \mathcal{H}$ je matice $\mathbb{X}$ singulární, a tudíž je její determinant nulový.

□

2.4 Asymptotické rozdělení L_1 odhadu regresních koeficientů

Pro data velikosti $n \in \mathbb{N}$ si označíme regresní matici

$$\mathbb{X}_n = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}.$$

Nyní můžeme dokázat větu o asymptotickém normálním rozdělení L_1 odhadu regresních koeficientů v lineárním modelu, která je zobecněním bodu (ii) Věty 2 a kterou poprvé zformulovali a dokázali Bassett a Koenker (1978).

Věta 7. *Nechť $\widehat{\boldsymbol{\beta}}_n, n \in \mathbb{N}$ je posloupnost řešení úlohy (2.5) a platí:*

(i) F_ε je spojitá a má spojitou a kladnou hustotu f v mediánu, tedy $f(0) > 0$;

(ii) $\lim_{n\to\infty} \frac{1}{n}\mathbb{X}_n^\top \mathbb{X}_n = \mathbb{Q}$, *kde* $\mathbb{Q}$ *je* $k \times k$ *pozitivně definitní matice;*

(iii) $||\mathbb{X}_n||_\infty = \max_{\substack{i=1,\dots,n,\\ j=1,\dots,k}} |X_{ij}| = o(\sqrt{n})$, $n \to +\infty$.

Potom

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_k\Big(0, \frac{1}{[2f(0)]^2}\mathbb{Q}^{-1}\Big),\ n \to +\infty.$$

Důkaz. Větu dokážeme ve dvou krocích:

(1) Odvodíme přesné rozdělení náhodného k-rozměrného vektoru $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, které bude vyjádřeno pomocí hustoty $\phi_n(\boldsymbol{\delta})$, $\boldsymbol{\delta} = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, tvaru

$$\phi_n(\boldsymbol{\delta}) = n^{-\frac{k}{2}} \sum_{H\in\mathcal{H}} |\mathbb{X}(H)| \prod_{i\in H} f(n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta})\, \mathsf{P}(\boldsymbol{Z}_n(\boldsymbol{\delta}, H) \in C(\mathbf{0},1)), \tag{2.6}$$

kde

$$\boldsymbol{Z}_n(\boldsymbol{\delta}, H) = \sum_{i\in\bar{H}} \boldsymbol{z}_i(\boldsymbol{\delta}, H) = \sum_{i\in\bar{H}} \operatorname{sgn}(u_i - n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta})\boldsymbol{X}_i^\top\mathbb{X}(H)^{-1}$$

a $u_i = Y_i - \boldsymbol{X}_i\widehat{\beta}$ jsou rezidua.

(2) Ukážeme, že

$$\phi_n(\boldsymbol{\delta}) \xrightarrow{n\to+\infty} \frac{|[2f(0)]^2\mathbb{Q}|^{\frac{1}{2}}}{(2\pi)^{\frac{K}{2}}} \exp\Big(-\frac{1}{2}[2f(0)]^2\boldsymbol{\delta}^\top\mathbb{Q}\boldsymbol{\delta}\Big), \tag{2.7}$$

tedy že přesné rozdělení konverguje pro $n \to \infty$ k požadované k-rozměrné normální hustotě.

Z Scheffého věty o konvergenci hustot pak dostaneme požadovanou konvergenci v distribuci.

(1) V této části důkazu je $n \in \mathbb{N}$ pevně dané a při značení jej proto vynecháváme. Označíme

$$\begin{aligned}\boldsymbol{\delta}^* &= \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta},\\ \boldsymbol{b}(H) &= \mathbb{X}(H)^{-1}\boldsymbol{Y}(H),\\ \boldsymbol{d}(H) &= \boldsymbol{b}(H) - \boldsymbol{\beta} = \mathbb{X}(H)^{-1}\boldsymbol{u}(H),\end{aligned}$$

kde

$$\boldsymbol{u}(H) = \mathbb{X}(H)\boldsymbol{b}(H) - \mathbb{X}(H)\boldsymbol{\beta} = \boldsymbol{Y}(H) - \mathbb{X}(H)\boldsymbol{\beta}$$

$\implies Y_i - \boldsymbol{X}_i^\top\boldsymbol{b}(H) = u_i - \boldsymbol{X}_i^\top\boldsymbol{d}(H)$, kde u_i je i-tá složka vektoru $\boldsymbol{u}(H)$.

Uvažujme jevy

$$\begin{aligned}E_1(H, \boldsymbol{\delta}, \epsilon) &= \{\boldsymbol{u} \in \mathbb{R}^n \mid \boldsymbol{d}(H) \in C(\boldsymbol{\delta}, \epsilon)\},\ \epsilon > 0,\\ E_2(H) &= \{\boldsymbol{u} \in \mathbb{R}^n \mid \boldsymbol{\zeta}(H, \boldsymbol{v}) \in C(\mathbf{0}, 1), \forall \boldsymbol{v} \neq \mathbf{0}\}.\end{aligned}$$

Potom z Lemmat 3 a 4

$$\mathsf{P}(\boldsymbol{\delta}^* \in C(\boldsymbol{\delta}, \epsilon)) = \mathsf{P}\Big(\bigcup_{H\in\mathcal{H}} [E_1(H, \boldsymbol{\delta}, \epsilon) \cap E_2(H)]\Big), \tag{2.8}$$

protože pak pro každé $\boldsymbol{u}(H) \in E_2(H)$ existuje právě jedno $\widehat{\boldsymbol{\beta}}$ takové, že

$$\widehat{\boldsymbol{\beta}} = \mathbb{X}(H)^{-1}\boldsymbol{Y}(H) = \mathbb{X}(H)^{-1}\boldsymbol{u}(H) + \boldsymbol{\beta},$$

a tedy

$$\boldsymbol{d}(H) = \mathbb{X}(H)^{-1}\boldsymbol{u}(H) = \mathbb{X}(H)^{-1}\boldsymbol{Y}(H) - \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\delta}^*,$$

z čehož díky $E_1(H, \boldsymbol{\delta}, \epsilon)$ plyne, že $\boldsymbol{\delta}^* \in C(\boldsymbol{\delta}, \epsilon)$ pro dané $H \in \mathcal{H}$.
Položíme $M = k||\mathbb{X}||_\infty$ a definujeme jev

$$E_3(H, \boldsymbol{\delta}, \epsilon) = \{\boldsymbol{u} \in R^n \mid |u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}| > \epsilon M \ \ \forall i \in \bar{H}\}, \ \epsilon > 0. \tag{2.9}$$

Zřejmě platí

$$\mathsf{P}(E_1 \cap E_2) = \mathsf{P}(E_1 \cap E_2 \cap E_3) + \mathsf{P}(E_1 \cap E_2 \cap (\mathbb{R}^n \setminus E_3)). \tag{2.10}$$

Jev $E_1(H, \boldsymbol{\delta}, \epsilon)$ implikuje $|u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}| < \epsilon M$ pro každé $i \in H$, z čehož plyne, že pro $H \neq H'$, $i \in H \setminus H'$ má platit

$$\begin{aligned} |u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}| &< \epsilon M \text{ dle } E_1, \\ |u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}| &> \epsilon M \text{ dle } E_3. \end{aligned}$$

Obdobný argument lze udělat pro $i \in H' \setminus H$, a dostáváme tedy, že

$$E_1(H, \boldsymbol{\delta}, \epsilon) \cap E_3(H', \boldsymbol{\delta}, \epsilon) = \emptyset.$$

Pak lze psát

$$\begin{aligned} \mathsf{P}(\bigcup_{H \in \mathcal{H}} E_1 \cap E_2 \cap E_3) &= \sum_{H \in \mathcal{H}} \mathsf{P}(E_1 \cap E_2 \cap E_3) \\ &= \sum_{H \in \mathcal{H}} \mathsf{P}(E_2 | E_1 \cap E_3)\, \mathsf{P}(E_3 | E_1)\, \mathsf{P}(E_1). \end{aligned} \tag{2.11}$$

Položíme

$$E_2'(H, \delta) = \{\boldsymbol{u} \in \mathbb{R}^n \mid \boldsymbol{Z}_n(\sqrt{n}\boldsymbol{\delta}, H) \in C(\boldsymbol{0}, 1)\}$$

a ukážeme, že

$$\mathsf{P}(E_2') = \lim_{\epsilon \to 0_+} \mathsf{P}(E_2 | E_1 \cap E_3). \tag{2.12}$$

Nechť $i \in \bar{H}$ a $\boldsymbol{u} \in E_1 \cap E_3$, pak dostaneme

$$\begin{aligned} Y_i - \boldsymbol{X}_i^\top \boldsymbol{b}(H) &= u_i - \boldsymbol{X}_i^\top \boldsymbol{d}(H) \\ &= u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta} + \boldsymbol{X}_i^\top (\boldsymbol{\delta} - \boldsymbol{d}(H)) \xrightarrow{\epsilon \to 0_+} u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}, \end{aligned}$$

protože $\boldsymbol{\delta} - \boldsymbol{d}(H) \xrightarrow{\epsilon \to 0_+} \boldsymbol{0}$ díky E_1, a dále platí, že $u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta} \neq 0$, protože $|u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}| > \epsilon M > 0$ z E_3. Potom můžeme zjednodušit

$$\boldsymbol{\zeta}(H, \boldsymbol{v}) = \sum_{i \in \bar{H}} \operatorname{sgn}^*(Y_i - \boldsymbol{X}_i^\top \mathbf{b}(H); -\boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}\boldsymbol{v}) \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}$$

pro $\boldsymbol{u} \in [E_2|E_1 \cap E_3]$ na

$$\boldsymbol{Z}_n(\sqrt{n}\boldsymbol{\delta}, H) = \sum_{i \in \bar{H}} \operatorname{sgn}(u_i - \boldsymbol{X}_i^\top \boldsymbol{\delta}) \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}.$$

Dále protože M je kladné reálné číslo, z definice E_3 v (2.9) plyne

$$\lim_{\epsilon \to 0_+} \mathsf{P}(E_3(H, \boldsymbol{\delta}, \epsilon)) = 1,$$

díky čemuž z (2.10), (2.11) a (2.12) dostáváme

$$\begin{aligned}
\lim_{\epsilon \to 0_+} \mathsf{P}\Big(\bigcup_{H \in \mathcal{H}} E_1 \cap E_2 \Big) &= \lim_{\epsilon \to 0_+} \Big[\mathsf{P}\Big(\bigcup_{H \in \mathcal{H}} E_1 \cap E_2 \cap E_3 \Big) \\
&\qquad + \mathsf{P}\Big(\bigcup_{H \in \mathcal{H}} E_1 \cap E_2 \cap (\mathbb{R}^n \setminus E_3) \Big) \Big] \\
&= \sum_{H \in \mathcal{H}} \lim_{\epsilon \to 0_+} \mathsf{P}(E_1 \cap E_2 \cap E_3) \\
&= \sum_{H \in \mathcal{H}} \lim_{\epsilon \to 0_+} \mathsf{P}(E_2 | E_1 \cap E_3)\, \mathsf{P}(E_3 | E_1)\, \mathsf{P}(E_1) \\
&= \sum_{H \in \mathcal{H}} \mathsf{P}(E_2')\, \lim_{\epsilon \to 0_+} E_1(H, \boldsymbol{\delta}, \epsilon). \qquad (2.13)
\end{aligned}$$

