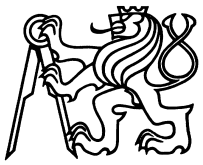
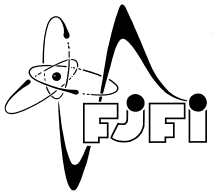


České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Robustní odhady a testy v modelech poissonovské regrese

Robust Estimation and Inference in Poisson Regression Models

Diplomová práce

Autor: **Bc. Jana Novotná**

Vedoucí práce: **doc. Ing. Tomáš Hobza, Ph.D.**

Akademický rok: **2020/2021**

Katedra: matematiky Akademický rok: 2019/2020

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student:	Bc. Jana Novotná
Studijní program:	Aplikace přírodních věd
Studijní obor:	Aplikované matematicko-stochastické metody
Název práce (česky):	Robustní odhady a testy v modelech poissonovské regrese
Název práce (anglicky):	Robust Estimation and Inference in Poisson Regression Models

Pokyny pro vypracování:

1) Seznamte se s metodami robustního odhadování a testování v zobecněných lineárních modelech, zaměřte se zejména na model poissonovské regrese.
2) V rámci modelů poissonovské regrese pokračujte ve studiu zobecněného mediánového odhadu, navrženého ve Vašem výzkumném úkolu. Pomocí přístupu v práci Hobza et al. (2017) se pokuste odvodit asymptotické rozdělení tohoto odhadu.
3) Na základě zobecněného mediánového odhadu navrhněte vhodnou testovací statistiku pro robustní testy hypotéz o parametrech modelu poissonovské regrese.
4) Pomocí odvozeného asymptotického rozdělení zobecněného mediánového odhadu, případně pomocí vhodně zvolených simulací, stanovte asymptotické rozdělení navržené testovací statistiky.
5) Celou metodiku implementujte a pomocí simulačních experimentů ověřte platnost odvozených asymptotických výsledků a prostudujte empirické chování navržených testů. Zaměřte se na jejich citlivost na odlehlá pozorování, případně pákové body.

Doporučená literatura:

1) T. Hobza, N. Martín, L. Pardo, A Wald-type test statistic based on robust modified median estimator in logistic regression models. Journal of Statistical Computation and Simulation 87(12), 2017, 2309–2333.

2) E. Cantoni, E. Ronchetti, Robust Inference for Generalized Linear Models. Journal of the American Statistical Association 96 (455), 2001, 1022–1030.

3) M. Valdora, V. J. Yohai, Robust estimators for generalized linear models. Journal of Statistical Planning and Inference 146, 2014, 31–48.

4) S. N. Lo, E. Ronchetti, Robust and accurate inference for generalized linear models. Journal of Multivariate Analysis 100, 2009, 2126–2136.

5) Ch. E. McCulloch, S. R. Searle, J. M. Neuhaus, Generalized, Linear, and Mixed Models. John Wiley & Sons, New Jersey, 2008.

6) P. Bohuslav, Robustní odhady v zobecněných lineárních modelech. Diplomová práce, Katedra matematiky FJFI ČVUT, Praha, 2016.

Jméno a pracoviště vedoucího diplomové práce:

doc. Ing. Tomáš Hobza, Ph.D.
Katedra matematiky, FJFI ČVUT, Trojanova 13, 120 00 Praha 2

Jméno a pracoviště konzultanta:

Datum zadání diplomové práce: 28.2.2020

Datum odevzdání diplomové práce: 6.1.2021

Doba platnosti zadání je dva roky od data zadání.

Poděkování:
Chtěla bych zde poděkovat především svému školiteli doc. Ing. Tomáši Hobzovi, Ph.D. za ochotu, trpělivost, podnětné rady a odborné i lidské zázemí při vedení mé diplomové práce.

Čestné prohlášení:
Prohlašuji, že jsem tuto práci vypracovala samostatně a uvedla jsem všechnu použitou literaturu.

V Praze dne 6. ledna 2021 Bc. Jana Novotná

Název práce:

Robustní odhady a testy v modelech poissonovské regrese

Autor: Bc. Jana Novotná

Obor: Aplikované matematicko-stochastické metody

Druh práce: Diplomová práce

Vedoucí práce: doc. Ing. Tomáš Hobza, Ph.D., Katedra matematiky, FJFI ČVUT

Abstrakt: Tato práce se zabývá robustním odhadováním a testováním v modelech poissonovské regrese. Největší pozornost je zde věnována nově zavedenému zobecněnému mediánovému odhadu. Po jeho definici a představení několika existujících metod odhadu je v rámci teoretické části práce odvozeno asymptotické rozdělení zobecněného mediánového odhadu. Dále je navržena na něm založená testovací statistika Waldova typu pro testy hypotéz o parametrech modelu a je odvozeno její asymptotické rozdělení. V simulační části práce je provedeno několik simulací, které ověřují odvozené teoretické výsledky a porovnávají vlastnosti nového odhadu při testování hypotéz s již existujícími metodami odhadu. Ukazuje se, že námi zavedený odhad jim dokáže konkurovat a v některých případech dokáže být dokonce lepší. V rámci této práce je tedy odvozena potřebná teorie pro využití zobecněného mediánového odhadu a provedené simulace ukazují, že lze tento odhad použít v praxi.

Klíčová slova: Model poissonovské regrese, robustní odhady, robustní testy hypotéz, statistika Waldova typu, zobecněný mediánový odhad

Title:

Robust Estimation and Inference in Poisson Regression Models

Author: Bc. Jana Novotná

Abstract: This thesis deals with robust estimation and inference in Poisson regression models. The greatest attention is paid to a newly introduced modified median estimator. After its definition and an introduction of several existing estimation methods, an asymptotic distribution of the modified median estimate is derived in the theoretical part of the thesis. Furthermore, Wald-type test statistic based on the modified median estimator for tests of hypotheses about model parameters is proposed and its asymptotic distribution is derived. In the simulation part of the thesis, several simulations are performed. These simulations verify the derived theoretical results and compare the properties of the new estimator with existing estimation methods when testing hypotheses. It turns out that our estimator can compete with them and in some cases can be even better. In this thesis, the necessary theory for the use of the modified median estimator is derived and the performed simulations show that this estimator can be applied in practice.

Key words: Poisson regression model, robust estimators, robust hypothesis testing, Wald-type test statistic, modified median estimator

Obsah

Úvod **7**

1 Zobecněné lineární modely **9**
1.1 Exponenciální rodina distribucí 9
1.2 Zobecněné lineární modely 11
1.3 Poissonovská regrese 11

2 Odhady parametrů **13**
2.1 Maximálně věrohodný odhad 13
2.2 Mediánový odhad 14
2.3 Zobecněný mediánový odhad 18
2.4 Mallowsův odhad 19
2.5 M-odhad založený na transformaci odezvy 20

3 Asymptotické rozdělení zobecněného mediánového odhadu **22**
3.1 Značení 22
3.2 Pomocné výpočty 22
3.3 Centrální předpoklady 25
3.4 Věta o asymptotickém rozdělení 26
3.5 Konzistentní odhad asymptotické kovarianční matice 30

4 Testování hypotéz **32**

5 Simulační experimenty **36**
5.1 Model 36
5.2 Konzistence zobecněného mediánového odhadu 37
5.3 Algoritmus experimentů na testování hypotéz 39
5.4 Chyba 1. druhu v závislosti na počtu pozorování 39
5.5 Testování hypotéz pro čistá data 41
5.6 Vliv odlehlých pozorování 44
5.7 Vliv pákových bodů 49
5.8 Poznámka k simulačním experimentům 55

Závěr **56**

Literatura **58**

Příloha: Tabulka hodnot C_k **59**

Úvod

Jedním z hojně využívaných nástrojů datové analýzy jsou zobecněné lineární modely. Abychom je mohli v praxi využít, je naším prvním úkolem odhadnout jejich parametry. K tomu je často využíván maximálně věrohodný odhad. U tohoto odhadu je ale známo, že je velmi citlivý na přítomnost znečištění v datech. Je tedy potřeba mít k dispozici i jiné metody odhadu, které jsou robustnější než maximálně věrohodný odhad. Mnoho prací se zabývá výzkumem robustních odhadů parametrů modelů logistické regrese, pomocí kterých modelujeme binární proměnné. Vývoji robustních metod odhadů parametrů pro zobecněné lineární modely, které popisují jiné než binární proměnné, se nevěnuje taková pozornost.

Pokud již máme vhodnou robustní metodu odhadu, potřebujeme být schopni pomocí ní provést testy hypotéz o parametrech modelu. Je tedy třeba navrhnout vhodnou testovací statistiku a určit její rozdělení. Poté už můžeme metodu používat v praxi a využít všechny možnosti, které nám nabízejí zobecněné lineární modely.

Tato práce se proto zaměřuje na robustní odhady parametrů a testy hypotéz pro modely poissonovské regrese, které, jak název napovídá, popisují proměnné, které mají Poissonovo rozdělení. Poissonovo rozdělení může například popisovat počet stížností za rok nebo počet nezaměstnaných v určité oblasti.

Inspirujeme se v oblasti odhadů parametrů pro modely logistické regrese a definujeme nový odhad pro modely poissonovské regrese. V práci [9] byl představen mediánový odhad pro model logistické regrese. V článku [8] se autorům podařilo mediánový odhad ještě významně vylepšit a vznikl tak zobecněný mediánový odhad. Obdoba mediánového odhadu pro modely poissonovské regrese již byla představena v práci [3]. My nyní navážeme na tuto práci a pomocí myšlenek z článku [8] zavedeme zobecněný mediánový odhad pro modely poissonovské regrese a navrhneme vhodný test pro testování hypotéz o parametrech modelu.

První kapitola nám poslouží jako úvod do problematiky zobecněných lineárních modelů. Zadefinujeme si nejdříve obecně zobecněné lineární modely a pak se zaměříme na modely poissonovské regrese, kterým se věnuje tato práce.

Ve druhé kapitole se přesuneme k metodám odhadu parametrů modelu poissonovské regrese. Nejprve si představíme maximálně věrohodný odhad a mediánový odhad. Poté navážeme na mediánový odhad a odvodíme novou metodu odhadu – zobecněný mediánový odhad. Na závěr kapitoly se seznámíme se dvěma již existujícími robustními metodami odhadu parametrů, konkrétně Mallowsovým odhadem a M-odhadem založeným na transformaci odezvy.

Třetí kapitola je věnována odvození asymptotického rozdělení zobecněného mediánového odhadu. V první části provedeme potřebné výpočty a na závěr kapitoly se zaměříme na konzistentní odhad odvozené teoretické asymptotické kovarianční matice, který je důležitý pro následné využití odvozených teoretickým poznatků v praxi.

Čtvrtá kapitola využívá výsledky odvozené ve třetí kapitole a zaměřuje se na testování hypotéz. Nejprve zde navrhneme testovací statistiku Waldova typu a poté určíme její asymptotické

rozdělení. Ilustrujeme si zde také, jak využít navržený test v praxi.

Poté se již přesuneme k simulačním experimentům, kterým se věnuje celá pátá kapitola. Představíme si nejprve model, se kterým budeme pracovat, a poté provedeme několik simulačních experimentů. První z nich ilustruje konzistenci zobecněného mediánového odhadu. Další experimenty se již věnují testování hypotéz a uvažují pro porovnání kromě zobecněného mediánového odhadu také maximálně věrohodný odhad, Mallowsův odhad a M-odhad založený na transformaci odezvy. Nejprve provedeme experiment, v rámci kterého budeme studovat chybu 1. druhu v závislosti na počtu pozorování. Dále budeme testovat hypotézy pro čistá data a také budeme zkoumat vliv přítomnosti znečištění na testování hypotéz, konkrétně budeme uvažovat znečištění odlehlými pozorováními a pákovými body.

Kapitola 1

Zobecněné lineární modely

Tato kapitola je věnována úvodu do problematiky zobecněných lineárních modelů. Nejprve si představíme exponenciální rodinu distribucí, kterou dále využijeme při definici zobecněných lineárních modelů. Ty definujeme nejprve obecně a poté se zaměříme na modely poissonovské regrese, kterými se zabývá celá tato práce. První dvě podkapitoly jsou částečně převzaty z práce [12].

1.1 Exponenciální rodina distribucí

Mějme náhodnou veličinu Y, jejíž rozdělení závisí pouze na jednom parametru $\theta \in \mathbb{R}$. Říkáme, že rozdělení veličiny Y patří do exponenciální rodiny distribucí, pokud lze hustotu pravděpodobnosti, případně pravděpodobnostní funkci, veličiny Y zapsat ve tvaru

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}, \tag{1.1}$$

kde $a(y), b(\theta), s(y)$ a $t(\theta)$ jsou známé funkce a $s(y) = \exp\{d(y)\}, t(\theta) = \exp\{c(\theta)\}$.

Pokud platí, že $a(y) = y$, říkáme, že distribuce je v kanonickém tvaru a $b(\theta)$ se nazývá přirozený parametr rozdělení.

Vyskytují-li se v rozdělení ještě jiné parametry než námi uvažovaný parametr θ, nazýváme je šumovými parametry a pohlížíme na ně jako na konstanty. Tyto parametry mohou být zahrnuty v libovolné funkci $a(y), b(\theta), c(\theta)$ nebo $d(y)$.

Do exponenciální rodiny distribucí patří například normální, exponenciální, Poissonovo, gama rozdělení nebo třeba binomické a Bernoulliho rozdělení. Všechna výše jmenovaná rozdělení mají kanonický tvar. My si přepis do kanonického tvaru představíme pouze pro normální, binomické, Bernoulliho a Poissonovo rozdělení.

Normální rozdělení

Normální rozdělení je jedno z nejpoužívanějších rozdělení pro spojitou symetrickou náhodnou veličinu. Značíme ho $\mathcal{N}(\mu, \sigma^2)$ a pro jeho hustotu pravděpodobnosti platí

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} = \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}, \tag{1.2}$$

kde $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ a na $\sigma^2 > 0$ nyní pohlížíme jako na šumový parametr. Vidíme, že platí $a(y) = y$, tudíž jsme získali kanonický tvar a pro přirozený parametr rozdělení platí $b(\mu) = \mu/\sigma^2$. Pro zbylé dvě funkce platí $c(\mu) = -\mu^2/(2\sigma^2) - \frac{1}{2}\ln(2\pi\sigma^2)$ a $d(y) = -y^2/(2\sigma^2)$.

Binomické rozdělení

Binomické rozdělení používáme, pokud chceme popsat četnost výskytu náhodného jevu v n nezávislých pokusech za předpokladu, že pravděpodobnost výskytu jevu je ve všech pokusech rovna $\pi \in (0,1)$. Binomické rozdělení $\mathrm{Bi}(n,\pi)$ má pravděpodobnostní funkci rovnu

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \exp\left\{y\ln\pi - y\ln(1-\pi) + n\ln(1-\pi) + \ln\binom{n}{y}\right\}, \tag{1.3}$$

kde y nabývá hodnot $0,1,2,\ldots,n$. Pravděpodobnostní funkce je v kanonickém tvaru a pro přirozený parametr platí $b(\pi) = \ln[\pi/(1-\pi)]$.

Bernoulliho rozdělení

Bernoulliho rozdělení je speciální případ binomického rozdělení, kdy místo n pokusů uvažujeme pouze pokus jeden. Pravděpodobnostní funkce Bernoulliho rozdělení $\mathrm{Be}(\pi)$ má tvar

$$f(y;\pi) = \pi^y(1-\pi)^{1-y} = \exp\left\{y\ln\pi - y\ln(1-\pi) + \ln(1-\pi)\right\}, \quad y \in \{0,1\}. \tag{1.4}$$

Jedná se opět o zápis v kanonickém tvaru a stejně jako v případě binomického rozdělení je přirozený parametr rozdělení roven $b(\pi) = \ln[\pi/(1-\pi)]$.

Poissonovo rozdělení

Poissonovo rozdělení využíváme například pro popis počtu událostí nebo částic v určitém intervalu, ať už časovém, nebo prostorovém. Poissonovo rozdělení $\mathrm{Po}(\lambda)$ má pravděpodobnostní funkci ve tvaru

$$f(y;\lambda) = \frac{\lambda^y}{y!}e^{-\lambda} = \exp\left\{y\ln\lambda - \lambda - \ln(y!)\right\}, \quad y \in \mathbb{N}_0, \quad \lambda > 0. \tag{1.5}$$

Pravděpodobnostní funkci lze tedy zapsat v kanonickém tvaru a pro přirozený parametr rozdělení platí $b(\lambda) = \ln\lambda$. Pro zbylé dvě funkce pak platí $c(\lambda) = -\lambda$ a $d(y) = -\ln(y!)$.