Nechť $\lambda(\cdot)$ je Lebesgueova míra. Potom z (2.8) a (2.13) dostáváme

$$\begin{aligned}
\lim_{\epsilon \to 0_+} \frac{\mathsf{P}(\boldsymbol{\delta}^* \in C(\boldsymbol{\delta}, \epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} &= \sum_{H \in \mathcal{H}} \mathsf{P}(E_2')\, \lim_{\epsilon \to 0_+} \frac{\mathsf{P}(E_1(H, \boldsymbol{\delta}, \epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} \\
&= \sum_{H \in \mathcal{H}} \mathsf{P}(\boldsymbol{Z}_n(\sqrt{n}\boldsymbol{\delta}, H) \in C(\mathbf{0}, 1)) |\mathbb{X}(H)| \prod_{i \in H} f(\boldsymbol{X}_i^\top \boldsymbol{\delta}),
\end{aligned} \qquad (2.14)$$

protože $u_i = Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}, i \in H$, a tedy $u_i, i \in H$ mají rozdělení s distribuční funkcí F a hustotou f, a platí

$$\begin{aligned}
\lim_{\epsilon \to 0_+} \frac{\mathsf{P}(E_1(H, \boldsymbol{\delta}, \epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} &= \lim_{\epsilon \to 0_+} \frac{\mathsf{P}(\boldsymbol{d}(H) \in C(\boldsymbol{\delta},\epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} \\
&= \lim_{\epsilon \to 0_+} \frac{\mathsf{P}(\mathbb{X}(H)^{-1} \boldsymbol{u}(H) \in C(\boldsymbol{\delta},\epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} \\
&= |\mathbb{X}(H)| \prod_{i \in H} f(\boldsymbol{X}_i^\top \boldsymbol{\delta}).
\end{aligned}$$

Přechod k $\sqrt{n}\boldsymbol{\delta}^*$ nám dává

$$\begin{aligned}
\lim_{\epsilon \to 0_+} \frac{\mathsf{P}(\sqrt{n}\boldsymbol{\delta}^* \in C(\boldsymbol{\delta}, \epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} &= \sum_{H \in \mathcal{H}} \lim_{\epsilon \to 0_+} \frac{\mathsf{P}(\boldsymbol{\delta}^* \in C(n^{-\frac{1}{2}}\boldsymbol{\delta}, n^{-\frac{1}{2}}\epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))} \\
&= \sum_{H \in \mathcal{H}} \lim_{\epsilon \to 0_+} n^{-\frac{k}{2}} \frac{\mathsf{P}(\boldsymbol{\delta}^* \in C(n^{-\frac{1}{2}}\boldsymbol{\delta}, \epsilon))}{\lambda(C(\boldsymbol{\delta},\epsilon))}
\end{aligned}$$

a po dosazení do (2.14) již dostaneme požadované přesné rozdělení (2.6).

(2) Z rozdělení u_i plyne

$$\boldsymbol{z}_i(\boldsymbol{\delta}, H) = \begin{cases} \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} & \text{s pravděpodobností } 1 - F(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}), \\ -\boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} & \text{s pravděpodobností } F(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}). \end{cases}$$

Dále platí

$$\frac{1}{n} \sum_{i \in \bar{H}} \boldsymbol{X}_i \boldsymbol{X}_i^\top = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top - \frac{1}{n} \sum_{i \in H} \boldsymbol{X}_i \boldsymbol{X}_i^\top \xrightarrow{n \to +\infty} \mathbb{Q}.$$

Pak díky rozvoji F kolem 0 lze ukázat, že pro

$$n^{-\frac{1}{2}} \boldsymbol{Z}_n(\boldsymbol{\delta}, H) = n^{-\frac{1}{2}} \sum_{i \in \bar{H}} \boldsymbol{z}_i(\boldsymbol{\delta}, H)$$

platí

$$n^{-\frac{1}{2}} \sum_{i \in \bar{H}} \boldsymbol{z}_i(\boldsymbol{\delta}, H) \xrightarrow{d} \mathcal{N}(-2f(0)\boldsymbol{\delta}^\top \mathbb{Q} \mathbb{X}(H)^{-1}, \mathbb{X}^\top(H)^{-1} \mathbb{Q} \mathbb{X}(H)^{-1}).$$

Taylorův rozvoj F kolem 0 za použití předpokladu $F^{-1}(\frac{1}{2}) = 0$ můžeme rozepsat jako

$$F(y) = F(0) + f(0)y + o(y) = \frac{1}{2} + f(0)y + o(y). \tag{2.15}$$

Do (2.15) dosadíme $y = n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}$:

$$F(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}) = \frac{1}{2} + f(0) n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta} + o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}).$$

Pak můžeme odvodit střední hodnotu $\boldsymbol{z}_i(\boldsymbol{\delta}, H)$ jako

$$\begin{aligned} \mathsf{E}\ \boldsymbol{z}_i(\boldsymbol{\delta}, H) =& \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}(1 - F(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta})) - \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} F(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}) \\ =& \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}(1 - 2F(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta})) \\ =& \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}(1 - 2\frac{1}{2} - 2f(0) n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta} + o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta})) \\ =& -2f(0) n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} \boldsymbol{X}_i^\top \boldsymbol{\delta} + \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}) \\ =& -2f(0) n^{-\frac{1}{2}} \boldsymbol{\delta}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} + \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}). \end{aligned} \tag{2.16}$$

Poslední člen v (2.16) $\boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta})$ lze rozepsat jako

$$\begin{aligned} \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}) &= \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta} \frac{o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta})}{n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}} \\ &= n^{-\frac{1}{2}} \boldsymbol{\delta}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1} \frac{o(n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta})}{n^{-\frac{1}{2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}}. \end{aligned} \tag{2.17}$$

Dohromady tak z (2.16) a (2.17) dostáváme

$$\mathsf{E}\ \boldsymbol{z}_i(\boldsymbol{\delta}, H) = n^{-\frac{1}{2}}\boldsymbol{\delta}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}\Big(-2f(0) + \frac{o(n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta})}{n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta}}\Big),$$

což můžeme dosadit do

$$\begin{aligned} n^{-\frac{1}{2}}\,\mathsf{E}\ \boldsymbol{Z}_n(\boldsymbol{\delta}, H) &= n^{-\frac{1}{2}} \sum_{i\in\bar{H}} \mathsf{E}\ \boldsymbol{z}_i(\boldsymbol{\delta}, H) \\ &= \frac{1}{n}\sum_{i\in\bar{H}} \boldsymbol{\delta}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}\Big(-2f(0) + \frac{o(n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta})}{n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta}}\Big) \\ &\xrightarrow{n\to+\infty} -2f(0)\boldsymbol{\delta}^\top \mathbb{Q}\mathbb{X}(H)^{-1}. \end{aligned}$$

Obdobně lze odvodit asymptotický rozptyl:

$$\begin{aligned} \mathsf{E}\,[\boldsymbol{z}_i(\boldsymbol{\delta}, H)]^2 &= \mathbb{X}^\top(H)^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^\top\mathbb{X}(H)^{-1}, \\ \operatorname{var}\boldsymbol{z}_i(\boldsymbol{\delta}, H) &= \mathbb{X}^\top(H)^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^\top\mathbb{X}(H)^{-1} \\ &\quad - \frac{1}{n}\boldsymbol{\delta}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbb{X}(H)^{-1}\Big(-2f(0) + \frac{o(n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta})}{n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta}}\Big). \\ \operatorname{var} n^{-\frac{1}{2}}\boldsymbol{Z}_n(\boldsymbol{\delta}, H) &= \frac{1}{n}\sum_{i\in\bar{H}} \operatorname{var}\boldsymbol{z}_i(\boldsymbol{\delta}, H) \xrightarrow{n\to+\infty} \mathbb{X}^\top(H)^{-1}\mathbb{Q}\mathbb{X}(H)^{-1}. \end{aligned}$$

Můžeme rozepsat centrovaný sčítanec jako

$$\begin{aligned} n^{-\frac{1}{2}}(\boldsymbol{z}_i(\boldsymbol{\delta}, H) - \mathsf{E}\ \boldsymbol{z}_i(\boldsymbol{\delta}, H)) = \Big(&\operatorname{sgn}(u_i - n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta}) - 2f(0)n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta} \\ &+ n^{-\frac{1}{2}}\boldsymbol{\delta}\boldsymbol{X}_i\frac{o(n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta})}{n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta}}\Big)\boldsymbol{X}_i^\top\mathbb{X}(H)n^{-\frac{1}{2}}. \end{aligned}$$

Protože funkce $\operatorname{sgn}(\cdot)$ je omezená a díky předpokladu $||\mathbb{X}_n||_\infty = o(\sqrt{n})$ ostatní členy konvergují stejnoměrně k nule pro $n \to +\infty, i \in \bar{H}$, celý výraz konverguje stejnoměrně k nule, což implikuje Lindebergovu podmínku, a asymptotické normální rozdělení dostaneme z centrální limitní věty.

Nechť G_n značí pravděpodobnostní míru indukovanou na $\mathbb{R}^k$ náhodnou veličinou $\widetilde{\boldsymbol{Z}}_n = \frac{1}{2}n^{-\frac{1}{2}}\boldsymbol{Z}_n$ a nechť $G_n \to G$.

Definujeme k-dimenzionální otevřené nadkrychle se středem v počátku

$$C_n = C\Big(\mathbf{0}, \frac{1}{2\sqrt{n}}\Big) = \Big\{\boldsymbol{c} \in \mathbb{R}^k : \max_{j\in 1,\dots,k} c_j < \frac{1}{2\sqrt{n}}\Big\}$$

s Lebesgueovou mírou $\lambda(C_n) = n^{-\frac{k}{2}}$. Platí

$$\lim_{n\to+\infty}\frac{G_n(C_n)}{\lambda(C_n)} = \lim_{n\to+\infty} n^{\frac{k}{2}}\,\mathsf{P}(\widetilde{\boldsymbol{Z}}_n \in C_n) = g(\mathbf{0}),$$

kde $g(\mathbf{0})$ je hustota G v počátku.

Tedy z předchozích výsledků dostáváme

$$
\begin{aligned}
& n^{\frac{k}{2}}\,\mathsf{P}(\boldsymbol{Z}_n(\boldsymbol{\delta},H)\in C_n)\\
&\quad = \frac{1}{(2\pi)^{\frac{k}{2}}}\left|\frac{1}{4}\mathbb{X}^\top(H)^{-1}\mathbb{Q}\mathbb{X}(H)^{-1}\right|^{\frac{1}{2}}\\
&\qquad \times \exp\left\{-\frac{1}{2}f(0)^2\boldsymbol{\delta}^\top\mathbb{Q}\mathbb{X}^\top(H)^{-1}\left[\frac{1}{4}\mathbb{X}^\top(H)^{-1}\mathbb{Q}\mathbb{X}(H)^{-1}\right]^{-1}\mathbb{X}(H)^{-1}\mathbb{Q}\boldsymbol{\delta}\right\}\\
&\qquad + o(1)\\
&\quad = \frac{1}{(2\pi)^{\frac{k}{2}}}2^k|\mathbb{X}(H)||\mathbb{Q}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}[2f(0)]^2\boldsymbol{\delta}^\top\mathbb{Q}\boldsymbol{\delta}\right\}+o(1). \qquad (2.18)
\end{aligned}
$$

Spojitost hustoty v mediánu (tedy v nule) implikuje

$$\prod_{i\in H} f(n^{-\frac{1}{2}}\boldsymbol{X}_i^\top\boldsymbol{\delta}) = f^k(0)+o(1). \qquad (2.19)$$

Po dosazení (2.18) a (2.19) do přesného rozdělení (2.6) odvozeného v prvním kroku důkazu dostaneme

$$
\begin{aligned}
\phi_n(\boldsymbol{\delta}) = {} & \sum_{H\in\mathcal{H}} n^{-k}|\mathbb{X}(H)|^2|\mathbb{Q}|^{-\frac{1}{2}}\frac{1}{(2\pi)^{\frac{k}{2}}}[2f(0)]^k\exp\left\{-\frac{1}{2}[2f(0)]^2\boldsymbol{\delta}^\top\mathbb{Q}\boldsymbol{\delta}\right\}\\
& + o(1)\sum_{H\in\mathcal{H}} n^{-k}|\mathbb{X}(H)|^2. \qquad (2.20)
\end{aligned}
$$

Protože dle Lemmatu 6 $\sum_{H\in\mathcal{H}}|\mathbb{X}(H)|^2 = |\mathbb{X}^\top\mathbb{X}|$, pak spolu s členem n^{-k} dostaneme limitně $|\mathbb{Q}|$:

$$\sum_{H\in\mathcal{H}} n^{-k}|\mathbb{X}(H)|^2 \xrightarrow{n\to+\infty} |\mathbb{Q}|. \qquad (2.21)$$

Po dosazení (2.21) do (2.20) dostáváme

$$\phi_n(\boldsymbol{\delta}) \xrightarrow{n\to+\infty} \frac{1}{(2\pi)^{\frac{k}{2}}}|[2f(0)]^2\mathbb{Q}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}[2f(0)]^2\boldsymbol{\delta}^\top\mathbb{Q}\boldsymbol{\delta}\right\},$$

což je hustota požadovaného normálního rozdělení (2.7).

Podle Scheffého věty bodová konvergence hustot absolutně spojitých náhodných veličin implikuje konvergenci v distribuci těchto náhodných veličin, věta je tedy dokázána.

□

Poznámka. Větu lze dokázat i pro obecný kvantil, viz Koenker a Bassett (1978) nebo Koenker a kol. (2005).

Důsledek. Z Věty 7 plyne také konzistence L_1 odhadu regresních koeficientů, protože $\mathsf{E}\,\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}$ a $\operatorname{var}\widehat{\boldsymbol{\beta}}_n \xrightarrow{n\to+\infty} \mathbb{O}_{k\times k}$. Lze ji však dokázat i za slabších předpokladů, viz Zhao, Rao a Chen (1993) pro postačující podmínku a Chen a Wu (1993) pro nutnou podmínku.

V rámci důkazu Věty 7 jsme odvodili i přesné rozdělení regresních koeficientů, které je však kvůli sumě přes $H\in\mathcal{H}$ v praxi těžko použitelné a vyžadovalo by znalost hustoty f, tedy přibyl by silný předpoklad na rozdělení chyb. Uchylujeme se proto k asymptotickému rozdělení.

2.5 Inference

Pro využití L_1 regrese v praxi nám nestačí pouze znát asymptotické rozdělení odhadů regresních koeficientů, ale potřebujeme vhodné nástroje pro inferenci, zejména formální testy hypotéz a konfidenční intervaly, případně množiny, které lze z asymptotického rozdělení přímo odvodit. V této kapitole uvedeme test Waldova typu, který pro L_1 regresi poprvé sestrojili Koenker a Bassett (1982), a odvodíme konfidenční intervaly a množiny Waldova typu.

2.5.1 Testování hypotéz

Jednou z nejběžnějších otázek v regresním modelu je, zda odezva opravdu závisí na regresorech, případně zda závisí jen na některých z nich. Závislost na konkrétním regresoru je přítomna tehdy, pokud je odpovídající regresní koeficient nenulový. Koeficienty jsou však neznámé a k dispozici máme pouze bodový odhad $\widehat{\boldsymbol{\beta}}_n$, který skoro jistě nebude nulový ani v případě, že skutečné koeficienty jsou nulové. Je proto potřeba zkonstruovat formální testy, které mohou na pevně zvolené hladině α zamítnout nulovost skutečných koeficientů.

Nejprve budeme uvažovat test, kde nás zajímá, zda má alespoň jeden z regresorů nenulový koeficient. Uvažujme nulovou a alternativní hypotézu

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \quad H_1 : \boldsymbol{\beta} \neq \mathbf{0},$$

pak z Věty 7 za platnosti H_0 plyne

$$\sqrt{n}\widehat{\boldsymbol{\beta}}_n \xrightarrow{d} \mathcal{N}_k\Big(0, \frac{1}{[2f(0)]^2}\mathbb{Q}^{-1}\Big),\ n \to +\infty$$

a přímočaře dostáváme

$$n[2f(0)]\widehat{\boldsymbol{\beta}}_n^\top \mathbb{Q} \widehat{\boldsymbol{\beta}}_n \xrightarrow{d} \chi^2_k,\ n \to +\infty.$$

Matici $n\mathbb{Q}$ lze odhadnout pomocí napozorované matice $\mathbb{X}^\top\mathbb{X}$, nicméně odhad $f(0)$ je více problematický a možné přístupy budou shrnuty níže. Nulová hypotéza $H_0 : \boldsymbol{\beta} = \mathbf{0}$ je speciálním případem obecnější nulové hypotézy, která umožňuje nulovost pouze některých složek vektoru $\boldsymbol{\beta}$.

Rozdělíme k-rozměrný vektor neznámých regresních parametrů $\boldsymbol{\beta}$ na dvě složky $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, kde $\boldsymbol{\beta}_1^\top$ je $(k-p)$-rozměrný vektor a $\boldsymbol{\beta}_2^\top$ je p-rozměrný vektor, a regresní matici $\mathbb{X}$ rozdělíme na odpovídající podmatice $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$, kde $\mathbb{X}_1$ je rozměru $n \times (k-p)$ a $\mathbb{X}_2$ je rozměru $n \times p$. Lineární model pak můžeme zapsat ve tvaru

$$\boldsymbol{Y} = \mathbb{X}\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Zajímat nás bude následující nulová a alternativní hypotéza:

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}, \quad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}.$$

Dále pro regulární $(k \times k)$ matici $\mathbb{Q}$ označíme

$$\mathbb{Q} = \begin{pmatrix}\mathbb{Q}_{11} & \mathbb{Q}_{12} \\ \mathbb{Q}_{21} & \mathbb{Q}_{22}\end{pmatrix}, \qquad \mathbb{Q}^{-1} = \begin{pmatrix}\mathbb{Q}^{11} & \mathbb{Q}^{12} \\ \mathbb{Q}^{21} & \mathbb{Q}^{22}\end{pmatrix}.$$

Jako $\chi^2_\nu(\eta)$ budeme značit necentrální chí kvadrát rozdělení, kde ν je počet stupňů volnosti a η je parametr necentrality. Definováno je jako

$$\chi^2_\nu(\eta) = \sum_{i=1}^{\nu} X_i,$$

kde $X_i, i = 1, \ldots, \nu$ jsou nezávislé normální veličiny se střední hodnotou μ_i a jednotkovým rozptylem, $\eta = \sum_{i=1}^{\nu} \mu_i^2$. Pokud $\mu_i = 0, i = 1, \ldots, \nu$, X_i mají normované normální rozdělení a $\chi^2_\nu(\eta) = \chi^2_\nu$ je klasické chí kvadrát rozdělení.

Pomocí Věty 7 sestrojíme test Waldova typu s testovou statistikou

$$W_n = n[2f(0)]^2 \widehat{\boldsymbol{\beta}}_2^\top (\mathbb{Q}^{22})^{-1} \widehat{\boldsymbol{\beta}}_2, \tag{2.22}$$

kde $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_1^\top, \widehat{\boldsymbol{\beta}}_2^\top)^\top$ je řešení úlohy (2.5) a jeho složka $\widehat{\boldsymbol{\beta}}_2$ je odhad $\boldsymbol{\beta}_2$.

Asymptotické rozdělení této testové statistiky je odvozeno v následující větě.

Věta 8. *Nechť W_n je testová statistika definovaná v (2.22) a platí předpoklady Věty 7.*

(i) Pokud platí nulová hypotéza $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, pak

$$W_n \xrightarrow{d} \chi^2_p, \ n \to +\infty.$$

(ii) Pokud existuje pevné $\boldsymbol{\gamma} \in \mathbb{R}^p$ takové, že $\boldsymbol{\beta}_2 = \frac{\gamma}{\sqrt{n}}$ pro všechny velikosti výběru n, pak

$$W_n \xrightarrow{d} \chi^2_p(\eta), \ n \to +\infty,$$

kde $\eta = [2f(0)]^2 \boldsymbol{\gamma}^\top (\mathbb{Q}^{22})^{-1} \boldsymbol{\gamma}$.

Důkaz. Z Věty 7 víme, že

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_k\Big(0, \frac{1}{[2f(0)]^2} \mathbb{Q}^{-1}\Big),$$

z čehož plyne

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \xrightarrow{d} \mathcal{N}_p\Big(0, \frac{1}{[2f(0)]^2} \mathbb{Q}^{22}\Big).$$

(i) Za H_0 platí

$$\begin{aligned}
\sqrt{n}\widehat{\boldsymbol{\beta}}_2 &\xrightarrow{d} \mathcal{N}_p\Big(0, \frac{1}{[2f(0)]^2} \mathbb{Q}^{22}\Big), \\
\sqrt{n}[2f(0)]\widehat{\boldsymbol{\beta}}_2 (\mathbb{Q}^{22})^{-\frac{1}{2}} &\xrightarrow{d} \mathcal{N}_p(0, \mathbb{I}_p), \\
n[2f(0)]^2 \widehat{\boldsymbol{\beta}}_2^\top (\mathbb{Q}^{22})^{-1} \widehat{\boldsymbol{\beta}}_2 &\xrightarrow{d} \chi^2_p.
\end{aligned}$$

(ii) Z předpokladu existence $\boldsymbol{\gamma} = \sqrt{n}\boldsymbol{\beta}_2$ dostáváme

$$\begin{aligned}
\sqrt{n}\widehat{\boldsymbol{\beta}}_2 &\xrightarrow{d} \mathcal{N}_p\Big(\boldsymbol{\gamma}, \frac{1}{[2f(0)]^2} \mathbb{Q}^{22}\Big), \\
\sqrt{n}[2f(0)]\widehat{\boldsymbol{\beta}}_2 (\mathbb{Q}^{22})^{-\frac{1}{2}} &\xrightarrow{d} \mathcal{N}_p\Big([2f(0)]\boldsymbol{\gamma}(\mathbb{Q}^{22})^{-\frac{1}{2}}, \mathbb{I}_p\Big), \\
n[2f(0)]^2 \widehat{\boldsymbol{\beta}}_2^\top (\mathbb{Q}^{22})^{-1} \widehat{\boldsymbol{\beta}}_2 &\xrightarrow{d} \chi^2_p(\eta).
\end{aligned}$$