Vlastnosti exponenciální rodiny distribucí

Pro exponenciální rodinu distribucí se dají odvodit vzorce (viz [5]) pro výpočet střední hodnoty a rozptylu funkce $a(Y)$, které platí pro libovolné rozdělení z této rodiny. Vzorce jsou tvaru

$$\mathrm{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}, \tag{1.6}$$

$$\mathrm{var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}. \tag{1.7}$$

Pokud máme rozdělení kanonického tvaru, je $a(Y) = Y$, a tedy využitím uvedených vzorců získáme střední hodnotu a rozptyl přímo náhodné veličiny Y.

Například pro Poissonovo rozdělení platí $b'(\lambda) = 1/\lambda$, $b''(\lambda) = -1/\lambda^2$, $c'(\lambda) = -1$ a $c''(\lambda) = 0$. Pro střední hodnotu a rozptyl veličiny $Y \sim \mathrm{Po}(\lambda)$ pak pomocí vztahů (1.6) a (1.7) získáme

$$\mathrm{E}(Y) = -\frac{-1}{\frac{1}{\lambda}} = \lambda, \tag{1.8}$$

$$\mathrm{var}(Y) = \frac{-\frac{1}{\lambda^2}\cdot(-1) - 0\cdot\frac{1}{\lambda}}{\left(\frac{1}{\lambda}\right)^3} = \lambda. \tag{1.9}$$

1.2 Zobecněné lineární modely

V této části si zadefinujeme obecně zobecněné lineární modely a v následující části se zaměříme na modely poissonovské regrese, kterými se budeme zabývat i dále v práci.

Nejprve si definujme klasický lineární model. Lineární model má tvar

$$Y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, 2, \ldots, n, \tag{1.10}$$

kde Y_i, $i = 1, 2, \ldots, n$, jsou nezávislé náhodné proměnné, které se nazývají vysvětlované. Vektorem vysvětlujících proměnných či vektorem regresorů nazýváme $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i1}, x_{i2}, \ldots, x_{id})$ a považujeme ho za vektor známých hodnot. Vektor $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d)^{\mathrm{T}}$ se nazývá vektorem neznámých koeficientů (či parametrů), případně vektorem regresních koeficientů. Veličiny e_i, $i = 1, 2, \ldots, n$, se nazývají náhodné chyby a předpokládáme o nich, že jsou nezávislé.

Klasický lineární model má alternativní zápis

$$\mathrm{E}(Y_i) = \mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \quad Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad Y_i \text{ nezávislé}, \quad i = 1, 2, \ldots, n. \tag{1.11}$$

Zobecněný lineární model nabízí, jak název napovídá, jisté zobecnění oproti lineárnímu modelu. První rozdíl je, že nemusíme modelovat pouze střední hodnotu Y_i. V případě zobecněného lineárního modelu modelujeme funkci střední hodnoty veličiny Y_i, tj. modelujeme

$$g\big(\mathrm{E}(Y_i)\big) = g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}. \tag{1.12}$$

Funkce g se nazývá spojovací funkce a předpokládáme, že je monotónní a diferencovatelná.

Další zobecnění spočívá v tom, že uvažujeme Y_i s rozdělením z exponenciální rodiny distribucí. Netrváme tedy na normálním rozdělení. Vysvětlované proměnné musí být ale stejného typu rozdělení v kanonickém tvaru. Jednotlivá rozdělení ale mohou mít různý parametr θ_i. Například tedy platí $Y_i \sim \mathrm{Po}(\theta_i)$, $i = 1, 2, \ldots, n$.

1.3 Poissonovská regrese

Model poissonovské regrese je zobecněný lineární model pro nezávislá pozorování $Y_1, Y_2, \ldots, Y_n$, pro který platí

$$Y_i \sim \mathrm{Po}(\mu_i), \quad \mathrm{E}(Y_i) = \mu_i, \quad i = 1, 2, \ldots, n, \tag{1.13}$$

$$g(\mu_i) = \ln \mu_i, \quad i = 1, 2, \ldots, n. \tag{1.14}$$

Jako spojovací funkci tedy volíme přirozený logaritmus, což je přirozený parametr Poissonova rozdělení. Díky tomu je logaritmus věrohodnostní funkce, o které se zmíníme v části o maximálně věrohodném odhadu, konkávní a má tedy pouze jedno maximum, které lze nalézt pomocí numerických metod na hledání maxima.

Pro μ_i předpokládáme, že platí

$$\mu_i = s_i \lambda_i, \tag{1.15}$$

kde s_i značí známou velikost vzorku. Střední hodnotu μ_i tedy modelujeme jako funkci vektoru nezávislých proměnných $\boldsymbol{x}_i$ a velikosti vzorku, ze kterého bylo Y_i získáno. Model poissonovské regrese má tedy tvar

$$\ln \mu_i = \ln s_i + \ln \lambda_i = \ln s_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n. \tag{1.16}$$

Pokud ze vztahu (1.16) vyjádříme μ_i, získáme vztah

$$\mu_i = \mu_i(\boldsymbol{\beta}) = \mu(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = s_i \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}. \tag{1.17}$$

Model poissonovské regrese nám umožňuje modelovat proměnnou, která má význam například počtu událostí nebo částic v určitém časovém nebo prostorovém intervalu. Vysvětlovaná proměnná může tedy být např. počet virů v roztoku nebo počet nezaměstnaných v určité oblasti. Pro lepší pochopení uveďme konkrétní model poissonovské regrese.

Modelujeme počet stížností na lékaře pracujících na pohotovosti za rok, což je naše vysvětlovaná proměnná Y. U každého lékaře známe počet provedených ošetření za rok, což je velikost vzorku s. Dále máme čtyři vysvětlující proměnné neboli regresory: roční plat (x_2), počet odpracovaných hodin (x_3), pohlaví (x_4) a absolvování speciální stáže na pohotovosti (x_5). Tyto regresory společně s regresorem reprezentujícím intercept ($x_1 = 1$) nám tvoří vektor vysvětlujících veličin $\boldsymbol{x}$. Vektor $\boldsymbol{\beta}$, který odhadujeme, má tedy 5 složek a pro $Y \sim \mathrm{Po}(\mu)$ platí

$$\ln \mu = \ln s + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 = \ln s + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}. \tag{1.18}$$

Pomocí takového modelu se můžeme například snažit zjistit, které vysvětlující proměnné mají vliv na počet stížností na lékaře.

Výhodou modelu poissonovské regrese je, že má velmi dobře interpretovatelné parametry. Uvažujme j-tý regresor x_j, ke kterému náleží parametr β_j. Pokud necháme ostatní regresory nezměněné, tak veličina $\exp\{\beta_j\}$ značí relativní riziko spojené s expozicí, pokud je x_j binární, nebo relativní riziko spojené se zvýšením x_j o jednu jednotku, pokud je x_j spojitý regresor. Jinými slovy nám veličina $\exp\{\beta_j\}$ udává relativní nárůst střední hodnoty vysvětlované veličiny pro individuum vystavené riziku (regresor $x_j = 1$) oproti individuu nevystavenému riziku (regresor $x_j = 0$) pro binární proměnnou, případně pro spojitou proměnnou udává relativní nárůst střední hodnoty vysvětlované veličiny v případě, kdy se hodnota regresoru x_j zvýší o 1.

Kapitola 2

Odhady parametrů

V této kapitole se budeme zabývat odhady parametrů modelů poissonovské regrese. Nejprve si představíme maximálně věrohodný odhad a poté přejdeme k zástupcům robustních odhadů. Seznámíme se s mediánovým odhadem a zavedeme nový odhad – zobecněný mediánový odhad. Nakonec si představíme Mallowsův odhad a M-odhad založený na transformaci odezvy.

2.1 Maximálně věrohodný odhad

Maximálně věrohodný odhad (maximum likelihood estimate – MLE) je založen na maximalizaci věrohodnostní (případně logaritmické věrohodnostní) funkce, která je definována pomocí realizací $y_1, y_2, \ldots, y_n$ náhodných veličin $Y_1, Y_2, \ldots, Y_n$. Pro náš model má věrohodnostní funkce tvar

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\mu_i(\boldsymbol{\beta})^{y_i}}{y_i!} \exp\{-\mu_i(\boldsymbol{\beta})\} \tag{2.1}$$

a logaritmická věrohodnostní funkce je tvaru

$$l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \left[\frac{\mu_i(\boldsymbol{\beta})^{y_i}}{y_i!} \exp\{-\mu_i(\boldsymbol{\beta})\} \right]. \tag{2.2}$$

Maximálně věrohodný odhad je tedy v našem modelu dán předpisem

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} d_i(\boldsymbol{\beta}), \tag{2.3}$$

kde

$$\begin{aligned} d_i(\boldsymbol{\beta}) &= -\ln \left[\frac{\mu_i(\boldsymbol{\beta})^{y_i}}{y_i!} \exp\{-\mu_i(\boldsymbol{\beta})\} \right] \\ &= -y_i \ln\big(\mu_i(\boldsymbol{\beta})\big) + \ln(y_i!) + \mu_i(\boldsymbol{\beta}). \end{aligned} \tag{2.4}$$

Hledání argumentu minima probíhá tak, že výraz $\sum_{i=1}^{n} d_i(\boldsymbol{\beta})$ zderivujeme jednotlivě podle všech složek parametru $\boldsymbol{\beta}$ a tyto parciální derivace položíme rovny nule. Vznikne nám tak soustava d rovnic. Může se stát, že nejsme schopni nalézt analytické řešení této soustavy a musíme využít numerické řešení. Mezi používané metody k nalezení numerického řešení patří například Newton-Raphsonova metoda nebo Fisher-scoring algoritmus. Obě tyto metody jsou popsány v práci [11].

Maximálně věrohodný odhad je v praxi velmi používaný, protože je asymptoticky eficientní. Je ale známo, že je MLE citlivý na znečištěná data a není tudíž robustní. Představíme si proto další odhady, které by neměly být tak citlivé na znečištění dat. V simulační části práce porovnáme chování MLE s většinou těchto odhadů a ověříme, zda jsou opravdu robustnější než MLE.

2.2 Mediánový odhad

Mediánový odhad pro modely poissonovské regrese při svém vzniku vycházel z mediánového odhadu pro modely logistické regrese, který byl zaveden v práci [9]. Mediánový odhad pro modely poissonovské regrese byl publikován v práci [3]. My si ho zde představíme i s odvozením, abychom se seznámili s jeho hlavními myšlenkami.

Nejprve si ale definujme M-odhady. Později uvidíme, že mediánový odhad také patří mezi M-odhady neboli odhady maximálně věrohodného typu. Tyto odhady vznikly jako reakce na citlivost MLE na znečištěná data a jsou tedy obecně robustnější než MLE. M-odhady pro modely poissonovské regrese jsou definovány předpisem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Phi\big(Y_i, \mu(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big), \tag{2.5}$$

kde $\Phi : \mathbb{N}_0 \times (0, +\infty) \to \mathbb{R}$ je vhodně zvolená funkce. Díky této definici a použití funkce Φ jsme schopni zmenšit vliv netypických pozorování na výsledný odhad regresního koeficientu.

Nyní už přejděme k odvození mediánového odhadu. V odvození uvažujeme velikost vzorku $s_i = 1$, $i = 1, 2, \ldots, n$. Pokud bychom chtěli přejít k obecnému s_i, nahradíme λ_i vztahem $s_i\lambda_i$.

Mějme nezávislé náhodné veličiny $Y_1, Y_2, \ldots, Y_n$, pro něž platí

$$Y_i \sim \mathrm{Po}(\lambda_i), \quad i = 1, 2, \ldots, n, \tag{2.6}$$

a které jsou spojeny s vysvětlujícími veličinami prostřednictvím vztahu

$$\lambda_i = \lambda_i(\boldsymbol{\beta}) = \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}, \quad i = 1, 2, \ldots, n. \tag{2.7}$$

Jak název odhadu napovídá, využívá mediánový odhad mediánovou funkci. Mediánovou funkcí rozumíme funkci, která pro dané λ_i vrací medián daného rozdělení. Pokud označíme jako $\tilde{m}(\lambda_i)$ mediánovou funkci pro veličinu $Y_i \sim \mathrm{Po}(\lambda_i)$, $i = 1, 2, \ldots, n$, chtěli bychom mediánový odhad definovat vztahem

$$\hat{\boldsymbol{\beta}}_{\tilde{\mathrm{M}}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \big|Y_i - \tilde{m}\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big|, \tag{2.8}$$

který splňuje definici M-odhadu.

Odvoďme tedy nyní mediánovou funkci pro veličinu $Y \sim \mathrm{Po}(\lambda)$. Pravděpodobnostní funkce a distribuční funkce veličiny Y jsou tvaru

$$\mathrm{P}[Y = j] = e^{-\lambda}\frac{\lambda^j}{j!}, \quad j = 0, 1, 2, \ldots, \tag{2.9}$$

$$F_Y(y; \lambda) = e^{-\lambda}\sum_{j=0}^{\lfloor y \rfloor}\frac{\lambda^j}{j!}, \quad y \geq 0, \tag{2.10}$$

kde $\lfloor\cdot\rfloor$ značí dolní celou část. Mediánová funkce $\tilde{m}(\lambda)$ veličiny Y splňuje

$$\tilde{m}(\lambda) = 0 \quad \Leftrightarrow \quad F_Y(0;\lambda) \geq \frac{1}{2}, \tag{2.11}$$

$$\tilde{m}(\lambda) = k, \text{ pro } k \in \mathbb{N} \quad \Leftrightarrow \quad F_Y(k;\lambda) \geq \frac{1}{2} \quad \wedge \quad F_Y(k-1;\lambda) < \frac{1}{2}, \tag{2.12}$$

což je ekvivalentní zápisu

$$\tilde{m}(\lambda) = 0 \quad \Leftrightarrow \quad e^{-\lambda}\sum_{j=0}^{0}\frac{\lambda^j}{j!} = e^{-\lambda} \geq \frac{1}{2}, \tag{2.13}$$

$$\tilde{m}(\lambda) = k, \text{ pro } k \in \mathbb{N} \quad \Leftrightarrow \quad e^{-\lambda}\sum_{j=0}^{k}\frac{\lambda^j}{j!} \geq \frac{1}{2} \quad \wedge \quad e^{-\lambda}\sum_{j=0}^{k-1}\frac{\lambda^j}{j!} < \frac{1}{2}. \tag{2.14}$$

Pro získání krajních hodnot intervalů, ve kterých má mediánová funkce konstantní hodnotu, musíme řešit rovnice

$$e^{-\lambda}\sum_{j=0}^{k}\frac{\lambda^j}{j!} = \frac{1}{2} \tag{2.15}$$

pro $k \in \mathbb{N}_0$. V případě $k = 0$ je řešením rovnice (2.15) $\lambda = \ln 2$, které označíme jako C_0. Pro $k > 0$ nemá výše uvedená rovnice analytické řešení. Řešení této rovnice budeme tedy počítat numericky a zavedeme následující značení.

Definice 1. Kladné řešení rovnice

$$e^{-\lambda}\sum_{j=0}^{k}\frac{\lambda^j}{j!} = \frac{1}{2} \tag{2.16}$$

pro neznámou λ a parametr $k \in \mathbb{N}_0$ označíme jako C_k.

Ukázkovou tabulku hodnot C_k lze najít v příloze.

Díky zavedení hodnot C_k můžeme zapsat mediánovou funkci ve tvaru

$$\tilde{m}(\lambda) = \begin{cases} 0, & \text{je-li } 0 < \lambda \leq C_0, \\ k, & \text{pro } \lambda \in (C_{k-1}, C_k\rangle,\ k \in \mathbb{N}. \end{cases} \tag{2.17}$$

Vidíme, že mediánová funkce je po částech konstantní, tudíž odhad (2.8) nedetekuje malé změny ve velikosti λ a je tedy nevhodný k použití při odhadování parametrů modelu poissonovské regrese.