□

Asymptotický rozptyl $n[f(0)]^2(\mathbb{Q}^{22})^{-1}$ je neznámý, pro sestrojení testu je tedy nutné jej v testové statistice W_n nahradit jeho odhadem. K odhadu $n(\mathbb{Q}^{22})^{-1}$ využijeme známé identity

$$\mathbb{Q}^{22} = (\mathbb{Q}_{22} - \mathbb{Q}_{21}\mathbb{Q}_{11}^{-1}\mathbb{Q}_{12})^{-1},$$

ze které dostáváme

$$n(\mathbb{Q}^{22})^{-1} = n(\mathbb{Q}_{22} - \mathbb{Q}_{21}\mathbb{Q}_{11}^{-1}\mathbb{Q}_{12}).$$

Do podmatice $\mathbb{Q}_{ij}$ dosadíme pozorované matice $\mathbb{X}_1$ a $\mathbb{X}_2$:

$$\begin{aligned} \widehat{n(\mathbb{Q}^{22})^{-1}} &= n\Big(\frac{1}{n}\mathbb{X}_2^\top\mathbb{X}_2 - \frac{1}{n}\mathbb{X}_2^\top\mathbb{X}_1\Big(\frac{1}{n}\mathbb{X}_1^\top\mathbb{X}_1\Big)^{-1}\frac{1}{n}\mathbb{X}_1^\top\mathbb{X}_2\Big) \\ &= \mathbb{X}_2^\top\mathbb{X}_2 - \mathbb{X}_2^\top\mathbb{X}_1(\mathbb{X}_1^\top\mathbb{X}_1)^{-1}\mathbb{X}_1^\top\mathbb{X}_2 \\ &= \mathbb{X}_2^\top(\mathbb{I}_n - \mathbb{X}_1(\mathbb{X}_1^\top\mathbb{X}_1)^{-1}\mathbb{X}_1^\top)\mathbb{X}_2. \end{aligned}$$

Možné přístupy odhadu $f(0)$ jsou shrnuty v kapitole 4.10.1, Koenker a kol. (2005). Pro obecný τ-kvantil definujeme

$$s(\tau) = \frac{1}{f(F^{-1}(\tau))} = \frac{\mathrm{d}F^{-1}(\tau)}{\mathrm{d}\tau}.$$

Pro $s(\tau)$ se pak nabízí jednoduchý odhad v podobě diferenčních poměrů empirické kvantilové funkce, tedy

$$\widehat{s}(\tau) = \frac{\widehat{F_n^{-1}}(\tau + h_n) - \widehat{F_n^{-1}}(\tau - h_n)}{2h_n}, \tag{2.23}$$

kde h_n je posloupnost vyhlazovacích parametrů, $h_n \to 0, n \to +\infty$.

Označíme rezidua

$$\widehat{u}_i = Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n,\ i = 1, \ldots, n.$$

Pak můžeme za $\widehat{F_n^{-1}}$ v (2.23) dosadit empirickou kvantilovou funkci reziduí

$$\widehat{F_n^{-1}}(\tau) = \widehat{u}_{(i)},\ \tau \in \Big[\frac{i-1}{n}, \frac{i}{n}\Big),$$

nebo ekvivalentně

$$\widehat{F_n^{-1}}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[\widehat{u}_i \le \tau].$$

Je důležité si uvědomit, že díky vlastnosti z Lemmatu 3 bude k reziduí nulových, a je tak třeba volit dostatečně velké h_n (problém nastává, pokud je $\frac{k}{n}$ relativně velké vůči h_n). Alternativně se dá těchto k pozorování vynechat a odhad sestrojit pouze pro zbývajících $(n-p)$ pozorování.

Zbývá již jen vhodně určit posloupnost vyhlazovacích parametrů h_n. Pro účely konfidenčních intervalů a testových statistik na hladině $\alpha \in (0,1)$ autoři Hall a Sheather (1988) doporučují

$$h_n = n^{-\frac{1}{3}}\Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big)\Big[\frac{3}{2}\frac{\phi^2(\Phi^{-1}(\tau))}{2\Phi^{-1}(\tau)^2 + 1}\Big]^{\frac{1}{3}},$$

kde Φ^{-1} je kvantilová funkce a ϕ je hustota standardního normálního rozdělení.

Protože $s(\frac{1}{2}) = \frac{1}{f(0)}$, jako odhad $f(0)$ můžeme brát

$$\widehat{f}(0) = \frac{1}{\widehat{s}(\frac{1}{2})}.$$

Označíme

$$\widehat{W}_n = n[2\widehat{f}(0)]^2 \widehat{\boldsymbol{\beta}}_2^\top \widehat{(\mathbb{Q}^{22})^{-1}} \widehat{\boldsymbol{\beta}}_2.$$

Nulovou hypotézu $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ na hladině α pak zamítáme pro

$$\widehat{W}_n > \chi^2_p(1-\alpha).$$

Poznámka. Lze také zformulovat test pro obecnou lineární hypotézu

$$H_0 : \ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{r},$$

kde $\boldsymbol{r} \in \mathbb{R}^m, \mathbb{L} \in \mathbb{R}^{m\times k}$ a $rank(\mathbb{L}) = m$, tedy matice $\mathbb{L}$ má plnou řádkovou hodnost.

Z Věty 7 dostáváme

$$\sqrt{n}(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_k\Big(0, \frac{1}{[2f(0)]^2}\mathbb{L}\mathbb{Q}^{-1}\mathbb{L}^\top\Big),$$

což lze za platnosti H_0 upravit na

$$\sqrt{n}(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{r}) \xrightarrow{d} \mathcal{N}_m\Big(0, \frac{1}{[2f(0)]^2}\mathbb{L}\mathbb{Q}^{-1}\mathbb{L}^\top\Big)$$

a obdobným postupem jako v důkazu Věty 8 dostaneme

$$n[f(0)]^2(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{r})^\top(\mathbb{L}\mathbb{Q}^{-1}\mathbb{L}^\top)^{-1}(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{r}) \xrightarrow{d} \chi^2_m.$$

Člen asymptotického rozptylu $n(\mathbb{L}\mathbb{Q}^{-1}\mathbb{L}^\top)^{-1}$ můžeme odhadnout pomocí matice $(\mathbb{L}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{L}^\top)^{-1}$ a pomocí odhadu $f(0)$ popsaného výše můžeme sestrojit test analogicky.

2.5.2 Konfidenční intervaly a množiny

Testy mohou nulovou hypotézu pouze zamítnout, ale nemohou ji přijmout. V případě, že zamítáme nulovou hypotézu, navíc stále nevíme nic dalšího o skutečné hodnotě parametru, která může být nulové hypotéze více, či méně vzdálená. Pro lepší představu o skutečné hodnotě parametru tak mohou sloužit konfidenční intervaly a množiny, které skutečnou hodnotu pokrývají se spolehlivostí $1-\alpha$ a lze je rovněž odvodit pomocí Věty 7

Pro $j \in \{1, \ldots, k\}$ označíme q^j j-tý prvek na diagonále matice $\mathbb{Q}^{-1}$ a $\widehat{\beta}_j$ j-tou složku vektoru $\widehat{\boldsymbol{\beta}}_n$. Pak

$$\sqrt{n}(\widehat{\beta}_j - \beta_j) \xrightarrow{d} \mathcal{N}\Big(0, \frac{q^j}{[2f(0)]^2}\Big).$$

Dostáváme tak standardní konfidenční interval Waldova typu, který má asymptotické pokrytí $1-\alpha$:

$$\left(\widehat{\beta}_j - \Phi^{-1}\Big(1-\frac{1}{\alpha}\Big)\frac{\sqrt{nq^j}}{2f(0)},\ \widehat{\beta}_j + \Phi^{-1}\Big(1-\frac{1}{\alpha}\Big)\frac{\sqrt{nq^j}}{2f(0)}\right).$$

Člen q^j můžeme odhadnout pomocí j-tého prvku na diagonále matice $\left(\frac{1}{n}\mathbb{X}^\top\mathbb{X}\right)^{-1}$ a $f(0)$ odhadneme obdobně jako pro testy v podkapitole 2.5.1.

Konfidenční množinu s asymptotickým pokrytím $(1-\alpha)$ pro celý vektor $\boldsymbol{\beta}$ získáme pomocí chí kvadrát rozdělení

$$S(\alpha) = \{\boldsymbol{\beta} \in \mathbb{R}^k : n[2f(0)]^2(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top \mathbb{Q}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) < \chi^2_k(1-\alpha)\},$$

kde $[2f(0)]^2\mathbb{Q}$ můžeme opět nahradit odhady navrhovanými výše. Množina $S(\alpha)$ tvoří elipsoid se středem v $\widehat{\boldsymbol{\beta}}_n$.

Lze sestrojit konfidenční množinu i pro daný podvektor $\boldsymbol{\beta}$ pomocí rozdělení podvektoru odvozeného v důkazu Věty 8.

2.5.3 Další možné přístupy

Existují další přístupy pro inferenci, které nejsou přímo odvozené z asymptotické normality ve Větě 7, a tudíž odpadá problematika odhadu $f(0)$. Jednou z alternativ jsou testy založené na pořadí (anglicky *rank*), které jsou analogií Wilcoxonova a Mannova-Whitneyovatestu pro lineární model. Poprvé rank testy zformulovali Gutenbrunner, Jurečková, Koenker a Portnoy (1993) a inference založená na této metodě je shrnuta v kapitole 3.5, Koenker a kol. (2005). Další možností pro inferenci je bootstrap, jehož možné využití je shrnuto v kapitole 3.9, Koenker a kol. (2005) a v kapitole 3.10 jsou pomocí simulací srovnány vlastnosti různých typů bootstrapu s přístupy Waldova typu a rank přístupy.

3. Simulační studie

Vlastnosti L_1 odhadů pro konečný náhodný výběr budeme srovnávat s L_2 odhady (odhady metodou nejmenších čtverců) v simulační studii, která byla zpracována v prostředí R (R Core Team, 2019) za pomoci balíčku `quantreg` (Koenker, 2019). V balíčku jsou mimo jiné implementovány testy na podmodel Waldova typu, které jsou popsané v kapitole 2.5, a základním výstupem jsou odpovídající odhady směrodatných odchylek a p-hodnoty testu nulové hypotézy $H_0 : \beta_j = 0$ pro jednotlivé parametry $\beta_j, j = 1, \ldots, k$, které slouží jako ukazatel jejich statistické významnosti.

Provedeme dvě oddělené části simulační studie – jednu s modelem s jedním regresorem a druhou s modelem s více regresory. V první části se zaměříme především na testy významnosti a konfidenční intervaly jednoho parametru, ve druhé části na testy na podmodel a konfidenční množiny. Studie je rozdělená do dvou částí z toho důvodu, že v první části budeme zkoumat závislost síly testu na skutečné hodnotě parametru, což vede na větší výpočetní náročnost a složitost interpretace výsledků. U modelu s více regresory kvůli zjednodušení výpočtu i interpretace výsledků již tuto závislost uvažovat nebudeme.