Využijeme proto takzvané statistické vyhlazování náhodné veličiny, které bylo zavedeno v práci [9]. Jeho princip spočívá v tom, že původně diskrétní veličiny Y_i převedeme na spojité přičtením realizace náhodné veličiny U_i z rovnoměrného rozdělení na $(0,1)$. Vzniknou nám tedy veličiny

$$Z_i = Y_i + U_i, \quad Y_i \sim \mathrm{Po}\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big), \quad U_i \sim \mathcal{U}(0,1), \quad i = 1,2,\ldots,n. \tag{2.18}$$

Náhodné proměnné Y_i a U_h (i U_h mezi sebou) jsou nezávislé pro všechna $i = 1,2,\ldots,n$, $h = 1,2,\ldots,n$. Poznamenejme, že původní data lze zpětně získat aplikací operace dolní celá část na Z_i.

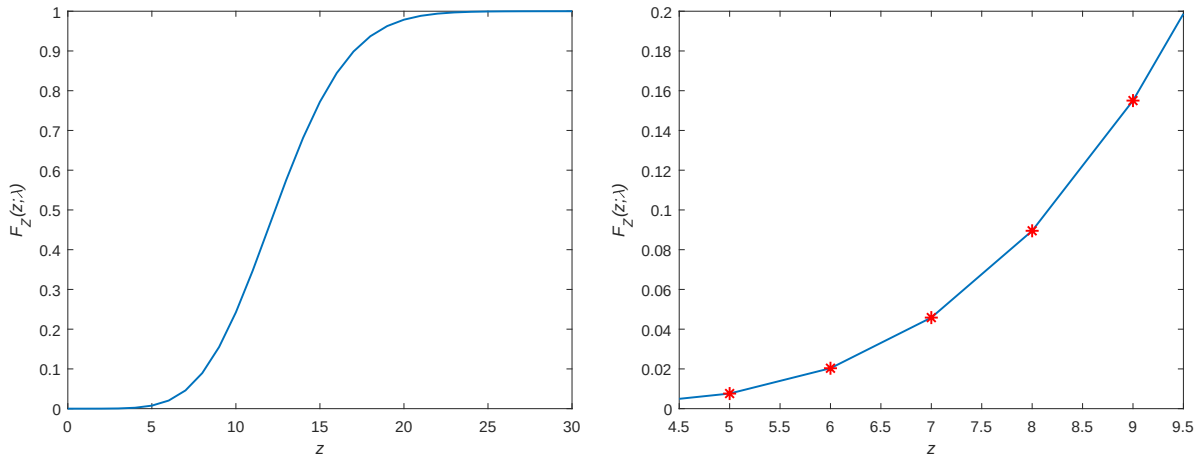
Pro výpočet distribuční funkce veličiny $Z = Y + U$, kde $Y \sim \mathrm{Po}(\lambda)$ a $U \sim \mathcal{U}(0,1)$ jsou nezávislé, nyní předpokládejme, že $z \in \langle k, k+1)$, $k \in \mathbb{N}_0$. Pak platí

$$\begin{aligned} F_Z(z;\lambda) &= \mathrm{P}[Y+U \le z] = \mathrm{P}\left[\left(\bigcup_{j=0}^{k-1}[Y=j]\right) \cup \left([Y=k] \cap [U \le z-k]\right)\right] \\ &= \sum_{j=0}^{k-1} \mathrm{P}[Y=j] + \mathrm{P}[Y=k] \cdot \mathrm{P}[U \le z-k]. \end{aligned} \tag{2.19}$$

U poslední rovnosti jsme využili disjunktnosti jevů a nezávislosti veličin Y a U. Využijeme ještě znalosti rozdělení veličin Y a U a získáme výsledný tvar distribuční funkce

$$F_Z(z;\lambda) = \sum_{j=0}^{k-1} e^{-\lambda}\frac{\lambda^j}{j!} + e^{-\lambda}\frac{\lambda^k}{k!}(z-k), \quad z \in \langle k, k+1), \, k \in \mathbb{N}_0, \tag{2.20}$$

kde v případě $k = 0$ chápeme sumu jako prázdnou, a tedy rovnou nule. Graf této funkce si můžeme prohlédnout na obrázku 2.1. Je to po částech lineární a spojitá funkce.



Obrázek 2.1: Graf distribuční funkce $F_Z(z;\lambda)$ pro $\lambda = 12$. Pravý obrázek je výřez z levého obrázku a jsou na něm patrné zlomy grafu v označených bodech.

Jelikož se nám podařilo získat předpis distribuční funkce, můžeme se přesunout k odvození mediánové funkce veličiny Z. Aby medián ležel v intervalu $\langle k, k+1)$, musí platit

$$F_Z(k;\lambda) \le \frac{1}{2} \quad \wedge \quad F_Z(k+1;\lambda) > \frac{1}{2}. \tag{2.21}$$

Pro $k = 0$ se uplatní pouze druhá z podmínek a dosazením ze vztahu (2.20) získáme podmínku ve tvaru

$$e^{-\lambda}\sum_{j=0}^{0}\frac{\lambda^j}{j!} = e^{-\lambda} > \frac{1}{2}, \tag{2.22}$$

která je ekvivalentní podmínce $\lambda < C_0$. Pro $k > 0$ po dosazení ze vztahu (2.20) dostaneme podmínky ve tvaru

$$e^{-\lambda} \sum_{j=0}^{k-1} \frac{\lambda^j}{j!} \leq \frac{1}{2} \quad \wedge \quad e^{-\lambda} \sum_{j=0}^{k} \frac{\lambda^j}{j!} > \frac{1}{2}, \tag{2.23}$$

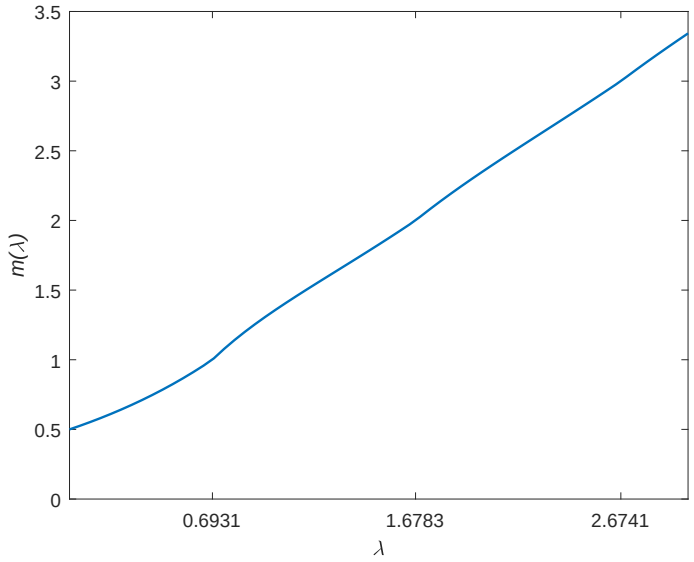
které jsou ekvivalentní podmínce $\lambda \in \langle C_{k-1}, C_k)$. Vyjádřením z z rovnice

$$F_Z(z; \lambda) = \sum_{j=0}^{k-1} e^{-\lambda} \frac{\lambda^j}{j!} + e^{-\lambda} \frac{\lambda^k}{k!}(z - k) = \frac{1}{2} \tag{2.24}$$

získáme předpis pro mediánovou funkci

$$m(\lambda) = \begin{cases} \dfrac{1}{2} e^{\lambda}, & \text{je-li } 0 < \lambda < C_0, \\ k + \dfrac{k!}{\lambda^k} \left(\dfrac{1}{2} e^{\lambda} - \displaystyle\sum_{j=0}^{k-1} \frac{\lambda^j}{j!} \right), & \text{pro } \lambda \in \langle C_{k-1}, C_k), \ k \in \mathbb{N}. \end{cases} \tag{2.25}$$

Mediánová funkce je spojitá a ostře rostoucí. Důkaz tohoto tvrzení lze nalézt v práci [3]. Graf mediánové funkce si můžeme prohlédnout na obrázku 2.2.



Obrázek 2.2: Graf mediánové funkce $m(\lambda)$. Vyznačené hodnoty pro λ jsou konstanty C_0, C_1 a C_2.

Nyní již známe vše potřebné k definici mediánového odhadu pro model poissonovské regrese.

Definice 2. Mějme veličiny $Z_1, Z_2, \ldots, Z_n$ definované v (2.18). Pak mediánový odhad parametrů modelu poissonovské regrese (1.13), (1.16) s $s_i = 1$, $i = 1, 2, \ldots, n$, definujeme předpisem

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| Z_i - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})\big) \right| = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| Y_i + U_i - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})\big) \right|, \tag{2.26}$$

kde $m(\lambda)$ značí mediánovou funkci definovanou v (2.25) a $\lambda(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$ je dána vztahem (2.7).

Z výše uvedené definice je zřejmé, že mediánový odhad patří mezi M-odhady a platí pro něj

$$\Phi_{\mathrm{M}}\big(Y_i, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big) = \big|Y_i + U_i - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big|, \quad i = 1, 2, \ldots, n, \tag{2.27}$$

kde U_i, $i = 1, 2, \ldots, n$, jsou definovány v (2.18).

Velkou nevýhodou mediánového odhadu je potřeba generovat navíc náhodné veličiny U_i, $i = 1, 2, \ldots, n$. Tím dochází k vnášení nežádoucího šumu do odhadování. V práci [9] byly kromě zavedení mediánového odhadu pro modely logistické regrese provedeny simulace, které ukázaly, že tento odhad funguje dobře pro velká n a větší rozsahy znečištění. Simulace ale dále ukázaly, že mediánový odhad pro modely logistické regrese má velký rozptyl odhadů. Aby se odstranily nedostatky týkající se rozptylu odhadů a potřeby generovat náhodné veličiny, pokračovalo se v dalším výzkumu. My budeme tyto kroky následovat a představíme si vylepšení mediánového odhadu.

2.3 Zobecněný mediánový odhad

V této části zavedeme novou metodu odhadu parametrů modelu poissonovské regrese – zobecněný mediánový odhad. K vytvoření tohoto nového odhadu využijeme znalost mediánového odhadu pro poissonovskou regresi a myšlenku, která stála za zrodem zobecněného mediánového odhadu pro odhad parametrů modelu logistické regrese, viz článek [8]. Nejdříve se zvýší počet statistických vyhlazování a následně se provede v jistém smyslu limitní přechod, díky kterému nám suma přejde v integrál. Nebude pak nutné generovat další náhodné veličiny a nebudeme tedy vnášet do odhadování další šum.

V předpisu opět uvažujeme $s_i = 1$, $i = 1, 2, \ldots, n$. Pokud bychom chtěli přejít k obecnému s_i, nahradíme λ_i vztahem $s_i\lambda_i$.

Zavedeme nejdříve k-posílený mediánový odhad pro modely poissonovské regrese předpisem

$$\hat{\boldsymbol{\beta}}_{\mathrm{M},k} = \arg\min_{\boldsymbol{\beta}} \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k} \big|Y_i + U_{ij} - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big|, \tag{2.28}$$

kde k je námi zvolená konstanta a $U_{ij} \sim \mathcal{U}(0,1)$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, jsou vzájemně a na $Y_1, Y_2, \ldots, Y_n$ nezávislé náhodné veličiny. Posílený mediánový odhad tedy vznikl pomocí opakovaného statistického vyhlazování. V práci [9], která se týkala modelů logistické regrese, bylo ukázáno, že pro určité modely se tímto postupem dá dosáhnout výrazného snížení rozptylu odhadu a v některých jednoduchých případech dokonce klesne rozptyl až na úroveň rozptylu maximálně věrohodného odhadu.

Využijeme-li nyní zákon velkých čísel, dostaneme pro $k \to +\infty$

$$\frac{1}{k}\sum_{j=1}^{k} \big|Y_i+U_{ij}-m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big| \xrightarrow{\mathrm{P}} \mathrm{E}_U\big|Y_i+U-m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big| = \int_0^1 \big|Y_i+u-m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big|\,\mathrm{d}u, \tag{2.29}$$

kde $U \sim \mathcal{U}(0,1)$. Pomocí získaného vztahu již můžeme definovat zobecněný mediánový odhad.

Definice 3. Uvažujme model poissonovské regrese (1.13), (1.16) s $s_i = 1$, $i = 1, 2, \ldots, n$. Pak zobecněný mediánový odhad parametrů modelu poissonovské regrese definujeme předpisem

$$\hat{\boldsymbol{\beta}}_{\mathrm{MM}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \int_0^1 \big|Y_i + u - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\big|\,\mathrm{d}u, \tag{2.30}$$

kde $m(\lambda)$ značí mediánovou funkci definovanou v (2.25) a $\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ je dána vztahem (2.7).

Uvažujme nyní náhodnou veličinu $U \sim \mathcal{U}(0,1)$ a definujme funkci $\varphi(t) = \mathrm{E}|U+t| = \int_0^1 |u+t|\,\mathrm{d}u$, pro kterou platí

$$\varphi(t) = \begin{cases} t^2 + t + \frac{1}{2}, & -1 < t \leq 0, \\ t + \frac{1}{2}, & t > 0, \\ -t - \frac{1}{2}, & t \leq -1. \end{cases} \tag{2.31}$$

Pomocí funkce $\varphi(t)$ můžeme přepsat zobecněný mediánový odhad do tvaru

$$\hat{\boldsymbol{\beta}}_{\mathrm{MM}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \varphi\Big(Y_i - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\Big). \tag{2.32}$$

Tento přepis je důležitý pro následnou implementaci zobecněného mediánového odhadu. Díky němu totiž můžeme integrál, který se vyskytuje v definici odhadu, spočíst analyticky a není třeba numerické integrace. Do odhadu tedy nevnášíme další chybu a celkově je odhadování rychlejší.

Předpis pro zobecněný mediánový odhad ještě můžeme přepsat do tvaru

$$\hat{\boldsymbol{\beta}}_{\mathrm{MM}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Phi_{\mathrm{MM}}\big(Y_i, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big), \tag{2.33}$$

kde

$$\Phi_{\mathrm{MM}}\big(Y_i, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big) = \varphi\Big(Y_i - m\big(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\Big), \quad i = 1, 2, \ldots, n. \tag{2.34}$$

Jedná se tedy o M-odhad a bude možné využít obecnou asymptotickou teorii těchto odhadů.

Pokud sčítance v sumě $\sum_{i=1}^{n} \Phi_{\mathrm{MM}}\big(Y_i, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)$ zderivujeme podle $\boldsymbol{\beta}$ a celou sumu položíme rovnu $\mathbf{0}_d$, vzniknou nám takzvané odhadovací rovnice se sčítanci $\boldsymbol{\Psi}\big(Y_i, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)$, $i = 1, 2, \ldots, n$, konkrétně

$$\sum_{i=1}^{n} \boldsymbol{\Psi}\big(Y_i, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big) = \mathbf{0}_d. \tag{2.35}$$

Přesný tvar sčítanců v odhadovacích rovnicích odvodíme v následující kapitole a využijeme ho k odvození asymptotického rozdělení zobecněného mediánového odhadu.

2.4 Mallowsův odhad

Mallowsův odhad je založen na zobecnění metody kvazi-věrohodnostního odhadu. Podrobnosti o kvazi-věrohodnostním odhadu lze nalézt např. v práci [15]. Mallowsův odhad byl poprvé publikován v článku [4] a zabývala se jím také práce [3].

Předtím než si zadefinujeme Mallowsův odhad pro model poissonovské regrese (1.13) a (1.16), zavedeme si označení pro pojmy, které budeme dále potřebovat.

Nejprve si jako $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ označíme matici, která v jednotlivých řádcích obsahuje vektory vysvětlujících proměnných $\boldsymbol{x}_i^{\mathrm{T}}$, kde $i = 1, 2, \ldots, n$. Dále jako h_i označíme i-tý diagonální prvek matice $\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ a definujeme váhy w jako $w(\boldsymbol{x}_i) = \sqrt{1 - h_i}$.

Dále definujeme Huberovu funkci $\psi_c(x)$ pro $x \in \mathbb{R}$ a $c > 0$ předpisem

$$\psi_c(x) = \begin{cases} x, & \text{je-li } |x| \leq c, \\ c\,\mathrm{sgn}(x), & \text{je-li } |x| > c, \end{cases} \tag{2.36}$$

kde sgn značí znaménkovou funkci signum. Hodnota konstanty c v Huberově funkci, která se, jak uvidíme později, vyskytuje v Mallowsově odhadu, má vliv na asymptotickou eficienci odhadu.