V obou částech uvažujeme různé velikosti výběru n a rozdělení náhodných chyb $\varepsilon_i,\ i = 1, \ldots, n$. Uvažované hodnoty n jsou $50, 100, 200, \ldots, 1000$ pro model s jedním regresorem a $50, 100, 500$ pro model s více regresory. Posloupnost velikostí výběru n je u modelu s jedním regresorem volena zejména kvůli grafickým výsledkům.

Uvažovaná rozdělení chyb jsou následující:

- F_1: standardní normální rozdělení $\mathcal{N}(0, 1)$;
- F_2: Cauchyho rozdělení;
- F_3: směs rozdělení ve tvaru

$$0.9\mathcal{N}(0, 1) + 0.1\mathcal{R}(-6{,}6),$$

  tedy směs standardního normálního rozdělení a rovnoměrného rozdělení na $(-6, 6)$, které představuje znečištění odlehlými pozorováními;
- F_4: posunuté log-normální rozdělení tak, aby byl medián roven nule (viz předpoklad na medián), tedy $(\varepsilon + 1) \sim \mathcal{LN}(0{,}1)$.

Normální rozdělení je voleno pro srovnání L_1 a L_2 regrese v případě, že jsou předpoklady L_2 regrese splněny (a to dokonce v silnější podobě normálního modelu). Cauchyho rozdělení je uvažováno jako příklad rozdělení s těžkými chvosty, které v tomto případě nemá ani střední hodnotu. Rozdělení F_3 sice splňuje předpoklady L_2 regrese, ale obsahuje odlehlá pozorování, slouží tedy k porovnání robustnosti obou metod. Rozdělení F_4 je uvažováno jako příklad rozdělení, které není symetrické. Log-normální rozdělení je posunuto tak, aby $F_4^{-1}(\frac{1}{2}) = 0$, ale neplatí nulovost střední hodnoty, což je potřeba brát v úvahu při analýze výsledků.

První tři rozdělení jsou symetrická kolem nuly, z čehož plyne, že s výjimkou Cauchyho rozdělení obě metody odhadují stejnou veličinu, protože je střední hodnota rovna mediánu. V případě Cauchyho rozdělení střední hodnota neexistuje,

takže L_2 odhad nedává dobrý smysl, nicméně medián existuje vždy a rozdělení splňuje předpoklady L_1 regrese. V případě čtvrtého rozdělení L_1 regrese odhaduje podmíněný medián, zatímco L_2 regrese odhaduje podmíněnou střední hodnotu, takže výsledky nejsou přímo porovnatelné, ale stále můžeme srovnávat vlastnosti regresních odhadů. Protože pro F_4 neplatí nulovost střední hodnoty, lze u L_2 odhadu očekávat vychýlení v absolutním členu o $e^{\frac{1}{2}} - 1$, což je rozdíl mezi střední hodnotou a mediánem v $\mathcal{LN}(0,1)$.

V obou částech simulační studie provedeme $iter = 10\,000$ iterací. Počet iterací je volen jako kompromis mezi výpočetním časem a dosažením co nejnižších monte carlo odchylek (MCE). Metriky použité pro analýzu výsledků simulací jsou:

- Střední čtvercová odchylka odhadů od skutečného odhadu, odhadnutá jako

$$\mathrm{MSE}(\beta_j) = \frac{1}{iter} \sum_{i=1}^{iter} (\beta_j - \widehat{\beta}_{j,i})^2.$$

- Síla testu, tedy pravděpodobnost zamítnutí neplatné nulové hypotézy, odhadnutá jako podíl zamítnutých neplatných nulových hypotéz:

$$\mathrm{Power}(H_1) = \frac{1}{iter} \sum_{i=1}^{iter} \mathbb{I}[p_i \leq \alpha],$$

  kde p_i je p-hodnota testu v i-té iteraci a α je uvažovaná hladina testu.

- Chyba I. typu, tedy pravděpodobnost zamítnutí platné nulové hypotézy, odhadnutá jako podíl zamítnutých platných nulových hypotéz (stejný výpočet jako u síly testu, pouze rozdíl v tom, zda platí, nebo neplatí nulová hypotéza):

$$\mathrm{Error}(H_0) = \frac{1}{iter} \sum_{i=1}^{iter} \mathbb{I}[p_i \leq \alpha].$$

- Pokrytí intervalů, tedy pravděpodobnost, že interval pokrývá skutečnou hodnotu parametru, odhadnuto jako podíl intervalů, které obsahují skutečnou hodnotu parametru:

$$\mathrm{Coverage}(\beta_j) = \frac{1}{iter} \sum_{i=1}^{iter} \mathbb{I}[\widehat{\beta}_{j,L,i} < \beta_j < \widehat{\beta}_{j,U,i}],$$

  kde $\widehat{\beta}_{j,L,i}$ a $\widehat{\beta}_{j,U,i}$ jsou dolní a horní meze konfidenčního intervalu j-tého regresního parametru v i-té iteraci se spolehlivostí $(1-\alpha)$, případně pokrytí konfidenční množiny odhadnuté jako

$$\mathrm{Coverage}(\boldsymbol{\beta}) = \frac{1}{iter} \sum_{i=1}^{iter} \mathbb{I}[\boldsymbol{\beta} \in S_i(\alpha)],$$

  kde $S_i(\alpha)$ je konfidenční množina odvozená v kapitole 2.5.2 pro i-tou iteraci.

Všechny testy a intervaly jsou konstruovány na hladině $\alpha = 0.05$. Žádoucí vlastnosti testů a intervalů jsou tedy chyba I. typu 5 % a pokrytí konfidenčních intervalů 95 %. U síly testu nezávisle na hladině požadujeme pro $n \to +\infty$ konvergenci k jedné, nicméně je zřejmé, že rychlost konvergence bude silně záviset na blízkosti nulové hypotéze.

3.1 Model s jedním regresorem

V první simulační studii pracujeme s modelem s absolutním členem a jedním regresorem, tedy s modelem ve tvaru

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n.$$

3.1.1 Nastavení a výpočet simulací

Uvažujeme různé hodnoty parametru β_1. Nastavení regresních parametrů je následující:

- pro všechny simulace je $\beta_0 = 1$,
- uvažované hodnoty β_1 jsou $0, 0.2, 0.5, 1$.

Různé hodnoty β_1 umožňují zkoumat závislost síly testu nulové hypotézy H_0 : $\beta_1 = 0$ na skutečné hodnotě β_1.

Pro každé n byl vygenerován jeden vektor regresorů $\boldsymbol{X}_n \sim \mathcal{N}_n(\mathbf{0}, \mathbb{I}_n)$ a pro každou kombinaci β_1 a n bylo spočteno $iter = 10\,000$ iterací následujícím způsobem:

1. Je vygenerován vektor pozic odlehlých pozorování pomocí alternativního rozdělení s parametrem $p = 0.1$.
2. Pro každé rozdělení vygenerujeme vektor náhodných chyb $\boldsymbol{\varepsilon}$ a spočteme odezvu $\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{X}_n + \boldsymbol{\varepsilon}$.
3. Pro všechny odezvy odhadneme L_1 a L_2 regresní modely. Pro každý z těchto modelů uložíme odhady regresních koeficientů, jejich odhadnuté směrodatné odchylky a p-hodnoty.

Z výsledků simulací jsou následně odhadnuty zkoumané metriky a jejich MCE.

3.1.2 Analýza výsledků

Výstupem jsou výsledky pro všechny kombinace rozdělení chyb, velikosti výběru a hodnot β_1, není tak reálné je do detailu uvést všechny. Nejzajímavější výsledky si proto shrneme v podobě grafů a tabulek.

Protože MSE jednotlivých odhadů nezávisí na skutečné hodnotě β_1, pro přehlednost uvedeme výsledky pouze pro případ $\beta_1 = 1$, které jsou pro některé velikosti výběru shrnuty v Tabulce 3.1 spolu s MSE pro absolutní člen β_0. Srovnatelné hodnoty L_1 a L_2 odhadů jsou vždy umístěny v tabulce nad sebou a v závorce je uvedena odpovídající MCE.

Dle očekávání je L_2 odhad přesnější v normálním modelu (F_1), kde je MSE u všech uvažovaných velikostí výběru přibližně o třetinu nižší než u L_1 odhadu. U Cauchyho rozdělení (F_2) je zjevné, že L_2 odhad nedává dobrý smysl a MSE se pohybují řádově v tisících až statisících, zatímco L_1 odhad stále dává rozumné výsledky. U normálního rozdělení znečištěného odlehlými pozorováními (F_3) MSE vychází vyšší pro L_2 odhad. U log-normálního rozdělení (F_4) je pro L_2 odhad uvažována skutečná hodnota $\beta_0 = e^{\frac{1}{2}}$, protože se do absolutního členu přesouvá

nenulovost střední hodnoty tohoto rozdělení. I po této korekci vychází MSE nižší pro L_1 odhad u obou parametrů.

Tabulka 3.1: Srovnání střední čtvercové chyby (MSE) L_1 a L_2 regresních odhadů v modelu s jedním regresorem.

			$n = 50$		$n = 100$		$n = 500$	
F_1	β_0	L_1	0.031	*(0.00045)*	0.016	*(0.00022)*	0.003	*(0.00004)*
		L_2	0.020	*(0.00028)*	0.010	*(0.00014)*	0.002	*(0.00003)*
	β_1	L_1	0.026	*(0.00037)*	0.011	*(0.00016)*	0.003	*(0.00004)*
		L_2	0.016	*(0.00023)*	0.007	*(0.00010)*	0.002	*(0.00003)*
F_2	β_0	L_1	0.056	*(0.00088)*	0.026	*(0.00038)*	0.005	*(0.00007)*
		L_2	5 336	*(3 142)*	317 483	*(28 229)*	283 499	*(27 559)*
	β_1	L_1	0.056	*(0.00099)*	0.020	*(0.00033)*	0.005	*(0.00007)*
		L_2	13 435	*(9 366)*	576 793	*(56 722)*	4 365	*(2 050)*
F_3	β_0	L_1	0.045	*(0.00065)*	0.022	*(0.00032)*	0.004	*(0.00006)*
		L_2	0.071	*(0.00102)*	0.035	*(0.00051)*	0.007	*(0.00009)*
	β_1	L_1	0.039	*(0.00059)*	0.016	*(0.00023)*	0.004	*(0.00006)*
		L_2	0.057	*(0.00086)*	0.025	*(0.00037)*	0.007	*(0.00009)*
F_4	β_0	L_1	0.035	*(0.00056)*	0.017	*(0.00026)*	0.003	*(0.00005)*
		L_2	0.097	*(0.00299)*	0.045	*(0.00073)*	0.009	*(0.00014)*
	β_1	L_1	0.025	*(0.00038)*	0.011	*(0.00016)*	0.003	*(0.00004)*
		L_2	0.076	*(0.00198)*	0.033	*(0.00069)*	0.009	*(0.00016)*

V Tabulce 3.2 jsou shrnuty odhadnuté chyby I. typu pro test nulové hypotézy $H_0 : \beta_1 = 0$. Pro každé rozdělení a metodu je uvedena pouze maximální monte carlo odchylka (MCE), protože se v závislosti na n příliš neliší. Z výsledků lze vyvodit, že L_2 test drží hladinu pro konečné n výrazně lépe než L_1 test, který je antikonzervativní (tedy má vyšší než požadovanou chybu I. řádu).

Tabulka 3.2: Srovnání chyby I. typu L_1 a L_2 testů v modelu s jedním regresorem.

		$n = 50$	$n = 100$	$n = 500$	*MCE*
Normální	L_1	7.82 %	7.32 %	5.91 %	*0.27 %*
	L_2	5.00 %	4.94 %	5.16 %	*0.22 %*
Cauchy	L_1	6.98 %	6.37 %	5.33 %	*0.25 %*
	L_2	5.98 %	4.64 %	4.96 %	*0.24 %*
Odlehlá poz.	L_1	7.20 %	6.98 %	5.69 %	*0.26 %*
	L_2	5.06 %	4.79 %	5.06 %	*0.22 %*
Log-normální	L_1	8.64 %	8.56 %	6.49 %	*0.28 %*
	L_2	5.79 %	5.04 %	5.27 %	*0.23 %*

Testy založené na metodě nejmenších čtverců sice drží lépe hladinu u všech uvažovaných rozdělení včetně Cauchyho rozdělení, pro které z pohledu MSE není L_2 regrese vhodná, nicméně je to za cenu velice nízké síly testu. I když se tak z pohledu chyby I. řádu mohou L_2 testy zdát lepší než L_1 testy, reálně v některých případech fungují výrazně hůře, což je vidět na grafech na následujících stranách. Na Obrázcích 3.1 až 3.4 jsou grafy, na kterých je vyzobrazena závislost síly testu na velikosti výběru n pro uvažované skutečné hodnoty β_1.

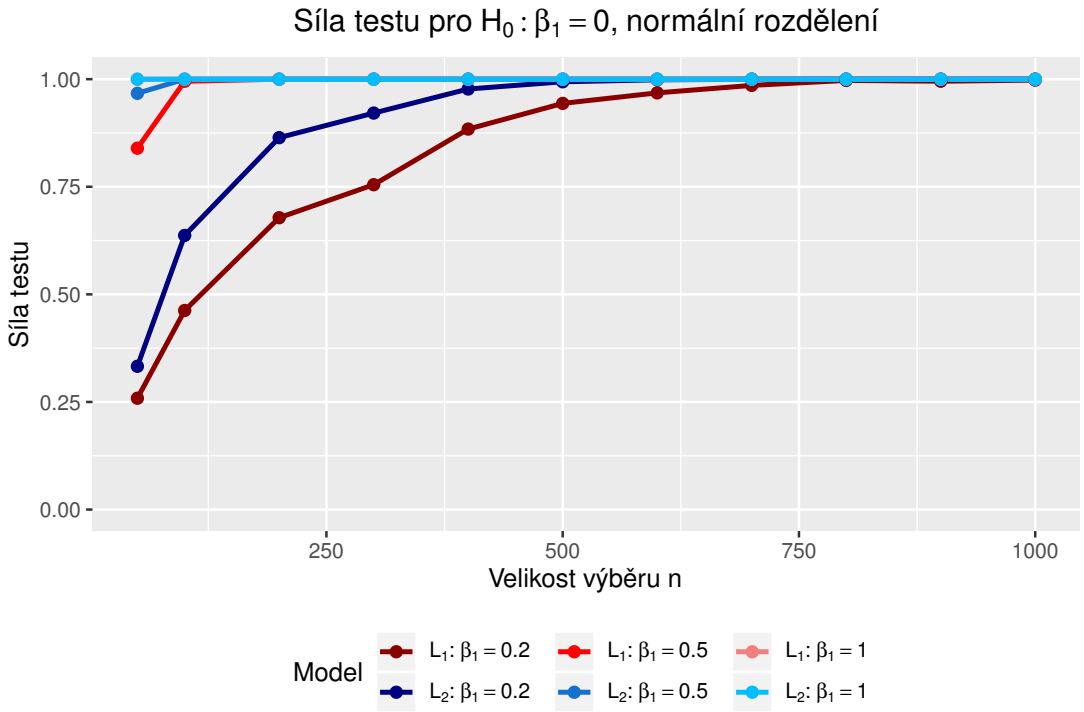
Na Obrázku 3.1 opět vidíme, že pro normální rozdělení se lépe chová L_2 test, který má obecně větší sílu, nicméně při větších velikostech výběru a vyšších skutečných hodnotách β_1 mají obě metody sílu blízkou jedné. Křivky pro $\beta_1 = 1$ se pro obě metody překrývají na jednotkové síle, které dosahují i pro nejmenší velikost výběru.

Oproti tomu na Obrázku 3.2 vidíme, že pro Cauchyho rozdělení není L_2 test vhodný vůbec, protože jeho síla byla vždy nízká a navíc není patrný růst s velikostí výběru, zatímco L_1 test stále dává dobré výsledky.

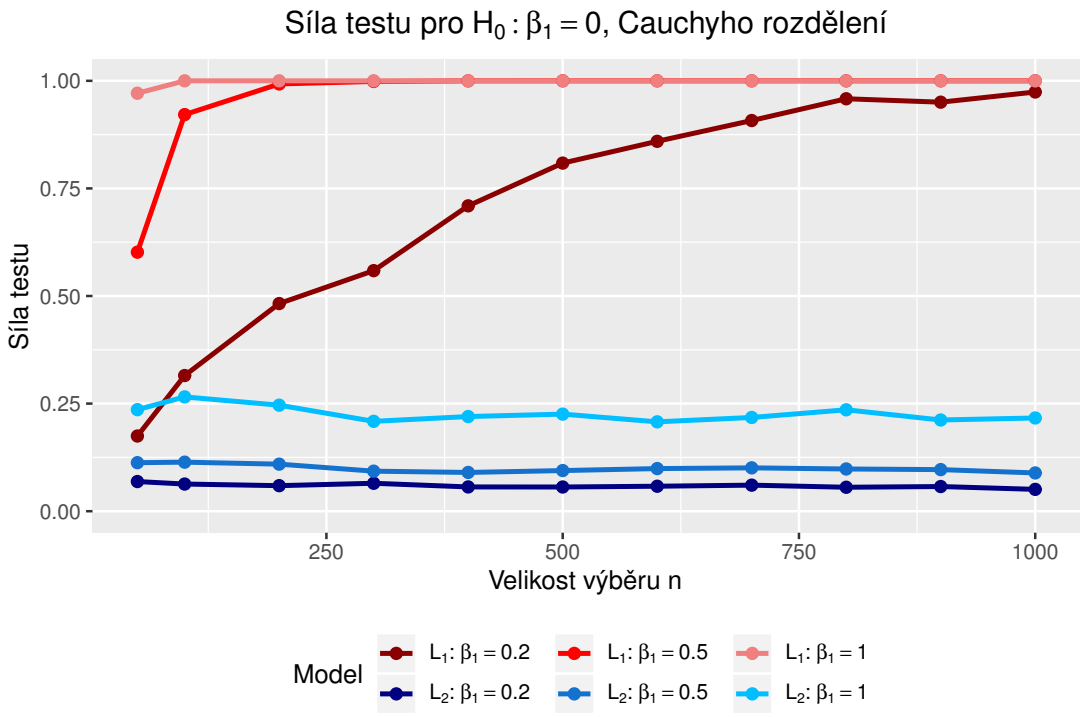
Na Obrázku 3.3 s rozdělením znečištěným odlehlými pozorováními vidíme, že vyšší sílu testu má L_1 metoda, nicméně pro větší výběry a vyšší hodnoty β_1 je síla pro obě metody blízká jedné.

Pro nesymetrické rozdělení na Obrázku 3.4 je patrné, že pro menší velikosti výběru n a pro β_1 blízké nulové hypotéze je výrazně lepší L_1 test, nicméně opět je pro rostoucí n a β_1 síla blízká jedné pro obě metody.

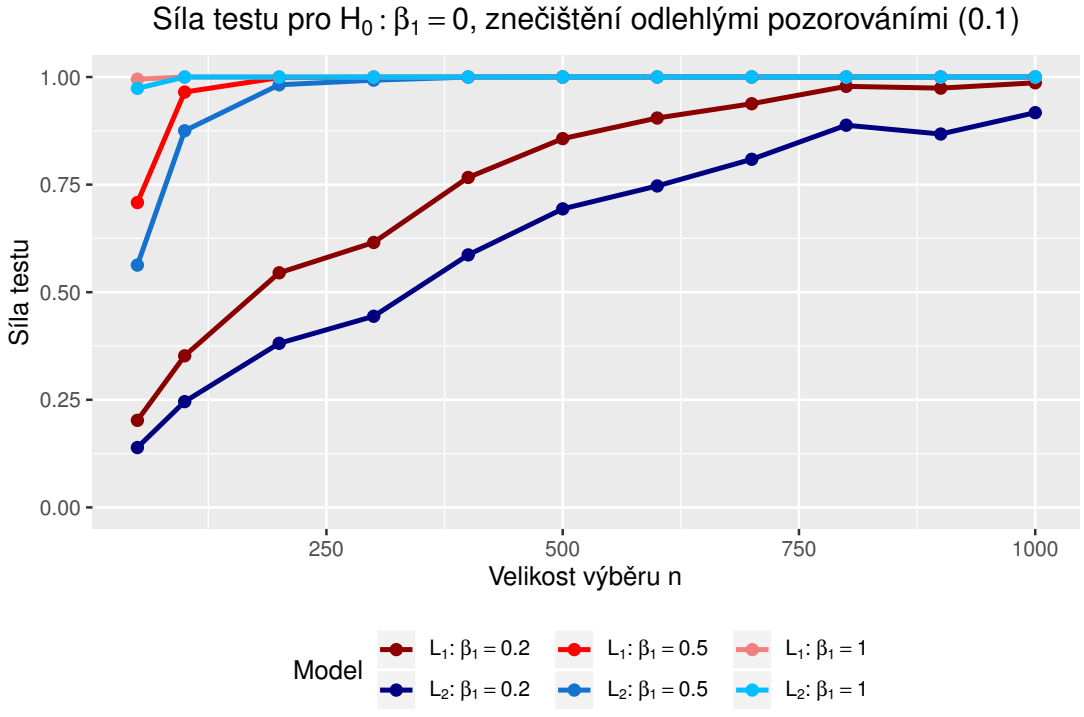
U všech rozdělení kromě Cauchyho je tedy u obou metod patrná konvergence síly testu k jedné pro rostoucí n. U Cauchyho rozdělení tuto vlastnost pozorujeme pouze u L_1 testů.