Mallowsův odhad pro modely poissonovské regrese definujeme jako řešení soustavy rovnic

$$\boldsymbol{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\psi_c(r_i)\, w(\boldsymbol{x}_i)\, \sqrt{\mu_i}\, \boldsymbol{x}_i - \boldsymbol{a}(\boldsymbol{\beta})] = \mathbf{0}, \tag{2.37}$$

kde

$$r_i = \frac{y_i - \mu_i}{\sqrt{\mu_i}}, \tag{2.38}$$

$$\boldsymbol{a}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[\psi_c(r_i)]\, w(\boldsymbol{x}_i)\, \sqrt{\mu_i}\, \boldsymbol{x}_i. \tag{2.39}$$

Výraz $\mathrm{E}[\psi_c(r_i)]$ lze vyjádřit explicitně ve tvaru

$$\mathrm{E}[\psi_c(r_i)] = c\,\big(\mathrm{P}(Y_i \geq j_2 + 1) - \mathrm{P}(Y_i \leq j_1)\big) + \sqrt{\mu_i}\,\big(\mathrm{P}(Y_i = j_1) - \mathrm{P}(Y_i = j_2)\big), \tag{2.40}$$

kde $j_1 = \lfloor \mu_i - c\sqrt{\mu_i} \rfloor$ a $j_2 = \lfloor \mu_i + c\sqrt{\mu_i} \rfloor$ a $\lfloor \cdot \rfloor$ značí dolní celou část.

Soustavu (2.37) lze vyřešit pomocí Fisher-scoring algoritmu. Řešení v k-té iteraci získáme z předpisu

$$\hat{\boldsymbol{\beta}}_{\mathrm{MAL}}^{(k)} = \hat{\boldsymbol{\beta}}_{\mathrm{MAL}}^{(k-1)} + \boldsymbol{\mathcal{I}}^{-1}\left(\hat{\boldsymbol{\beta}}_{\mathrm{MAL}}^{(k-1)}\right) \boldsymbol{u}\left(\hat{\boldsymbol{\beta}}_{\mathrm{MAL}}^{(k-1)}\right), \tag{2.41}$$

kde pro $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ platí $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{X}$ a $\boldsymbol{B}$ je diagonální matice s prvky

$$b_i = w(\boldsymbol{x}_i)\, \mu_i\, \mathrm{P}(j_1 \leq Y_i \leq j_2 - 1), \quad i = 1, 2, \ldots, n. \tag{2.42}$$

Mallowsův odhad je za určitých podmínek asymptoticky normální. Bližší informace o asymptotické normalitě Mallowsova odhadu lze najít v původním článku [4].

2.5 M-odhad založený na transformaci odezvy

M-odhad založený na transformaci odezvy byl publikován v práci [14]. Představíme si ho zde pro model (1.13) a (1.16), uvažujeme tedy $Y \sim \mathrm{Po}(\mu)$. Jak název napovídá, patří také mezi M-odhady. V názvu se navíc zmiňuje transformace odezvy. Znamená to, že hledáme funkci $t : \mathbb{R} \to \mathbb{R}$ takovou, aby rozptyl transformované veličiny $t(Y)$ byl téměř konstantní v závislosti na $\mathrm{E}(Y)$. Pro modely poissonovské regrese tento požadavek splňuje funkce

$$t(y) = \sqrt{y}. \tag{2.43}$$

Uvažujme dále spojitou a omezenou funkci $\rho : \mathbb{R} \to \mathbb{R}$, která má jediné lokální minimum v nule. V praxi se například využívá funkce

$$\rho(u) = \begin{cases} 1 - \left(1 - \left(\dfrac{u}{2{,}4}\right)^2\right)^4, & \text{je-li } |u| \leq 2{,}4, \\ 1, & \text{je-li } |u| > 2{,}4. \end{cases} \tag{2.44}$$

Definujeme

$$s(\mu) = \arg\min_{u} \mathrm{E}_Y\Big(\rho\big(t(Y) - u\big)\Big). \tag{2.45}$$

O $s(\mu)$ předpokládáme, že je to spojitá a jednoznačně definovaná funkce pro všechna μ.

Vážený M-odhad založený na transformaci odezvy pak definujeme předpisem

$$\hat{\boldsymbol{\beta}}_{\mathrm{WMT}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho\Big(t(Y_i) - s\big(\mu(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)\Big) w\Big(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n\Big), \tag{2.46}$$

kde $\mu(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, $i = 1, 2, \ldots, n$, je dáno v (1.17), $w\Big(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n\Big)$ je funkce Mahalanobisovy vzdálenosti, tedy

$$w\Big(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n\Big) = \omega\Big(\big((\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_n^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)\big)^{\frac{1}{2}}\Big), \tag{2.47}$$

$\hat{\boldsymbol{\mu}}_n$ a $\hat{\boldsymbol{\Sigma}}_n$ jsou robustní odhady (z daných n pozorování) střední hodnoty a kovarianční matice náležící rozdělení vysvětlujících proměnných a ω je nezáporná nerostoucí funkce. Účelem váhové funkce w je penalizovat výrazné pákové body. Jelikož je funkce ρ omezená, je odhad robustní i v případě, že jsou váhy rovny jedné. V takovém případě odhad značíme $\hat{\boldsymbol{\beta}}_{\mathrm{MT}}$. Použití vah může zlepšit i zhoršit odhad, záleží vždy na druhu přítomného znečištění. Jako váhová funkce se například používá funkce

$$\omega(t) = \begin{cases} 1, & \text{je-li } t \leq \chi_{0{,}965;\,5}, \\ \dfrac{\chi_{0{,}975;\,5} - t}{\chi_{0{,}975;\,5} - \chi_{0{,}965;\,5}}, & \text{je-li } \chi_{0{,}965;\,5} < t \leq \chi_{0{,}975;\,5}, \\ 0, & \text{je-li } t > \chi_{0{,}975;\,5}, \end{cases} \tag{2.48}$$

kde $\chi_{\alpha,p}$ značí α-kvantil rozdělení χ^2 s p stupni volnosti.

M-odhad založený na transformaci odezvy je za určitých podmínek asymptoticky normální. Bližší informace o asymptotické normalitě tohoto odhadu lze najít v původním článku [14].

Kapitola 3

Asymptotické rozdělení zobecněného mediánového odhadu

Cílem této kapitoly je odvodit asymptotické rozdělení zobecněného mediánového odhadu, které dále využijeme při testování hypotéz. Nejprve si představíme používané značení a pomocné výpočty. Poté vyslovíme předpoklady, které potřebujeme uvažovat při zkoumání asymptotického rozdělení zobecněného mediánového odhadu, a vyslovíme větu o tomto asymptotickém rozdělení. Na závěr kapitoly se zaměříme na konzistentní odhad teoretické asymptotické kovarianční matice, který potřebujeme znát, abychom mohli používat odvozenou teorii v praxi.

V rámci této kapitoly budeme uvažovat $s_i = 1$, $i = 1, 2, \ldots, n$. Pokud bychom chtěli přejít k obecnému s_i, nahradíme λ_i vztahem $s_i\lambda_i$.

3.1 Značení

Nejprve uveďme použité značení. Pro skalární funkci f a $\boldsymbol{\beta} \in \mathbb{R}^d$ definujeme

$$\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \left(\frac{\partial f}{\partial \beta_1}, \frac{\partial f}{\partial \beta_2}, \ldots, \frac{\partial f}{\partial \beta_d} \right)^{\mathrm{T}} \tag{3.1}$$

a pro vektorovou funkci $\boldsymbol{f} = (f_1, f_2, \ldots, f_g)^{\mathrm{T}}$ definujeme

$$\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}} \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{\beta}) = \left(\frac{\mathrm{D}f_1}{\mathrm{D}\boldsymbol{\beta}}, \frac{\mathrm{D}f_2}{\mathrm{D}\boldsymbol{\beta}}, \ldots, \frac{\mathrm{D}f_g}{\mathrm{D}\boldsymbol{\beta}} \right)^{\mathrm{T}} = \begin{pmatrix} \frac{\partial f_1}{\partial \beta_1} & \frac{\partial f_1}{\partial \beta_2} & \cdots & \frac{\partial f_1}{\partial \beta_d} \\ \frac{\partial f_2}{\partial \beta_1} & \frac{\partial f_2}{\partial \beta_2} & \cdots & \frac{\partial f_2}{\partial \beta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_g}{\partial \beta_1} & \frac{\partial f_g}{\partial \beta_2} & \cdots & \frac{\partial f_g}{\partial \beta_d} \end{pmatrix}. \tag{3.2}$$

3.2 Pomocné výpočty

V této části si představíme pomocné výpočty, které využijeme dále v práci. Nejprve spočteme derivaci mediánové funkce a poté vyjádříme tvar sčítanců $\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ v odhadovacích rovnicích. Jelikož už máme zavedené potřebné značení, definujme korektně funkci $\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$

jako

$$\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)=\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\Phi_{\mathrm{MM}}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big). \tag{3.3}$$

Funkce $\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ je nezbytná pro vyslovení věty o asymptotickém rozdělení zobecněného mediánového odhadu.

Věta 1. Derivace mediánové funkce (2.25) je rovna

$$m'(\lambda)=\frac{\mathrm{d}m(\lambda)}{\mathrm{d}\lambda}=\begin{cases}\dfrac{1}{2}e^{\lambda}, & \lambda\in(0,C_0),\\ m(\lambda)\left(1-\dfrac{k}{\lambda}\right)+\dfrac{k^2}{\lambda}+\dfrac{k}{\lambda}-k, & \lambda\in\langle C_{k-1},C_k),\ k\in\mathbb{N}.\end{cases} \tag{3.4}$$

Důkaz. Hodnota derivace pro variantu $\lambda\in(0,C_0)$ je zřejmá, provedeme tedy výpočet pouze pro případ $\lambda\in\langle C_{k-1},C_k)$, $k\in\mathbb{N}$. Platí

$$\begin{aligned}m'(\lambda)&\overset{(2.25)}{=}\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[k+\frac{k!}{\lambda^k}\left(\frac{1}{2}e^{\lambda}-\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}\right)\right]=-k\frac{k!}{\lambda^{k+1}}\left(\frac{1}{2}e^{\lambda}-\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}\right)+\frac{k!}{\lambda^k}\left(\frac{1}{2}e^{\lambda}-\sum_{j=1}^{k-1}\frac{\lambda^{j-1}}{(j-1)!}\right)\\&=-\frac{k}{\lambda}\left[\frac{k!}{\lambda^k}\left(\frac{1}{2}e^{\lambda}-\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}\right)\right]+\frac{k!}{\lambda^k}\left(\frac{1}{2}e^{\lambda}-\sum_{j=0}^{k-2}\frac{\lambda^j}{j!}\right)\\&=-\frac{k}{\lambda}\left[\frac{k!}{\lambda^k}\left(\frac{1}{2}e^{\lambda}-\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}\right)+k-k\right]+\frac{k!}{\lambda^k}\left(\frac{1}{2}e^{\lambda}-\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}+\frac{\lambda^{k-1}}{(k-1)!}\right)+k-k\\&\overset{(2.25)}{=}-\frac{k}{\lambda}m(\lambda)+\frac{k^2}{\lambda}+m(\lambda)+\frac{k!}{\lambda^k}\frac{\lambda^{k-1}}{(k-1)!}-k=m(\lambda)\left(1-\frac{k}{\lambda}\right)+\frac{k^2}{\lambda}+\frac{k}{\lambda}-k.\end{aligned}$$

□

Pro větší přehlednost v dalším textu označíme

$$\lambda=\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})=\exp\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\}. \tag{3.5}$$

Pro derivaci λ podle $\boldsymbol{\beta}$ pak platí

$$\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\lambda=\lambda\boldsymbol{x}. \tag{3.6}$$

Věta 2. Sčítance v odhadovacích rovnicích

$$\sum_{i=1}^{n}\boldsymbol{\Psi}\big(Y_i,\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)=\mathbf{0}_d \tag{3.7}$$

pro zobecněný mediánový odhad jsou tvaru

$$\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)=\boldsymbol{\Psi}\big(Y,\lambda\big)=\begin{cases}\frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda}-1\big)\boldsymbol{x}, & \lambda\in(0,C_0)\wedge Y=0,\\ -\frac{1}{2}\lambda e^{\lambda}\boldsymbol{x}, & \lambda\in(0,C_0)\wedge Y>0,\\ m'(\lambda)\lambda\Big(2\big(m(\lambda)-k\big)-1\Big)\boldsymbol{x}, & \lambda\in\langle C_{k-1},C_k)\wedge Y=k,\\ -m'(\lambda)\lambda\boldsymbol{x}, & \lambda\in\langle C_{k-1},C_k)\wedge Y>k,\\ m'(\lambda)\lambda\boldsymbol{x}, & \lambda\in\langle C_{k-1},C_k)\wedge Y<k,\end{cases} \tag{3.8}$$

kde $k\in\mathbb{N}$ a funkce $m(\lambda)$ a $m'(\lambda)$ jsou definovány předpisy (2.25) a (3.4).

Důkaz. Nejprve využijeme definici $\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ a vztahu mezi funkcemi Φ_{MM} a φ. Získáme

$$\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) = \boldsymbol{\Psi}\big(Y,\lambda\big) \overset{(3.3)}{=} \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\Phi_{\mathrm{MM}}\big(Y,\lambda\big) \overset{(2.34)}{=} \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\varphi\big(Y-m(\lambda)\big).$$

Pro větší přehlednost dosadíme do definice funkce φ (viz (2.31)) argument $Y-m(\lambda)$. Pak pro $\varphi\big(Y-m(\lambda)\big)$ platí

$$\varphi\big(Y-m(\lambda)\big) = \begin{cases} \big(Y-m(\lambda)\big)^2 + Y - m(\lambda) + \dfrac{1}{2}, & -1 < Y-m(\lambda) \le 0, \\ Y-m(\lambda)+\dfrac{1}{2}, & Y-m(\lambda) > 0, \\ -Y+m(\lambda)-\dfrac{1}{2}, & Y-m(\lambda) \le -1. \end{cases}$$

Pokračujeme dále v úpravách $\boldsymbol{\Psi}\big(Y,\lambda\big)$. Získáme

$$\boldsymbol{\Psi}\big(Y,\lambda\big) \overset{(3.6)}{=} m'(\lambda)\lambda\boldsymbol{x}\begin{cases} 2\big(m(\lambda)-Y\big)-1, & m(\lambda)-1 < Y \le m(\lambda), \\ -1, & Y > m(\lambda), \\ 1, & Y \le m(\lambda)-1, \end{cases}$$

$$= \begin{cases} \frac{1}{2}e^{\lambda}\lambda\Big(2\big(\frac{1}{2}e^{\lambda}-Y\big)-1\Big)\boldsymbol{x}, & \lambda\in(0,C_0)\wedge m(\lambda)-1<Y\le m(\lambda), \\ -\frac{1}{2}e^{\lambda}\lambda\boldsymbol{x}, & \lambda\in(0,C_0)\wedge Y>m(\lambda), \\ \frac{1}{2}e^{\lambda}\lambda\boldsymbol{x}, & \lambda\in(0,C_0)\wedge Y\le m(\lambda)-1, \\ m'(\lambda)\lambda\Big(2\big(m(\lambda)-Y\big)-1\Big)\boldsymbol{x}, & \lambda\in\langle C_{k-1},C_k)\wedge m(\lambda)-1<Y\le m(\lambda), \\ -m'(\lambda)\lambda\boldsymbol{x}, & \lambda\in\langle C_{k-1},C_k)\wedge Y>m(\lambda), \\ m'(\lambda)\lambda\boldsymbol{x}, & \lambda\in\langle C_{k-1},C_k)\wedge Y\le m(\lambda)-1, \end{cases}$$

kde $k\in\mathbb{N}$.

Přejděme nyní k úpravě podmínek. Vzhledem k definici konstant C_k, $k\in\mathbb{N}_0$, viz (2.16), a definici funkce $m(\lambda)$ v (2.25) platí

$$m(0^+)=\frac{1}{2} \quad\wedge\quad m(C_{k-1})=k,\ k\in\mathbb{N}.$$

Navíc je $m(\lambda)$ spojitá a ostře rostoucí. Pro $\lambda\in(0,C_0)$ tedy platí $\frac{1}{2}<m(\lambda)<1$ a jednotlivé podmínky platí v případech

- $m(\lambda)-1<Y\le m(\lambda)\Leftrightarrow Y=0$,
- $Y>m(\lambda)\Leftrightarrow Y=1,2,3,\ldots\Leftrightarrow Y>0$,
- $Y\le m(\lambda)-1$ nelze splnit.

Pro $\lambda\in\langle C_{k-1},C_k)$, $k\in\mathbb{N}$, platí $k\le m(\lambda)<k+1$ a jednotlivé podmínky platí v případech

- $m(\lambda)-1<Y\le m(\lambda)\Leftrightarrow Y=k$,
- $Y>m(\lambda)\Leftrightarrow Y=k+1,k+2,k+3,\ldots\Leftrightarrow Y>k$,
- $Y\le m(\lambda)-1\Leftrightarrow Y=k-1,k-2,k-3,\ldots,0\Leftrightarrow Y<k$.

Pokračujme s úpravami $\boldsymbol{\Psi}\big(Y,\lambda\big)$. Dosadíme dle podmínek za Y odpovídající hodnotu a získáme finální předpis ve tvaru

$$\boldsymbol{\Psi}\big(Y,\lambda\big) = \begin{cases} \frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda}-1\big)\boldsymbol{x}, & \lambda \in (0,C_0) \wedge Y = 0, \\ -\frac{1}{2}\lambda e^{\lambda}\boldsymbol{x}, & \lambda \in (0,C_0) \wedge Y > 0, \\ m'(\lambda)\lambda\Big(2\big(m(\lambda)-k\big)-1\Big)\boldsymbol{x}, & \lambda \in \langle C_{k-1},C_k) \wedge Y = k, \\ -m'(\lambda)\lambda\boldsymbol{x}, & \lambda \in \langle C_{k-1},C_k) \wedge Y > k, \\ m'(\lambda)\lambda\boldsymbol{x}, & \lambda \in \langle C_{k-1},C_k) \wedge Y < k, \end{cases}$$

kde $k \in \mathbb{N}$. □

Pro přehlednost v dalším textu označíme

$$A(\lambda) = m'(\lambda)\lambda, \quad A'(\lambda) = \frac{\mathrm{d}A(\lambda)}{\mathrm{d}\lambda}. \tag{3.9}$$

Sčítance v odhadovacích rovnicích přejdou na tvar

$$\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) = \boldsymbol{\Psi}\big(Y,\lambda\big) = \begin{cases} \frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda}-1\big)\boldsymbol{x}, & \lambda \in (0,C_0) \wedge Y = 0, \\ -\frac{1}{2}\lambda e^{\lambda}\boldsymbol{x}, & \lambda \in (0,C_0) \wedge Y > 0, \\ A(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big)\boldsymbol{x}, & \lambda \in \langle C_{k-1},C_k) \wedge Y = k, \\ -A(\lambda)\boldsymbol{x}, & \lambda \in \langle C_{k-1},C_k) \wedge Y > k, \\ A(\lambda)\boldsymbol{x}, & \lambda \in \langle C_{k-1},C_k) \wedge Y < k, \end{cases} \tag{3.10}$$

kde $k \in \mathbb{N}$.

3.3 Centrální předpoklady

Při zkoumání asymptotického rozdělení zobecněného mediánového odhadu potřebujeme předpokládat, že

1. vektory vysvětlujících proměnných $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ pocházejí z kompaktní množiny $\mathcal{X} \subset \mathbb{R}^d$ a pravděpodobnostní míra
$$\hat{G}_n(B) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(\boldsymbol{x}_i \in B), \tag{3.11}$$
kde I značí indikátorovou funkci a B je borelovská podmnožina $\mathcal{X}$, konverguje v distribuci (s $n \to +\infty$) k pravděpodobnostní míře G na borelovských podmnožinách $\mathcal{X}$,

2. skutečná hodnota $\boldsymbol{\beta}$ označená
$$\boldsymbol{\beta}^0 = \big(\beta_1^0,\beta_2^0,\ldots,\beta_d^0\big)^{\mathrm{T}} \tag{3.12}$$
odpovídá jednoznačnému řešení $\big($pro $\boldsymbol{\beta} \in \mathbb{R}^d\big)$ rovnice
$$\int_{\mathcal{X}} \mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\Big]\,\mathrm{d}G(\boldsymbol{x}) = \mathbf{0}_d, \tag{3.13}$$

3. nejmenší vlastní číslo matice

$$\int_{\mathcal{X}} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \,\mathrm{d}G(\boldsymbol{x}) \tag{3.14}$$

je kladné, a proto pro každé $\boldsymbol{\beta} \neq \boldsymbol{\beta}^0$ platí $G\big(\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^0) \neq 0\big) > 0$.

Uvedené předpoklady uvažujeme jako platné v rámci celé práce a nebudeme je tedy opakovat ve tvrzeních jednotlivých vět.

3.4 Věta o asymptotickém rozdělení

Nyní již máme vše připraveno, abychom vyslovili větu o asymptotickém rozdělení zobecněného mediánového odhadu.

Věta 3. Za předpokladu, že zobecněný mediánový odhad $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ je konzistentním odhadem parametru $\boldsymbol{\beta}$, platí

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{\beta}^0) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\big(\mathbf{0}_d, \boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\big), \tag{3.15}$$

kde matice $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ a $\boldsymbol{Q}(\boldsymbol{\beta}^0)$ jsou tvaru

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}^0) = \int_{\mathcal{X}} \mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0)\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0)\big)\Big]\mathrm{d}G(\boldsymbol{x}), \tag{3.16}$$

$$\boldsymbol{Q}(\boldsymbol{\beta}^0) = \int_{\mathcal{X}} \mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}\right]\mathrm{d}G(\boldsymbol{x}). \tag{3.17}$$

Důkaz. Tvrzení plyne z Věty 10.11 v [10]. □

Díky předchozí větě víme, jaké má zobecněný mediánový odhad asymptotické rozdělení. Pro praktické využití ale potřebujeme znát explicitní vyjádření matic $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ a $\boldsymbol{Q}(\boldsymbol{\beta}^0)$. Provedeme proto další výpočty, abychom získali předpisy pro výpočet obou matic.

Věta 4. Pro střední hodnotu výrazu $\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ platí

$$\begin{aligned}
&\mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\Big] = \mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y, \lambda\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda\big)\Big] \\
&\quad = \begin{cases} \dfrac{1}{4}\lambda^2 e^{2\lambda}\big(e^{\lambda} - 1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda \in (0, C_0), \\ \big[m'(\lambda)\big]^2\lambda^2\left[4e^{-\lambda}\dfrac{\lambda^k}{k!}\big(m(\lambda) - k\big)\big(m(\lambda) - k - 1\big) + 1\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda \in \langle C_{k-1}, C_k), \; k \in \mathbb{N}. \end{cases}
\end{aligned} \tag{3.18}$$

Důkaz. V důkazu budeme uvažovat vyjádření $\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ pomocí (3.8). Nejprve dokážeme variantu pro $\lambda \in (0, C_0)$. Platí

$$\begin{aligned}
&\mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\Big] = \mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y, \lambda\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda\big)\Big] \\
&\quad = \left[\frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda} - 1\big)\right]^2 \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \underbrace{e^{-\lambda}}_{\mathrm{P}[Y=0]} + \left[-\frac{1}{2}\lambda e^{\lambda}\right]^2 \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \underbrace{\big(1 - e^{-\lambda}\big)}_{\mathrm{P}[Y\geq 1]} \\
&\quad = \frac{1}{4}\lambda^2 e^{2\lambda}\Big[\big(e^{2\lambda} - 2e^{\lambda} + 1\big)e^{-\lambda} + 1 - e^{-\lambda}\Big]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} = \frac{1}{4}\lambda^2 e^{2\lambda}\big(e^{\lambda} - 1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.
\end{aligned}$$

Nyní přejdeme k variantě $\lambda \in \langle C_{k-1}, C_k)$, $k \in \mathbb{N}$. V tomto případě platí

$$\begin{aligned}
&\mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\Big] = \mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y,\lambda\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda\big)\Big] \\
&\quad= \Big[m'(\lambda)\lambda\Big(2\big(m(\lambda)-k\big)-1\Big)\Big]^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{e^{-\lambda}\frac{\lambda^k}{k!}}_{\mathrm{P}[Y=k]} + \big[-m'(\lambda)\lambda\big]^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{\sum_{j=k+1}^{+\infty}e^{-\lambda}\frac{\lambda^j}{j!}}_{\mathrm{P}[Y\geq k+1]} \\
&\qquad+ \big[m'(\lambda)\lambda\big]^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{\sum_{j=0}^{k-1}e^{-\lambda}\frac{\lambda^j}{j!}}_{\mathrm{P}[Y\leq k-1]} \\
&\quad= \big[m'(\lambda)\big]^2\lambda^2\left[\left(4\left[\big(m(\lambda)-k\big)^2-m(\lambda)+k\right]+1\right)e^{-\lambda}\frac{\lambda^k}{k!}+1-\sum_{j=0}^{k}e^{-\lambda}\frac{\lambda^j}{j!}+\sum_{j=0}^{k-1}e^{-\lambda}\frac{\lambda^j}{j!}\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \\
&\quad= \big[m'(\lambda)\big]^2\lambda^2\left[4e^{-\lambda}\frac{\lambda^k}{k!}\left[\big(m(\lambda)-k\big)^2-m(\lambda)+k\right]+e^{-\lambda}\frac{\lambda^k}{k!}+1-e^{-\lambda}\frac{\lambda^k}{k!}\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \\
&\quad= \big[m'(\lambda)\big]^2\lambda^2\left[4e^{-\lambda}\frac{\lambda^k}{k!}\big(m(\lambda)-k\big)\big(m(\lambda)-k-1\big)+1\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.
\end{aligned}$$

□

Díky tomuto výpočtu jsme se přiblížili k vyjádření matice $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$. Nyní vyjádříme derivaci funkce $\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ a následně spočteme její střední hodnotu. Přiblížíme se tak k vyjádření matice $\boldsymbol{Q}(\boldsymbol{\beta}^0)$.

Věta 5. Výraz $\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ lze vyjádřit ve tvaru

$$\begin{aligned}
\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) &= \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda\big) \\
&= \begin{cases}
\frac{1}{2}\lambda e^{\lambda}\big(2\lambda e^{\lambda}+e^{\lambda}-\lambda-1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda\in(0,C_0)\wedge Y=0,\\
-\frac{1}{2}\lambda(\lambda+1)e^{\lambda}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda\in(0,C_0)\wedge Y>0,\\
\lambda\left[A'(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big)+2\big[m'(\lambda)\big]^2\lambda\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda\in\langle C_{k-1},C_k)\wedge Y=k,\\
-\lambda A'(\lambda)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda\in\langle C_{k-1},C_k)\wedge Y>k,\\
\lambda A'(\lambda)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda\in\langle C_{k-1},C_k)\wedge Y<k,
\end{cases}
\end{aligned} \tag{3.19}$$

kde $k \in \mathbb{N}$ a $A'(\lambda)$ je dáno vztahem (3.9).

Důkaz. V důkazu využijeme vyjádření $\boldsymbol{\Psi}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ pomocí (3.10). Pro variantu $\lambda \in (0, C_0)$ a $Y = 0$ platí

$$\begin{aligned}
\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) = \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y,\lambda\big) &= \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\left[\frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda}-1\big)\boldsymbol{x}^{\mathrm{T}}\right] \\
&\overset{(3.6)}{=} \left(\frac{1}{2}e^{\lambda}\big(e^{\lambda}-1\big)+\frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda}-1\big)+\frac{1}{2}\lambda e^{\lambda}e^{\lambda}\right)\lambda\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \\
&= \frac{1}{2}\lambda e^{\lambda}\big(e^{\lambda}-1+\lambda e^{\lambda}-\lambda+\lambda e^{\lambda}\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} = \frac{1}{2}\lambda e^{\lambda}\big(2\lambda e^{\lambda}+e^{\lambda}-\lambda-1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.
\end{aligned}$$

Nyní přejdeme k variantě $\lambda \in \langle C_{k-1}, C_k)$ a $Y = k$, $k \in \mathbb{N}$. V tomto případě platí

$$\begin{aligned}\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) = \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda\big) &= \frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\Big[A(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big)\boldsymbol{x}^{\mathrm{T}}\Big] \\ &\overset{(3.6)}{=} \Big[A'(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big) + 2A(\lambda)m'(\lambda)\Big]\lambda\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \\ &\overset{(3.9)}{=} \lambda\Big[A'(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big) + 2\left[m'(\lambda)\right]^2\lambda\Big]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.\end{aligned}$$

Ostatní varianty jsou zřejmé. □

Věta 6. Pro střední hodnotu výrazu $\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)$ platí

$$\begin{aligned}\mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\right] &= \mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda\big)\right] \\ &= \begin{cases} \frac{1}{2}\lambda^2 e^{\lambda}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda \in (0, C_0), \\ 2\left[m'(\lambda)\right]^2\lambda^2 e^{-\lambda}\frac{\lambda^k}{k!}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, & \lambda \in \langle C_{k-1}, C_k),\ k \in \mathbb{N}. \end{cases}\end{aligned} \tag{3.20}$$

Důkaz. Nejprve provedeme výpočet pro variantu $\lambda \in (0, C_0)$. V tomto případě platí

$$\begin{aligned}\mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\right] &= \mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda\big)\right] \\ &= \frac{1}{2}\lambda e^{\lambda}\big(2\lambda e^{\lambda} + e^{\lambda} - \lambda - 1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{e^{-\lambda}}_{\mathrm{P}[Y=0]} - \frac{1}{2}\lambda(\lambda+1)e^{\lambda}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{\big(1-e^{-\lambda}\big)}_{\mathrm{P}[Y\geq 1]} \\ &= \frac{1}{2}\lambda\Big[2\lambda e^{\lambda} + e^{\lambda} - \lambda - 1 - \big(\lambda+1\big)\big(e^{\lambda}-1\big)\Big]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \\ &= \frac{1}{2}\lambda\Big[2\lambda e^{\lambda} + e^{\lambda} - \lambda - 1 - \lambda e^{\lambda} + \lambda - e^{\lambda} + 1\Big]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \\ &= \frac{1}{2}\lambda^2 e^{\lambda}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.\end{aligned}$$

Nyní provedeme druhou část důkazu pro variantu $\lambda \in \langle C_{k-1}, C_k)$, $k \in \mathbb{N}$. Získáme

$$\begin{aligned}\mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\right] &= \mathrm{E}_Y\left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}}\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda\big)\right] \\ &= \lambda\Big[A'(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big) + 2\left[m'(\lambda)\right]^2\lambda\Big]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{e^{-\lambda}\frac{\lambda^k}{k!}}_{\mathrm{P}[Y=k]} - \lambda A'(\lambda)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{\sum_{j=k+1}^{+\infty} e^{-\lambda}\frac{\lambda^j}{j!}}_{\mathrm{P}[Y\geq k+1]} \\ &\quad + \lambda A'(\lambda)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\underbrace{\sum_{j=0}^{k-1} e^{-\lambda}\frac{\lambda^j}{j!}}_{\mathrm{P}[Y\leq k-1]} = \lambda e^{-\lambda}\left[A'(\lambda)\Big(2\big(m(\lambda)-k\big)-1\Big)\frac{\lambda^k}{k!} + 2\left[m'(\lambda)\right]^2\lambda\frac{\lambda^k}{k!}\right. \\ &\quad \left. - A'(\lambda)\left(e^{\lambda} - \sum_{j=0}^{k}\frac{\lambda^j}{j!}\right) + A'(\lambda)\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} = 2\left[m'(\lambda)\right]^2\lambda^2 e^{-\lambda}\frac{\lambda^k}{k!}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\end{aligned}$$

$$+ \lambda e^{-\lambda} A'(\lambda) \left[\underbrace{2\frac{\lambda^k}{k!}\big(m(\lambda)-k\big)}_{\overset{(2.25)}{=} e^{\lambda} - 2\sum_{j=0}^{k-1}\frac{\lambda^j}{j!}} - \frac{\lambda^k}{k!} - e^{\lambda} + \frac{\lambda^k}{k!} + 2\sum_{j=0}^{k-1}\frac{\lambda^j}{j!} \right] \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$$

$$= 2\left[m'(\lambda)\right]^2 \lambda^2 e^{-\lambda}\frac{\lambda^k}{k!}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}. \qquad \square$$

Pro účely dalšího textu zadefinujeme dělení množiny $\mathcal{X}$ v závislosti na $\boldsymbol{\beta}^0$.