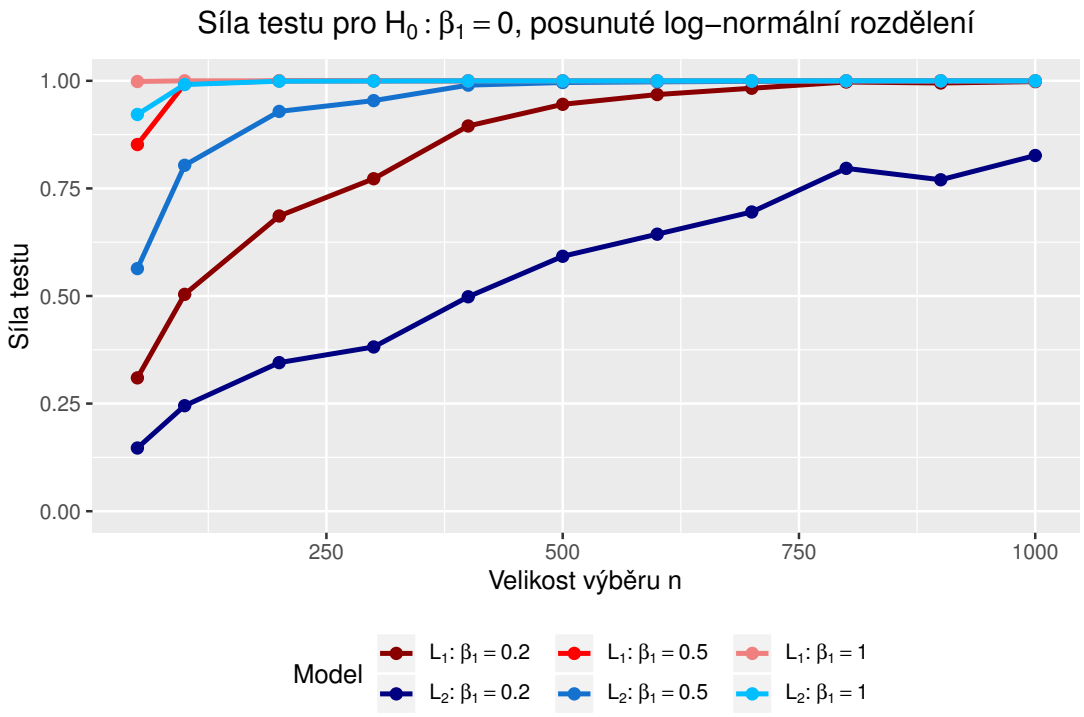
Obrázek 3.1: Závislost síly L_1 a L_2 testů na skutečné hodnotě β_1 a velikosti výběru pro normální rozdělení.



Obrázek 3.2: Závislost síly L_1 a L_2 testů na skutečné hodnotě β_1 a velikosti výběru pro Cauchyho rozdělení.



Obrázek 3.3: Závislost síly L_1 a L_2 testů na skutečné hodnotě β_1 a velikosti výběru pro rozdělení znečištěné odlehlými pozorováními s mírou znečištění 0.1.



Obrázek 3.4: Závislost síly L_1 a L_2 testů na skutečné hodnotě β_1 a velikosti výběru pro log-normální rozdělení.

Poslední metrikou zájmu je pokrytí konfidenčních intervalů, které nezávisí na skutečné hodnotě β_1, proto budeme opět analyzovat pouze výsledky pro $\beta_1 = 1$ shrnuté v Tabulce 3.3. Obdobně jako pro chybu I. typu uvádíme pouze maximální MCE pro každé rozdělení a metodu. Výsledky pro absolutní člen β_0 jsou velmi podobné, proto je již neuvádíme.

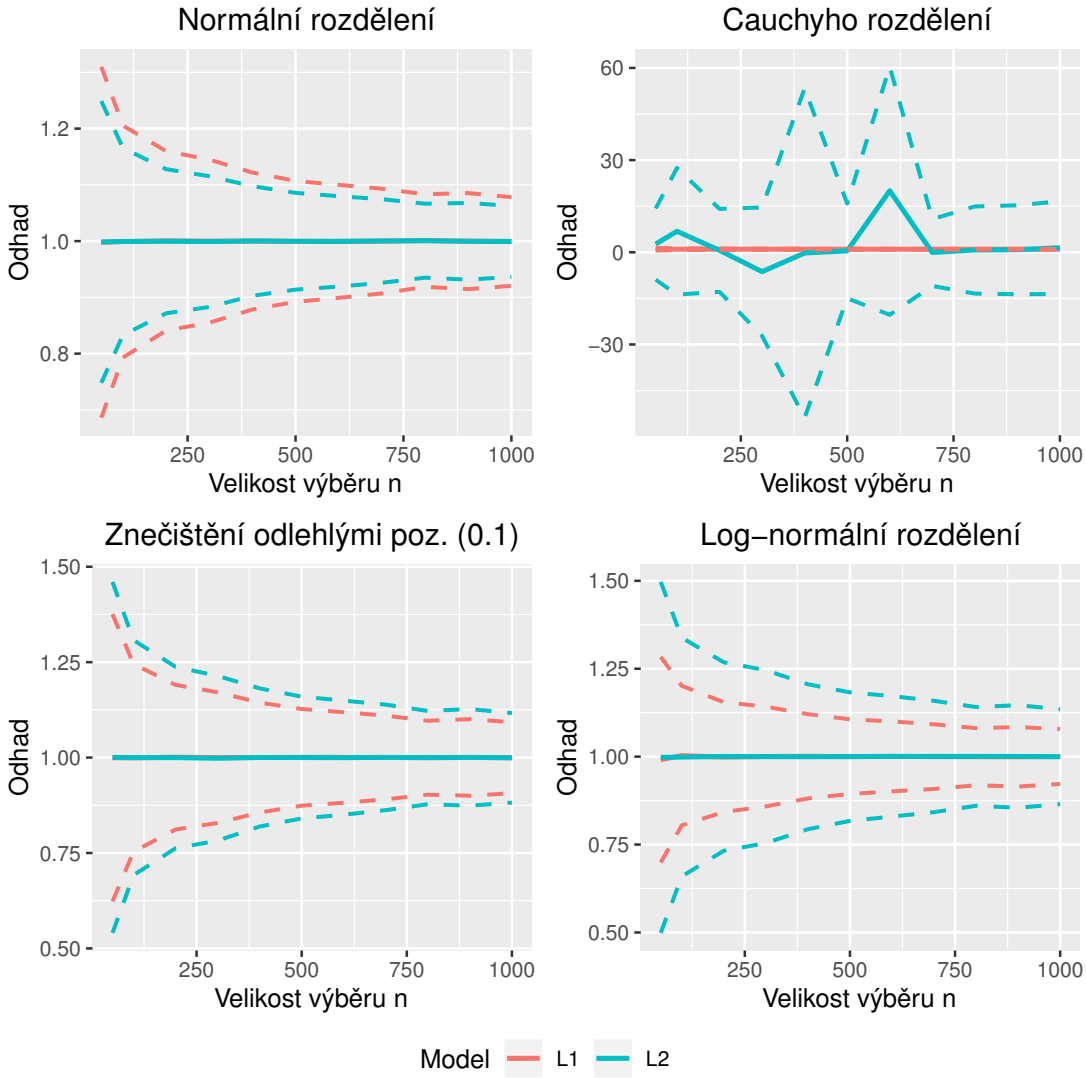
Tabulka 3.3: Srovnání pokrytí konfidenčních intervalů pro $\beta_1 = 1$ v modelu s jedním regresorem.

		$n = 50$	$n = 100$	$n = 500$	MCE
Normální	L_1	92.09 %	92.61 %	94.21 %	*0.27 %*
	L_2	94.82 %	94.91 %	95.36 %	*0.22 %*
Cauchy	L_1	91.92 %	93.22 %	93.96 %	*0.27 %*
	L_2	93.28 %	94.65 %	94.70 %	*0.25 %*
Odlehlá poz.	L_1	91.77 %	92.81 %	94.03 %	*0.27 %*
	L_2	94.10 %	95.05 %	94.50 %	*0.24 %*
Log-normální	L_1	91.31 %	91.26 %	93.94 %	*0.28 %*
	L_2	93.57 %	94.99 %	94.63 %	*0.25 %*

Závěry jsou analogické závěrům pro chybu I. typu, tedy že L_1 konfidenční intervaly jsou antikonzervativní, a mají tak nižší pokrytí než $(1 - \alpha)$. Obzvlášť

pro menší velikosti výběru tak mají L_2 intervaly lepší pokrytí než L_1 intervaly, nicméně u problematických rozdělení jsou výrazně širší než L_1 intervaly, což souvisí s menší sílou testů v grafech výše. Tyto analogie plynou z ekvivalence testů a konfidenčních intervalů.

Na Obrázku 3.5, kde jsou vykresleny průměrné konfidenční intervaly přes všechny iterace v závislosti na velikosti výběru pro jednotlivá rozdělení, je vidět, že pro log-normální rozdělení jsou L_2 intervaly výrazně širší než L_1 intervaly a u Cauchyho rozdělení jsou již řádově širší a nepřináší nám v podstatě žádnou informační hodnotu. U rozdělení s odlehlými pozorováními není rozdíl tak výrazný, ale L_1 intervaly jsou užší.



Obrázek 3.5: Konfidenční intervaly pro $\beta_1 = 0$ v modelu s jedním regresorem.

3.2 Model s více regresory

V druhé části simulační studie pracujeme s modelem s absolutním členem a čtyřmi regresory, tedy s modelem ve tvaru

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i, i = 1, \ldots, n.$$

3.2.1 Nastavení a výpočet simulací

Kvůli snížení výpočetní náročnosti a snazší interpretaci výsledků již budeme uvažovat pouze jednu kombinaci hodnot regresních koeficientů. Na základě poznatků ze simulací s jedním regresorem volíme $\beta_0 = \beta_1 = \beta_2 = 0.3$ a dále $\beta_3 = \beta_4 = 0$. Testovat budeme následující dvě hypotézy:

$$\begin{aligned} H_0 &: \beta_3 = \beta_4 = 0, & H_1 &: \beta_3 \neq 0 \vee \beta_4 \neq 0, \\ H_0' &: \beta_2 = \beta_3 = \beta_4 = 0, & H_1' &: \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0. \end{aligned}$$

Na nulové hypotéze H_0 odhadneme chybu I. typu testu na podmodel (testujeme platnou nulovou hypotézu) a na nulové hypotéze H_0' odhadneme sílu testu na podmodel (testujeme neplatnou nulovou hypotézu). Pro vektor regresních parametrů $\boldsymbol{\beta}$ zkonstruujeme konfidenční množinu a z výsledků simulací odhadneme její pokrytí.

Hodnoty nenulových parametrů jsou určeny tak, aby šlo dobře zkoumat sílu testu. Důležitý je v tomto ohledu zejména parametr $\beta_2 = 0.3$, který by neměl být příliš vzdálený nule, protože bychom pak pro H_0' téměř vždy zamítali a nemohli bychom metody efektivně porovnat.

Pro každé n byly vygenerovány čtyři vektory regresorů $\boldsymbol{X}_{n,j} \sim \mathcal{N}_n(\mathbf{0}, \mathbb{I}_n), j = 1, \ldots, 4$ a pro každou kombinaci β_1 a n bylo spočteno $iter = 10\,000$ iterací následujícím způsobem:

1. Je vygenerován vektor pozic odlehlých pozorování pomocí alternativního rozdělení s parametrem $p = 0.1$.

2. Pro každé rozdělení vygenerujeme vektor náhodných chyb $\boldsymbol{\varepsilon}$ a spočteme odezvu $\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{X}_{n,1} + \beta_2 \boldsymbol{X}_{n,2} + \beta_3 \boldsymbol{X}_{n,3} + \beta_4 \boldsymbol{X}_{n,4} + \boldsymbol{\varepsilon}$.

3. Pro všechny odezvy odhadneme L_1 a L_2 regresní modely. Pro každý z těchto modelů uložíme odhady regresních koeficientů a kovarianční matice a spočteme p-hodnoty nulových hypotéz H_0 a H_0'.

Z výsledků simulací jsou následně odhadnuty zkoumané metriky a jejich MCE.