Definice 4. Definujeme množiny $\Upsilon_k(\boldsymbol{\beta}^0)$, $k \in \mathbb{N}_0$, pomocí předpisů

$$\Upsilon_0(\boldsymbol{\beta}^0) = \big\{\boldsymbol{x} \in \mathcal{X} : \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0) \in (0, C_0)\big\} = \big\{\boldsymbol{x} \in \mathcal{X} : \exp\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0\} \in (0, C_0)\big\}, \tag{3.21}$$

$$\Upsilon_k(\boldsymbol{\beta}^0) = \big\{\boldsymbol{x} \in \mathcal{X} : \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0) \in \langle C_{k-1}, C_k)\big\} = \big\{\boldsymbol{x} \in \mathcal{X} : \exp\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0\} \in \langle C_{k-1}, C_k)\big\},\ k \in \mathbb{N}. \tag{3.22}$$

Pro větší přehlednost ještě označíme

$$\lambda^0 = \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0) = \exp\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0\}. \tag{3.23}$$

Nyní již můžeme vyslovit větu o explicitním tvaru matic $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ a $\boldsymbol{Q}(\boldsymbol{\beta}^0)$.

Věta 7. Matice $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ a $\boldsymbol{Q}(\boldsymbol{\beta}^0)$ z Věty 3 se dají vyjádřit ve tvaru

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\beta}^0) = &\int_{\Upsilon_0(\boldsymbol{\beta}^0)} \frac{1}{4}\big(\lambda^0\big)^2 e^{2\lambda^0}\big(e^{\lambda^0}-1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}G(\boldsymbol{x}) + \sum_{k=1}^{+\infty}\int_{\Upsilon_k(\boldsymbol{\beta}^0)} \left[m'\big(\lambda^0\big)\right]^2\big(\lambda^0\big)^2 \\ &\left[4e^{-\lambda^0}\frac{\big(\lambda^0\big)^k}{k!}\Big[\Big(m\big(\lambda^0\big)-k\Big)\Big(m\big(\lambda^0\big)-k-1\Big)\Big]+1\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}G(\boldsymbol{x}),\end{aligned} \tag{3.24}$$

$$\boldsymbol{Q}(\boldsymbol{\beta}^0) = \int_{\Upsilon_0(\boldsymbol{\beta}^0)} \frac{1}{2}\big(\lambda^0\big)^2 e^{\lambda^0}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}G(\boldsymbol{x}) + \sum_{k=1}^{+\infty}\int_{\Upsilon_k(\boldsymbol{\beta}^0)} 2\left[m'\big(\lambda^0\big)\right]^2\big(\lambda^0\big)^2 e^{-\lambda^0}\frac{\big(\lambda^0\big)^k}{k!}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}G(\boldsymbol{x}), \tag{3.25}$$

kde pro $\Upsilon_0(\boldsymbol{\beta}^0)$ a $\Upsilon_k(\boldsymbol{\beta}^0)$ platí vztahy (3.21) a (3.22).

Důkaz. V důkazu využijeme definice matic $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ a $\boldsymbol{Q}(\boldsymbol{\beta}^0)$, Věty 4 a 6 a definice $\Upsilon_0(\boldsymbol{\beta}^0)$ a $\Upsilon_k(\boldsymbol{\beta}^0)$ pomocí vztahů (3.21) a (3.22). Pro matici $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ platí

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\beta}^0) &\overset{(3.16)}{=} \int_{\mathcal{X}} \mathrm{E}_Y\Big[\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0)\big)\boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0)\big)\Big]\mathrm{d}G(\boldsymbol{x}) \\ &\overset{(3.18)}{=} \int_{\Upsilon_0(\boldsymbol{\beta}^0)} \frac{1}{4}\big(\lambda^0\big)^2 e^{2\lambda^0}\big(e^{\lambda^0}-1\big)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}G(\boldsymbol{x}) + \sum_{k=1}^{+\infty}\int_{\Upsilon_k(\boldsymbol{\beta}^0)} \left[m'\big(\lambda^0\big)\right]^2\big(\lambda^0\big)^2 \\ &\qquad \left[4e^{-\lambda^0}\frac{\big(\lambda^0\big)^k}{k!}\Big[\Big(m\big(\lambda^0\big)-k\Big)\Big(m\big(\lambda^0\big)-k-1\Big)\Big]+1\right]\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}G(\boldsymbol{x})\end{aligned}$$

a pro matici $\boldsymbol{Q}(\boldsymbol{\beta}^0)$ platí

$$\begin{aligned}\boldsymbol{Q}(\boldsymbol{\beta}^0) &\overset{(3.17)}{=} \int_{\mathcal{X}} \mathrm{E}_Y \left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}} \boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} \right] \mathrm{d}G(\boldsymbol{x}) \\ &\overset{(3.20)}{=} \int_{\Upsilon_0(\boldsymbol{\beta}^0)} \frac{1}{2}\big(\lambda^0\big)^2 e^{\lambda^0} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \,\mathrm{d}G(\boldsymbol{x}) + \sum_{k=1}^{+\infty} \int_{\Upsilon_k(\boldsymbol{\beta}^0)} 2\left[m'\big(\lambda^0\big)\right]^2 \big(\lambda^0\big)^2 e^{-\lambda^0} \frac{\big(\lambda^0\big)^k}{k!} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \,\mathrm{d}G(\boldsymbol{x}).\end{aligned}$$

□

3.5 Konzistentní odhad asymptotické kovarianční matice

V předchozí sekci jsme odvodili asymptotické rozdělení zobecněného mediánového odhadu. V praxi je ale běžné, že neznáme pravděpodobnostní míru G, a tudíž nemůžeme přesně vypočíst teoretickou asymptotickou kovarianční matici. Můžeme však provést konzistentní odhad této matice.

Definujme nejdříve náhodné rozdělení množiny $\hat{n} = \{1, 2, \ldots, n\}$, které vznikne na základě realizací $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ náhodného výběru s pravděpodobnostní mírou G, jako

$$\Lambda_{0,n}(\boldsymbol{\beta}^0) = \big\{i \in \hat{n} : \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0) \in (0, C_0)\big\} = \big\{i \in \hat{n} : \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0\} \in (0, C_0)\big\}, \tag{3.26}$$

$$\Lambda_{k,n}(\boldsymbol{\beta}^0) = \big\{i \in \hat{n} : \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0) \in \langle C_{k-1}, C_k)\big\} = \big\{i \in \hat{n} : \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0\} \in \langle C_{k-1}, C_k)\big\}, \tag{3.27}$$

kde $k \in \mathbb{N}$. Pro větší přehlednost si dále označíme

$$\lambda_i^0 = \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0) = \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0\}, \quad i = 1, 2, \ldots, n. \tag{3.28}$$

Díky slabému zákonu velkých čísel víme, že matice

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}^0) &= \int_{\mathcal{X}} \mathrm{E}_Y \Big[\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0)\big) \boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^0)\big) \Big] \mathrm{d}\hat{G}_n(\boldsymbol{x}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_Y \Big[\boldsymbol{\Psi}\big(Y, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)\big) \boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)\big) \Big] \\ &= \frac{1}{n} \left(\sum_{i \in \Lambda_{0,n}(\boldsymbol{\beta}^0)} \frac{1}{4}\big(\lambda_i^0\big)^2 e^{2\lambda_i^0} \Big(e^{\lambda_i^0} - 1\Big) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} + \sum_{k=1}^{+\infty} \sum_{i \in \Lambda_{k,n}(\boldsymbol{\beta}^0)} \left[m'\big(\lambda_i^0\big)\right]^2 \big(\lambda_i^0\big)^2 \right. \\ &\quad \left. \left[4e^{-\lambda_i^0} \frac{\big(\lambda_i^0\big)^k}{k!} \Big[\Big(m\big(\lambda_i^0\big) - k\Big) \Big(m\big(\lambda_i^0\big) - k - 1\Big) \Big] + 1 \right] \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right)\end{aligned} \tag{3.29}$$

a matice

$$\begin{aligned}\hat{\boldsymbol{Q}}_n(\boldsymbol{\beta}^0) &= \int_{\mathcal{X}} \mathrm{E}_Y \left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}} \boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} \right] \mathrm{d}\hat{G}_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_Y \left[\frac{\mathrm{D}}{\mathrm{D}\boldsymbol{\beta}} \boldsymbol{\Psi}^{\mathrm{T}}\big(Y, \lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} \right] \\ &= \frac{1}{n} \left(\sum_{i \in \Lambda_{0,n}(\boldsymbol{\beta}^0)} \frac{1}{2}\big(\lambda_i^0\big)^2 e^{\lambda_i^0} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} + \sum_{k=1}^{+\infty} \sum_{i \in \Lambda_{k,n}(\boldsymbol{\beta}^0)} 2\left[m'\big(\lambda_i^0\big)\right]^2 \big(\lambda_i^0\big)^2 e^{-\lambda_i^0} \frac{\big(\lambda_i^0\big)^k}{k!} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right)\end{aligned} \tag{3.30}$$

jsou konzistentní odhady $\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ a $\boldsymbol{Q}(\boldsymbol{\beta}^0)$, a tudíž

$$\hat{\boldsymbol{V}}_n(\boldsymbol{\beta}^0) = \hat{\boldsymbol{Q}}_n^{-1}(\boldsymbol{\beta}^0)\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}^0)\hat{\boldsymbol{Q}}_n^{-1}(\boldsymbol{\beta}^0) \tag{3.31}$$

je konzistentní odhad

$$\boldsymbol{V}(\boldsymbol{\beta}^0) = \boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0). \tag{3.32}$$

Jelikož v praxi $\boldsymbol{\beta}^0$ neznáme, můžeme ho nahradit odhadem $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ a matici $\boldsymbol{V}(\boldsymbol{\beta}^0)$ odhadnout pomocí

$$\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}}) = \hat{\boldsymbol{Q}}_n^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\hat{\boldsymbol{Q}}_n^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}}). \tag{3.33}$$

Protože jsme ale nedokázali konzistenci odhadu $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$, nemůžeme teoreticky nic tvrdit o konzistenci odhadu matice $\boldsymbol{V}(\boldsymbol{\beta}^0)$ pomocí (3.33). V simulační části práce, konkrétně v podkapitole 5.2, provedeme experiment, v rámci kterého budeme zkoumat konzistenci zobecněného mediánového odhadu. Výsledky simulace naznačují, že zatím není důvod se domnívat, že by zobecněný mediánový odhad nebyl konzistentním odhadem.

Předpokládejme tedy nyní, že $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ je skutečně konzistentní odhad $\boldsymbol{\beta}^0$. V článku [2] se zabývali podobným problémem v případě modelů logistické regrese a dokázali lemma, které tvrdí, že konzistentní odhad teoretické asymptotické kovarianční matice, do kterého dosadíme konzistentní odhad parametru, zůstává stále konzistentním odhadem teoretické asymptotické kovarianční matice. Lemma bylo vysloveno pouze pro modely logistické regrese. V článku [14] vyslovili a dokázali obecnější tvrzení pro M-odhad založený na transformaci odezvy, se kterým jsme se seznámili v části 2.5. Toto tvrzení tedy platí i pro modely poissonovské regrese. Vzhledem k uvedeným pracím a faktu, že zobecněný mediánový odhad patří také mezi M-odhady, budeme předpokládat, že uvedená tvrzení by šla zobecnit i pro zobecněný mediánový odhad. V takovém případě bychom mohli tvrdit, že (3.33) je konzistentním odhadem teoretické asymptotické kovarianční matice $\boldsymbol{V}(\boldsymbol{\beta}^0)$. Tuto vlastnost potřebujeme využít v následující kapitole o testování hypotéz.

Kapitola 4

Testování hypotéz

Tato kapitola se zabývá testováním hypotéz o parametrech modelu poissonovské regrese. Robustním testům se v literatuře obecně věnuje menší pozornost než robustním odhadům parametrů. V rámci této kapitoly proto definujeme testovací statistiku Waldova typu založenou na zobecněném mediánovém odhadu a pomocí jeho asymptotického rozdělení odvozeného v předchozí kapitole stanovíme asymptotické rozdělení této testovací statistiky. Na závěr ukážeme, jak odvozenou teorii využít v praxi při testování hypotéz.

V rámci modelů poissonovské regrese potřebujeme být schopni testovat hypotézy tvaru

$$H_0 : \boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}^0 = \boldsymbol{m} \quad \text{vs.} \quad H_1 : \boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}^0 \neq \boldsymbol{m}, \tag{4.1}$$

kde $\boldsymbol{K}^{\mathrm{T}} \in \mathbb{R}^{r\times d}$ je matice s hodností r a $\boldsymbol{m} \in \mathbb{R}^r$.

Testy založené na metodě maximální věrohodnosti jsou stejně jako samotná metoda citlivé na přítomnost odlehlých pozorování a pákových bodů, viz např. [7]. Je tedy potřeba využít robustní metodu odhadu a pomocí ní definovat robustní test. My pro tento účel definujeme testovací statistiku Waldova typu založenou na zobecněném mediánovém odhadu.

Připomeňme nejprve, jak je definována Waldova statistika, viz např. [13]. Pro maximálně věrohodný odhad $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ víme, že platí

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} \sim \mathcal{AN}\big(\boldsymbol{\beta}^0, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\beta}^0)\big), \tag{4.2}$$

kde značkou $\mathcal{AN}$ rozumíme asymptotickou normalitu a $\boldsymbol{\mathcal{I}}$ je Fisherova informační matice, která se dá v modelu určeném (1.13) a (1.16) spočíst pomocí vztahu

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}^0) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{X}, \tag{4.3}$$

kde $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ je matice, která v jednotlivých řádcích obsahuje vektory vysvětlujících proměnných $\boldsymbol{x}_i^{\mathrm{T}}$, $i = 1, 2, \ldots, n$, a $\boldsymbol{W}(\boldsymbol{\beta}^0)$ je diagonální matice, jejíž prvky na diagonále jsou v případě poissonovské regrese rovny $\mu_i^{-1}(\boldsymbol{\beta}^0)$, $i = 1, 2, \ldots, n$. Waldova statistika pro test hypotézy (4.1) je pak definována pro metodu maximálně věrohodného odhadu jako

$$(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} - \boldsymbol{m})^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\boldsymbol{K}\big)^{-1}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} - \boldsymbol{m}) \tag{4.4}$$

a víme o ní, že má asymptotické rozdělení $\chi^2(r)$.

Nyní přejděme k definici statistiky Waldova typu založené na zobecněném mediánovém odhadu. Využijeme k tomu myšlenku Waldovy statistiky, kterou dále zobecníme.

Definice 5. Nechť $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ je zobecněný mediánový odhad parametru $\boldsymbol{\beta}$ z n pozorování. Testovací statistiku Waldova typu pro testování hypotézy (4.1) definujeme vztahem

$$W_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}}) = n\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\big)^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{Q}}_n^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\hat{\boldsymbol{Q}}_n^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\big)^{-1}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\big), \tag{4.5}$$

kde $\hat{\boldsymbol{\Sigma}}_n, \hat{\boldsymbol{Q}}_n$ jsou dány vztahy (3.29), (3.30).

Abychom mohli využít testovací statistiku v praxi, potřebujeme znát její asymptotické rozdělení. Proto vyslovíme následující větu.

Věta 8. Předpokládejme, že zobecněný mediánový odhad $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ je konzistentním odhadem parametru $\boldsymbol{\beta}$. Pak za platnosti H_0, viz (4.1), má testovací statistika Waldova typu definovaná vztahem (4.5) asymptotické rozdělení $\chi^2(r)$. Platí tedy

$$W_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}}) \xrightarrow[n\to+\infty]{\mathcal{D}} \chi^2(r). \tag{4.6}$$

Důkaz. Díky Větě 3 víme, že platí

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{\beta}^0) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\big(\boldsymbol{0}_d, \boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\big). \tag{4.7}$$

Matice $\boldsymbol{K}^{\mathrm{T}}$ je konstantní matice a násobení konstantou je spojitá operace, která zachovává konvergenci v distribuci. Tudíž

$$\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}^0) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\big(\boldsymbol{0}_r, \boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{K}\big), \tag{4.8}$$

a proto za platnoti H_0 z (4.1) platí

$$\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\big(\boldsymbol{0}_r, \boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{K}\big). \tag{4.9}$$

Pro asymptotickou kovarianční matici ve vztahu výše platí, že je symetrická a regulární. Tudíž existuje regulární matice $\boldsymbol{C}$, viz např. [1], pro kterou platí

$$\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{K} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}. \tag{4.10}$$

Matice $\boldsymbol{C}$ je konstantní. Pokračujeme v úpravách a získáme postupně vztahy

$$\sqrt{n}\boldsymbol{C}^{-1}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\big(\boldsymbol{0}_r, \boldsymbol{C}^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{K}(\boldsymbol{C}^{-1})^{\mathrm{T}}\big), \tag{4.11}$$

$$\sqrt{n}\boldsymbol{C}^{-1}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\big(\boldsymbol{0}_r, \boldsymbol{I}\big). \tag{4.12}$$

Nyní využijeme definice rozdělení χ^2 a faktu, že skalární součin vektoru se sebou samým je spojitá operace, která zachovává konvergenci v distribuci. Můžeme tedy psát

$$\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m})^{\mathrm{T}}(\boldsymbol{C}^{-1})^{\mathrm{T}}\boldsymbol{C}^{-1}\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}) \xrightarrow[n\to+\infty]{\mathcal{D}} \chi^2(r). \tag{4.13}$$

Upravíme levou stranu výrazu

$$\begin{aligned}
&\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m})^{\mathrm{T}}(\boldsymbol{C}^{-1})^{\mathrm{T}}\boldsymbol{C}^{-1}\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}) \\
&\quad= \sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m})^{\mathrm{T}}(\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}})^{-1}\sqrt{n}(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}) \\
&\quad= \sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\big)^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{K}\big)^{-1}\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\big)
\end{aligned} \tag{4.14}$$

a získáme, že platí

$$\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big)^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\boldsymbol{Q}^{-1}(\boldsymbol{\beta}^0)\boldsymbol{K}\big)^{-1}\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big)\xrightarrow[n\to+\infty]{\mathcal{D}}\chi^2(r). \quad (4.15)$$

Využijeme značení (3.31), (3.32) a (3.33). Již víme, že $\hat{\boldsymbol{V}}_n(\boldsymbol{\beta}^0)$ je konzistentní odhad matice $\boldsymbol{V}(\boldsymbol{\beta}^0)$. Využijeme předpoklad věty o konzistenci odhadu a komentář z konce podkapitoly 3.5 a získáme, že i $\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})$ je konzistentní odhad matice $\boldsymbol{V}(\boldsymbol{\beta}^0)$, neboli

$$\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\xrightarrow[n\to+\infty]{\mathrm{P}}\boldsymbol{V}(\boldsymbol{\beta}^0). \quad (4.16)$$

Z předchozího vztahu plyne

$$\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\xrightarrow[n\to+\infty]{\mathrm{P}}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K}. \quad (4.17)$$

Využijeme znalosti, že invertování matice je spojitá operace, která zachovává konvergenci v pravděpodobnosti, a tedy

$$(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K})^{-1}\xrightarrow[n\to+\infty]{\mathrm{P}}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K})^{-1}. \quad (4.18)$$

Z konvergence v pravděpodobnosti plyne konvergence v distribuci, můžeme tedy psát i

$$(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K})^{-1}\xrightarrow[n\to+\infty]{\mathcal{D}}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K})^{-1}. \quad (4.19)$$

Nyní ukážeme předpoklady Slutského perturbačního teorému, viz např. Věta 6 (b) v knize [6]. Nejprve si označíme

$$X_n=\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big)^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K}\big)^{-1}\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big). \quad (4.20)$$

Z první části důkazu víme, že

$$X_n\xrightarrow[n\to+\infty]{\mathcal{D}}\chi^2(r). \quad (4.21)$$

Dále si označíme

$$W_n=W_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})=\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big)^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\big)^{-1}\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big). \quad (4.22)$$

Potřebujeme ukázat, že $(W_n-X_n)\xrightarrow{\mathrm{P}}0$. Pracujeme tedy s výrazem

$$W_n-X_n=\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big)^{\mathrm{T}}\Big[\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\big)^{-1}-\big(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K}\big)^{-1}\Big]\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big). \quad (4.23)$$

Využijeme označení (3.32) a dosadíme ho do vztahu (4.9). Víme tedy, že platí

$$\sqrt{n}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}-\boldsymbol{m}\big)\xrightarrow[n\to+\infty]{\mathcal{D}}\mathcal{N}(\boldsymbol{0}_r,\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K}). \quad (4.24)$$

Dále využijeme toho, že platí (4.19), a získáme

$$\Big[\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\big)^{-1}-\big(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\beta}^0)\boldsymbol{K}\big)^{-1}\Big]\xrightarrow[n\to+\infty]{\mathcal{D}}\boldsymbol{0}_{r\times r}. \quad (4.25)$$

Nyní použijeme Slutského větu, viz např. Důsledek v 6. kapitole knihy [6], a získáme výsledek

$$(W_n - X_n) \xrightarrow[n\to+\infty]{\mathcal{D}} 0. \tag{4.26}$$

Jelikož 0 je konstanta, platí i

$$(W_n - X_n) \xrightarrow[n\to+\infty]{\mathrm{P}} 0. \tag{4.27}$$

Ukázali jsme (4.21) a (4.27), využijeme tedy Slutského perturbační teorém a získáme finální tvrzení

$$W_n \xrightarrow[n\to+\infty]{\mathcal{D}} \chi^2(r), \tag{4.28}$$

neboli

$$n\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\big)^{\mathrm{T}}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{Q}}_n^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\hat{\boldsymbol{Q}}_n^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\big)^{-1}\big(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\big) \xrightarrow[n\to+\infty]{\mathcal{D}} \chi^2(r), \tag{4.29}$$

což je tvrzení věty. □

Ukažme si, jak vyslovenou větu využít v praxi. Můžeme chtít například otestovat hypotézu

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_d = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0 \text{ pro alespoň jeden index } i \in \{2, 3, \ldots, d\}. \tag{4.30}$$

V takovém případě je matice $\boldsymbol{K}^{\mathrm{T}}$ rozměru $(d-1) \times d$ a tvaru $\boldsymbol{K}^{\mathrm{T}} = (\boldsymbol{0}_{d-1}, \boldsymbol{I}_{d-1})$, kde $\boldsymbol{I}_{d-1}$ je jednotková matice rozměru $(d-1)\times(d-1)$, a pro vektor $\boldsymbol{m}$ platí $\boldsymbol{m} = \boldsymbol{0}_{d-1}$. Testovací statistika Waldova typu má pak asymptotické rozdělení $\chi^2(d-1)$.

V jiné situaci můžeme chtít například pro libovolné $i \in \{2, 3, \ldots, d\}$ otestovat hypotézu

$$H_0 : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0. \tag{4.31}$$

Nyní má $\boldsymbol{K}^{\mathrm{T}}$ podobu řádkového vektoru, jehož všechny prvky kromě i-tého jsou rovny nule a i-tý prvek je roven jedné. Vektor $\boldsymbol{m}$ má pouze jednu složku rovnou nule. Asymptotické rozdělení testovací statistiky Waldova typu je v tomto případě $\chi^2(1)$.

Kapitola 5

Simulační experimenty

Tato kapitola je věnována simulačním experimentům. Budeme v rámci ní zkoumat, jak fungují odvozené teoretické vlastnosti v praxi. Nejprve si představíme model, se kterým budeme pracovat. Poté budeme zkoumat konzistenci zobecněného mediánového odhadu. V další části se už přesuneme k testování hypotéz. Zde budeme zkoumat chybu 1. druhu a sílu testu v několika situacích. Budeme postupně pracovat s čistými daty a znečištěnými daty. Konkrétně budeme uvažovat znečištění odlehlými pozorováními nebo pákovými body. V části týkající se testování hypotéz budeme pro porovnání uvažovat kromě zobecněného mediánového odhadu (MMed) také maximálně věrohodný odhad (MLE) a dva zástupce již existujících robustních odhadů, konkrétně Mallowsův odhad (Mal) a M-odhad založený na transformaci odezvy (MT). Zkratky uvedené v závorkách využijeme v legendách grafů.

Veškeré simulační experimenty byly implementovány v jazyce R. Pro hledání maximálně věrohodného odhadu jsme použili funkci `glm()`. Implementace zobecněného mediánového odhadu vyžaduje použití funkce na hledání argumentu minima. My jsme využili funkci `fminsearch()` z balíčku neldermead, která je založena na simplexové metodě. Jako počáteční odhad jsme volili maximálně věrohodný odhad. Pro zbylé dva odhady jsme využili funkci `glmrob()` z balíčku robustbase. Pro Mallowsův odhad jsme nastavili `method = "Mqle"` a pro M-odhad založený na transformaci odezvy jsme nastavili `method = "MT"`.

5.1 Model

Uvažujeme model ve tvaru

$$\ln \lambda_i^0 = \beta_1^0 x_{i1} + \beta_2^0 x_{i2} + \beta_3^0 x_{i3} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^0, \quad i = 1, 2, \ldots, n, \tag{5.1}$$

kde pro vektory vysvětlujících proměnných platí

$$x_{i1} = 1, \quad x_{i2} \sim \mathcal{N}(0; 1), \quad x_{i2} \text{ nezávislé}, \quad x_{i3} = \frac{i-1}{n-1}, \quad i = 1, 2, \ldots, n. \tag{5.2}$$

Hodnoty x_{i3}, $i = 1, 2, \ldots, n$, jsou ekvidistantně rozdělené hodnoty z intervalu $\langle 0; 1 \rangle$. Třetí složka vektoru vysvětlujících proměnných nám tedy reprezentuje lineární trend.

Vektor regresních koeficientů má hodnotu

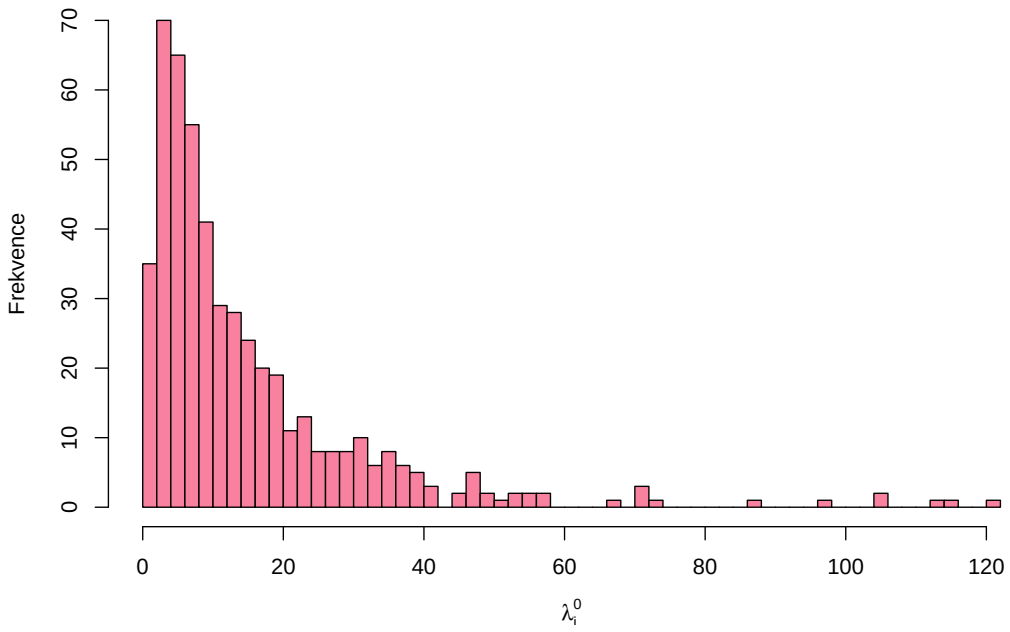
$$\boldsymbol{\beta}^0 = (2{,}5;\ 1;\ -0{,}5) \tag{5.3}$$

a pro vysvětlované proměnné, které jsou mezi sebou nezávislé, platí

$$Y_i \sim \mathrm{Po}(\lambda_i^0), \quad i = 1, 2, \ldots, n, \tag{5.4}$$

$$\lambda_i^0 = \exp\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^0\}, \quad i = 1, 2, \ldots, n. \tag{5.5}$$

Vektory regresorů generujeme pouze jednou, a tedy i hodnoty λ_i^0 jsou pevné. Představu o jejich rozsahu můžeme získat z obrázku 5.1. Při návrhu modelu jsme se inspirovali prací [3].



Obrázek 5.1: Histogram hodnot λ_i^0, $i = 1, 2, \ldots, n$, pro $n = 500$.

5.2 Konzistence zobecněného mediánového odhadu

V teoretické části práce jsme vybudovali teorii, která uvažovala předpoklad, že zobecněný mediánový odhad je konzistentní odhad. Konzistence odhadu nebyla teoreticky dokázána. V této části simulací ukážeme, že zatím není důvod se domnívat, že by zobecněný mediánový odhad nebyl konzistentním odhadem. Uvědomujeme si ale, že simulační experiment nikdy nemůže nahradit teoretický důkaz.

Budeme tedy zkoumat, zda se zobecněný mediánový odhad zpřesňuje se zvyšujícím se počtem pozorování.

Algoritmus experimentu

Experiment provedeme postupně pro počet pozorování $n \in \{50, 100, 150, 200, 250, 300, 400, 500, 650, 800, 1000\}$. Algoritmus experimentu lze popsat následujícími kroky.

1. Vygenerujeme vysvětlující veličiny $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, podle (5.2).
2. Vypočteme λ_i^0, $i = 1, 2, \ldots, n$, podle vztahu (5.5).
3. Pro $j = 1, \ldots, N$, kde $N = 10\,000$, opakujeme:
   (a) Vygenerujeme nezávislé vysvětlované proměnné $Y_i \sim \mathrm{Po}(\lambda_i^0)$, $i = 1, 2, \ldots, n$.
   (b) Odhadneme parametr $\hat{\boldsymbol{\beta}}^{(j)}$ pomocí zobecněného mediánového odhadu.

(c) Vypočteme průměrnou relativní chybu na složku v procentech

$$Z_j = \frac{1}{3}\left(\frac{|\hat{\beta}_1^{(j)} - \beta_1^0|}{|\beta_1^0|} + \frac{|\hat{\beta}_2^{(j)} - \beta_2^0|}{|\beta_2^0|} + \frac{|\hat{\beta}_3^{(j)} - \beta_3^0|}{|\beta_3^0|}\right) \cdot 100. \tag{5.6}$$

4. Vypočteme průměrnou relativní chybu odhadu podle vzorce

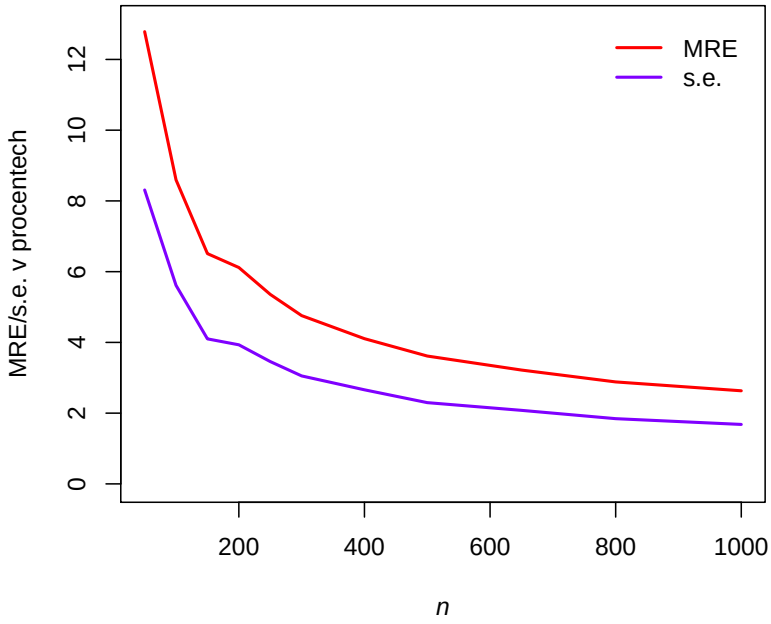
$$\text{MRE} = \bar{Z}_N = \frac{1}{N}\sum_{j=1}^{N} Z_j \tag{5.7}$$

a výběrovou směrodatnou odchylku chyb Z_j podle vzorce

$$\text{s.e.} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(Z_j - \bar{Z}_N)^2}. \tag{5.8}$$

Výsledky experimentu

Na obrázku 5.2 si můžeme prohlédnout výsledek simulačního experimentu. Vidíme, že se zvyšujícím se počtem pozorování se snižují obě sledované veličiny MRE a s.e. To jinými slovy znamená, že se zvyšujícím se počtem pozorování se odhad zpřesňuje a zároveň má menší rozptyl. Tento výsledek je typický pro konzistentní odhady a námi uvažovaný experiment tedy nevylučuje možnost, že je zobecněný mediánový odhad konzistentním odhadem. Obdobné výsledky jsme obdrželi i z dalších simulací týkajících se konzistence odhadu, které jsme provedli v rámci výzkumného projektu. Výsledky těchto dalších simulací zde nebudeme uvádět, protože nepřináší odlišné závěry od závěrů zde uvedených.