3.2.2 Analýza výsledků

Pro MSE pro parametr $\beta_1 = 0.3$ v Tabulce 3.4 (uvádíme pouze jeden parametr, protože pro ostatní jsou výsledky analogické) a pro chybu I. typu a sílu testu v Tabulkách 3.5 a 3.6 jsou závěry hodně podobné jako v modelu s jedním regresorem.

Tabulka 3.4: Srovnání střední čtvercové chyby (MSE) L_1 a L_2 odhadů regresních koeficientů pro $\beta_1 = 0.3$ v modelu s více regresory.

		$n = 50$		$n = 100$		$n = 500$	
F_1	L_1	0.026	(*0.00036*)	0.012	(*0.00018*)	0.003	(*0.00005*)
	L_2	0.017	(*0.00023*)	0.008	(*0.00011*)	0.002	(*0.00003*)
F_2	L_1	0.063	(*0.00112*)	0.024	(*0.00038*)	0.005	(*0.00007*)
	L_2	404 029	(*394 890*)	348 331	(*342 754*)	1 168 698	(*1 165 236*)
F_3	L_1	0.039	(*0.00062*)	0.018	(*0.00027*)	0.004	(*0.00006*)
	L_2	0.057	(*0.00089*)	0.028	(*0.00040*)	0.007	(*0.00011*)
F_4	L_1	0.025	(*0.00039*)	0.012	(*0.00017*)	0.003	(*0.00005*)
	L_2	0.076	(*0.00180*)	0.037	(*0.00078*)	0.010	(*0.00016*)

Tabulka 3.5: Srovnání chyby I. typu L_1 a L_2 testů na podmodel v modelu s více regresory.

		$n = 50$	$n = 100$	$n = 500$	MCE
Normální	L_1	8.51 %	8.21 %	6.09 %	*0.28 %*
	L_2	5.60 %	5.03 %	4.71 %	*0.23 %*
Cauchy	L_1	4.95 %	6.71 %	5.08 %	*0.25 %*
	L_2	4.83 %	6.11 %	4.98 %	*0.24 %*
Odlehlá poz.	L_1	6.85 %	8.05 %	5.94 %	*0.27 %*
	L_2	4.55 %	5.32 %	5.03 %	*0.22 %*
Log-normální	L_1	5.64 %	7.19 %	6.25 %	*0.26 %*
	L_2	4.66 %	5.39 %	5.38 %	*0.23 %*

Tabulka 3.6: Srovnání síly L_1 a L_2 testů na podmodel v modelu s více regresory.

		$n = 50$	$n = 100$	$n = 500$	MCE
Normální	L_1	39.55 %	55.87 %	99.85 %	*0.50 %*
	L_2	47.27 %	71.19 %	100.00 %	*0.50 %*
Cauchy	L_1	14.78 %	29.25 %	95.92 %	*0.45 %*
	L_2	5.58 %	6.81 %	5.85 %	*0.25 %*
Odlehlá poz.	L_1	26.29 %	42.73 %	98.43 %	*0.49 %*
	L_2	16.33 %	25.95 %	88.96 %	*0.44 %*
Log-normální	L_1	33.19 %	51.53 %	99.84 %	*0.50 %*
	L_2	17.56 %	24.60 %	79.07 %	*0.43 %*

U pokrytí konfidenčních množin v Tabulce 3.7 si můžeme všimnout, že u nenormálních rozdělení je pokrytí u L_2 regrese mnohem nižší než u jednorozměrných konfidenčních intervalů v modelu s jedním regresorem. Zatímco jednorozměrné konfidenční intervaly v Tabulce 3.3 mají pro všechna rozdělení požadované pokrytí i u L_2 metody, a to i u poměrně malých velikostí výběru, u modelu s více regresory vidíme, že ani pro $n = 500$ konfidenční množiny nedosahují uspokojivého pokrytí a jejich pokrytí naopak s rostoucí velikostí výběru klesá.

U normálního rozdělení je však vidět, že L_1 konfidenční množiny dosahují požadovaného pokrytí výrazně pomaleji než L_2 regrese, u které lze pro normální rozdělení sestrojit přesné konfidenční množiny a intervaly (nikoliv asymptotické).

Tabulka 3.7: Srovnání pokrytí L_1 a L_2 konfidenčních množin regresních koeficientů v modelu s více regresory.

		$n = 50$	$n = 100$	$n = 500$	MCE
Normální	L_1	85.93 %	87.30 %	92.21 %	*0.35 %*
	L_2	94.84 %	95.15 %	95.15 %	*0.22 %*
Cauchy	L_1	88.88 %	89.32 %	92.80 %	*0.31 %*
	L_2	1.20 %	0.29 %	0.00 %	*0.11 %*
Odlehlá poz.	L_1	86.41 %	88.01 %	91.55 %	*0.34 %*
	L_2	36.63 %	35.62 %	33.32 %	*0.48 %*
Log-normální	L_1	81.25 %	84.05 %	90.82 %	*0.39 %*
	L_2	35.90 %	30.50 %	23.17 %	*0.48 %*

V Tabulce 3.8 se můžeme podívat také na jednorozměrné konfidenční intervaly v modelu s více regresory, a to konkrétně u koeficientu $\beta_1 = 0.3$ (ostatní koeficienty dávají téměř identické výsledky). I v tomto případě pokrytí intervalů s rostoucí velikostí výběru klesá. Pokrytí u intervalu je dle očekávání vyšší než u konfidenčních množin, nicméně je výrazně nižší než v modelu s jedním regresorem. I když tedy pro model s jedním regresorem dávala L_2 regrese alespoň pro rozdělení s odlehlými pozorováními a log-normálním rozdělením rozumné výsledky, ve vyšších dimenzích již selhává.

Tabulka 3.8: Srovnání pokrytí L_1 a L_2 konfidenčního intervalu β_1 v modelu s více regresory.

		$n = 50$	$n = 100$	$n = 500$	MCE
Normální	L_1	92.45 %	92.81 %	94.13 %	*0.26 %*
	L_2	94.95 %	94.84 %	94.65 %	*0.23 %*
Cauchy	L_1	93.10 %	93.36 %	94.46 %	*0.25 %*
	L_2	22.91 %	16.57 %	6.64 %	*0.42 %*
Odlehlá poz.	L_1	92.26 %	92.66 %	94.12 %	*0.27 %*
	L_2	72.57 %	71.35 %	70.59 %	*0.46 %*
Log-normální	L_1	90.99 %	92.10 %	93.64 %	*0.29 %*
	L_2	72.08 %	68.56 %	63.99 %	*0.48 %*

Závěr

V diplomové práci jsme v Kapitole 2 shrnuli základní teorii pro L_1 regresi, která se nabízí jako alternativa k metodě nejmenších čtverců, nejběžnější metodě odhadu v regresním modelu. Hlavní větou teoretické části je Věta 7, kde je dokázána asymptotická normalita L_1 odhadu regresních koeficientů. Asymptotická normalita je zásadní pro odvození nástrojů pro inferenci v podkapitole 2.5, které jsou poté využity v simulační studii v Kapitole 3.

Teorie uvedená v práci byla odvozena a publikována zejména Rogerem Koenkerem, nicméně v práci jsou podrobněji odvozeny některé kroky důkazů, které byly v dosavadních publikacích jen zmíněné, zejména pak některé kroky v důkazu Věty 7. Dalším vlastním přínosem práce je simulační studie, která srovnává vlastnosti L_1 odhadů s vlastnostmi L_2 odhadů v lineárním regresním modelu.

Simulační studie ukázala, že metoda nejmenších čtverců dává dle očekávání lepší výsledky v případě normálního modelu – odhady mají menší čtvercovou chybu, testy mají vyšší sílu (resp. síla rychleji konverguje k jedné) a konfidenční intervaly mají vyšší pokrytí. L_1 regrese vede na antikonzervativní testy, zatímco L_2 testy i pro konečné výběry dobře drží hladinu. V případě rozdělení s těžkými chvosty dává výrazně lepší výsledky L_1 regrese. U Cauchyho rozdělení metoda nejmenších čtverců nedává dobrý smysl, protože rozdělení nemá střední hodnotu, což potvrzují výsledky simulací. Pro rozdělení s příměsí odlehlých pozorování sice dává L_2 regrese v modelu s jedním regresorem uspokojivé výsledky, nicméně L_1 regrese má lepší vlastnosti. V modelu s více regresory je pak jednoznačně lepší L_1 regrese, protože L_2 regrese selhává v případě konfidenčních množin i intervalů, které nemají požadované pokrytí. Stejné závěry lze udělat i pro posunuté log-normální rozdělení, které bylo uvažováno jako příklad nesymetrického rozdělení.

U všech zkoumaných rozdělení L_1 regrese vede na antikonzervativní testy a konfidenční intervaly, resp. množiny, nicméně všechna uvažovaná rozdělení splňují její předpoklady, a je tak potvrzeno očekávání, že odhady, testy a konfidenční intervaly, resp. konfidenční množiny mají uspokojivé vlastnosti. Pokud tedy máme podezření, že náhodné chyby mohou být z rozdělení s těžkými chvosty či máme v datech odlehlá pozorování, je vhodné použít L_1 regresi i přesto, že za ideálních podmínek nemá tak dobré vlastnosti jako L_2 regrese. Doporučení je o to silnější, je-li cílem inference v modelu s více než jedním regresorem.

V diplomové práci byly odvozeny a použity pouze testy a konfidenční intervaly (resp. množiny) Waldova typu, které byly srovnávány s analogickými nástroji pro inferenci L_2 regrese. Práce by se tedy dala rozšířit i o další možné přístupy pro inferenci L_1 regrese, které jsou stručně představené v podkapitole 2.5.3.

Seznam použité literatury

Bassett, G. a Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, **73**(363), 618–622.

Bhattacharya, R., Lin, L. a Patrangenaru, V. (2016). *A Course in Mathematical Statistics and Large Sample Theory*. doi: 10.1007/978-1-4939-4032-5.

Chen, X. a Wu, Y. (1993). On a necessary condition for the consistency of the l1 estimates in linear regression models. *Communications in Statistics - Theory and Methods*, **22**(3), 631–639.

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

Gutenbrunner, C., Jurečková, J., Koenker, R. a Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, **2**(4), 307–331.

Hall, P. a Sheather, S. J. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society. Series B (Methodological)*, **50**(3), 381–391.

Koenker, R., Chesher, A. a Jackson, M. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. ISBN 9780521608275.

Koenker, R. (2019). *quantreg: Quantile Regression*. URL `https://CRAN.R-project.org/package=quantreg`. R package version 5.54.

Koenker, R. a Bassett, G. (1978). Regression quantiles. *Econometrica*, **46** (1), 33–50.

Koenker, R. a Bassett, G. (1982). Tests of linear hypotheses and l1 estimation. *Econometrica*, **50**(6), 1577–1583.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`. R version 3.6.0.

Rao, C. (1973). *Linear statistical inference and its applications*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley. ISBN 9780471708230.

Wagner, H. M. (1959). Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, **54**(285), 206–212.

Zhao, L., Rao, C. a Chen, X. (1993). A note on the consistency of M-estimates in linear models. In *Stochastic Processes: A Festschrift in Honour of Gopinath Kallianpur*. Springer.

Seznam obrázků

Seznam tabulek