Obrázek 5.2: Graf znázorňující průběh veličin MRE a s.e. pro zobecněný mediánový odhad v závislosti na počtu pozorování n.

5.3 Algoritmus experimentů na testování hypotéz

V předchozí části jsme zkoumali konzistenci zobecněného mediánového odhadu. Následující simulace se už budou týkat testování hypotéz a budeme uvažovat kromě zobecněného mediánového odhadu také maximálně věrohodný odhad (MLE), Mallowsův odhad a M-odhad založený na transformaci odezvy.

V teoretické části práce jsme navrhli testovací statistiku pro zobecněný mediánový odhad, takzvanou statistiku Waldova typu, a určili jsme její asymptotické rozdělení. Nyní potřebujeme být schopni testovat hypotézy i pomocí zbylých uvažovaných metod odhadu. U MLE využijeme Waldovu testovací statistiku a u zbylých dvou odhadů využijeme, stejně jako u zobecněného mediánového odhadu, statistiku Waldova typu s příslušnou kovarianční maticí. Odhad této matice nám vracejí příslušné funkce v R.

Uveďme nyní obecný algoritmus experimentů týkajících se testování hypotéz. Experiment provedeme pro počet pozorování n, úroveň znečištění ε a vektor regresních koeficientů $\boldsymbol{\beta}$. Algoritmus experimentu lze popsat následujícími kroky.

1. Vygenerujeme vysvětlující veličiny $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, podle (5.2).
2. Vypočteme parametry λ_i, $i = 1, 2, \ldots, n$, pomocí vzorce $\lambda_i = \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}$.
3. V případě znečištění pákovými body nahradíme vysvětlující veličiny vysvětlujícími veličinami s pákovými body.
4. Pro $j = 1, \ldots, N$, opakujeme:
   (a) Vygenerujeme nezávislé vysvětlované proměnné Y_i, $i = 1, 2, \ldots, n$.
   (b) Odhadneme parametr $\hat{\boldsymbol{\beta}}^{(j)}$ pomocí MLE, zobecněného mediánového odhadu (MMed), Mallowsova odhadu (Mal) a M-odhadu založeného na transformaci odezvy (MT).
   (c) Pro všechny metody odhadu využijeme příslušnou testovací statistiku a otestujeme danou nulovou hypotézu. Podle výsledku nastavíme indikátor zamítnutí jako
   $$\delta^{(j)} = \begin{cases} 1, & \text{pokud } H_0 \text{ zamítneme,} \\ 0, & \text{pokud } H_0 \text{ nezamítneme.} \end{cases} \tag{5.9}$$
5. Pro všechny metody odhadu spočteme procento zamítnutí nulové hypotézy jako
   $$\delta = \frac{1}{N}\sum_{j=1}^{N} \delta^{(j)} \cdot 100. \tag{5.10}$$

5.4 Chyba 1. druhu v závislosti na počtu pozorování

V této části budeme zkoumat vývoj chyby 1. druhu v závislosti na počtu pozorování. Hladinu významnosti testu α jsme v kódech nastavili na hodnotu 0,05, pravděpodobnost chyby 1. druhu by se tedy měla pohybovat kolem této hodnoty. Pro získání dostatečné přesnosti odhadu pravděpodobnosti chyby 1. druhu zvolíme počet opakování $N = 10\,000$.

V tomto experimentu budeme uvažovat pouze čistá data a žádné vychýlení od původního modelu. Platí tedy $\varepsilon = 0$, $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ a vysvětlované proměnné generujeme podle vztahu (5.4).

Jelikož nás zajímá závislost chyby 1. druhu na počtu pozorování, provedeme experiment postupně pro hodnoty $n \in \{50, 100, 150, 200, 250, 300, 400, 500, 650, 800, 1000\}$.

Budeme testovat dvě nulové hypotézy, jedna se týká druhé složky vektoru regresních koeficientů, druhá se týká třetí složky vektoru regresních koeficientů. Konkrétně budeme testovat

$$H_0 : \beta_2^0 = 1 \quad \text{vs.} \quad H_1 : \beta_2^0 \neq 1, \tag{5.11}$$

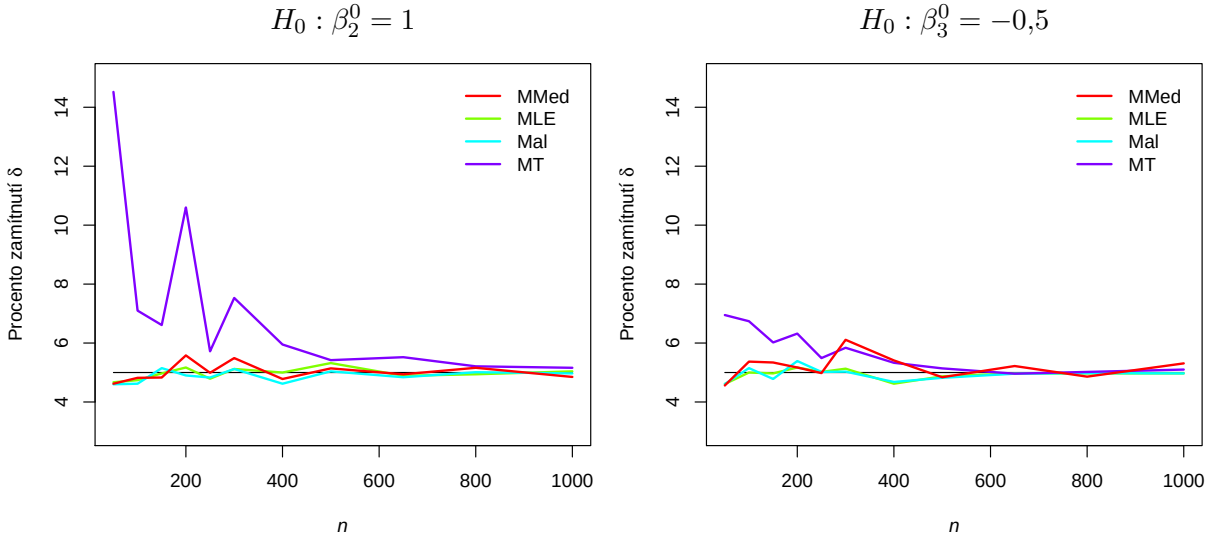
$$H_0 : \beta_3^0 = -0{,}5 \quad \text{vs.} \quad H_1 : \beta_3^0 \neq -0{,}5. \tag{5.12}$$

Obě nulové hypotézy jsou pravdivé, a tedy sledovaná veličina δ nám udává odhad pravděpodobnosti chyby 1. druhu.

Výsledky experimentu

Na obrázku 5.3 si můžeme prohlédnout výsledek experimentu. Nejlépe se 5% chyby drží maximálně věrohodný odhad a Mallowsův odhad. Oba dva odhady mají velmi podobné grafy a pro obě testované hypotézy se chovají stejně stabilně. Zobecněný mediánový odhad v případě testu hypotézy $H_0 : \beta_2^0 = 1$ vykazuje také velmi stabilní chování, v případě testu hypotézy $H_0 : \beta_3^0 = -0{,}5$ se výrazněji odchýlí od 5% hranice pouze v případě $n = 300$.

Největší odchýlení od 5% hranice pozorujeme u M-odhadu založeného na transformaci odezvy. V případě testu hypotézy $H_0 : \beta_2^0 = 1$ vykazuje nestabilní chování pro počet pozorování menší než 500. Není zde ani patrný stabilní klesající trend, na obrázku si můžeme všimnout výrazných extrémů. Nejhoršího odhadu pravděpodobnosti chyby 1. druhu dosahuje v případě $n = 50$, kdy je odhad pravděpodobnosti chyby 1. druhu nad 14 %, a v případě $n = 200$, kdy je odhad pravděpodobnosti chyby 1. druhu mezi 10 a 11 %. Pokud se zaměříme na test hypotézy $H_0 : \beta_3^0 = -0{,}5$, tak zde dosahuje M-odhad založený na transformaci odezvy výrazně lepších výsledků. Nejvyšší odhadnutá pravděpodobnost chyby 1. druhu má hodnotu kolem 7 % pro případ $n = 50$. Se zvyšujícím se počtem pozorování je i patrný klesající trend a kolem $n = 500$ už se odhad drží 5% hranice.



Obrázek 5.3: Odhad pravděpodobnosti chyby 1. druhu pro test hypotéz $H_0 : \beta_2^0 = 1$ (levý obrázek) a $H_0 : \beta_3^0 = -0{,}5$ (pravý obrázek) v závislosti na počtu pozorování n. Černá čára značí 5% hranici.

Výsledek simulace také ukazuje, že náš odhad asymptotické kovarianční matice, který jsme použili při výpočtu testovací statistiky, funguje velmi dobře a není důvod se domnívat, že by nebyl konzistentní.

5.5 Testování hypotéz pro čistá data

Počínaje tímto experimentem budeme uvažovat vychýlení od původního modelu. Konkrétně budeme hýbat s vektorem regresních koeficientů podle vztahu

$$\boldsymbol{\beta} = \boldsymbol{\beta}^0 + \Delta \boldsymbol{c}, \tag{5.13}$$

kde $\Delta \in \langle -0{,}3; 0{,}3 \rangle$ a pro $\boldsymbol{c}$ uvažujeme tři varianty. První varianta je, že hýbeme pouze s druhou složkou regresního koeficientu, a tedy $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$. Druhá varianta je, že hýbeme pouze se třetí složkou regresního koeficientu, a tedy $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$. Poslední varianta je, že hýbeme naráz s druhou i třetí složkou regresního koeficientu. Pro $\boldsymbol{c}$ pak platí $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$.

Budeme testovat stejné hypotézy jako v předchozí části, ale nyní pro vychýlený model. Zajímají nás tedy hypotézy

$$H_0 : \beta_2 = \beta_2^0 = 1 \quad \text{vs.} \quad H_1 : \beta_2 \neq \beta_2^0 = 1, \tag{5.14}$$

$$H_0 : \beta_3 = \beta_3^0 = -0{,}5 \quad \text{vs.} \quad H_1 : \beta_3 \neq \beta_3^0 = -0{,}5. \tag{5.15}$$

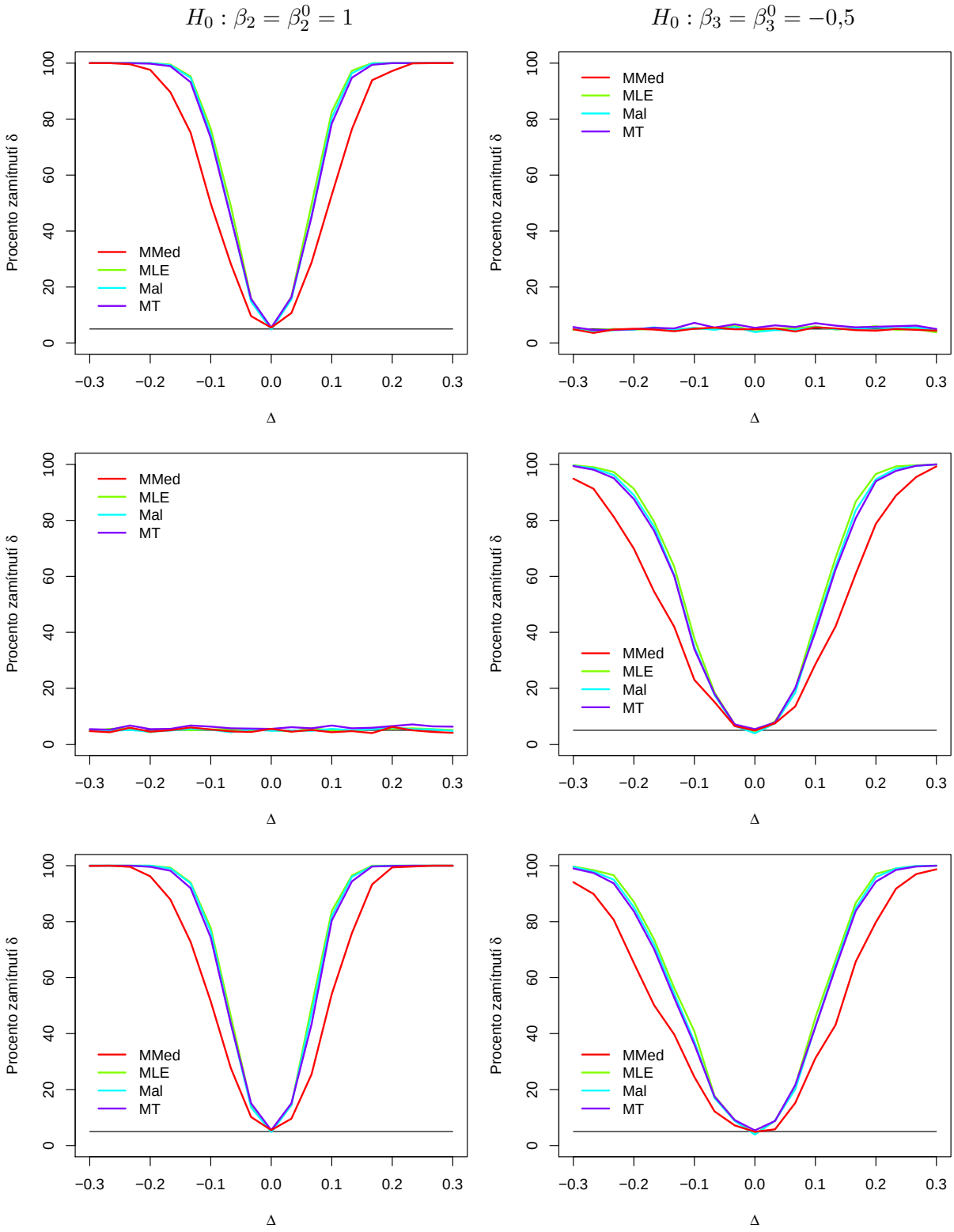
Pokud nehýbeme s danou složkou vektoru regresních koeficientů, je testovaná hypotéza pravdivá a my zkoumáme chybu 1. druhu. Pokud s danou složkou vektoru regresních koeficientů hýbeme, tak nulová hypotéza přestává být pravdivá a my zkoumáme sílu testu. Při návrhu vychýlení od modelu jsme se inspirovali prací [8].

Experiment provedeme pro všechny varianty posunu $\boldsymbol{c}$ a různé hodnoty Δ. Vektor $\boldsymbol{\beta}$ nám určuje vztah (5.13). Uvažujeme pouze čistá data, platí tedy $\varepsilon = 0$ a $Y_i \sim \mathrm{Po}(\lambda_i)$, $\lambda_i = \exp\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\}$, $i = 1, 2, \ldots, n$. Experiment provedeme pro dvě varianty počtu pozorování, konkrétně $n \in \{250, 500\}$. Počet opakování nastavíme jako $N = 1000$. Algoritmus experimentu je podrobně popsán v části 5.3.

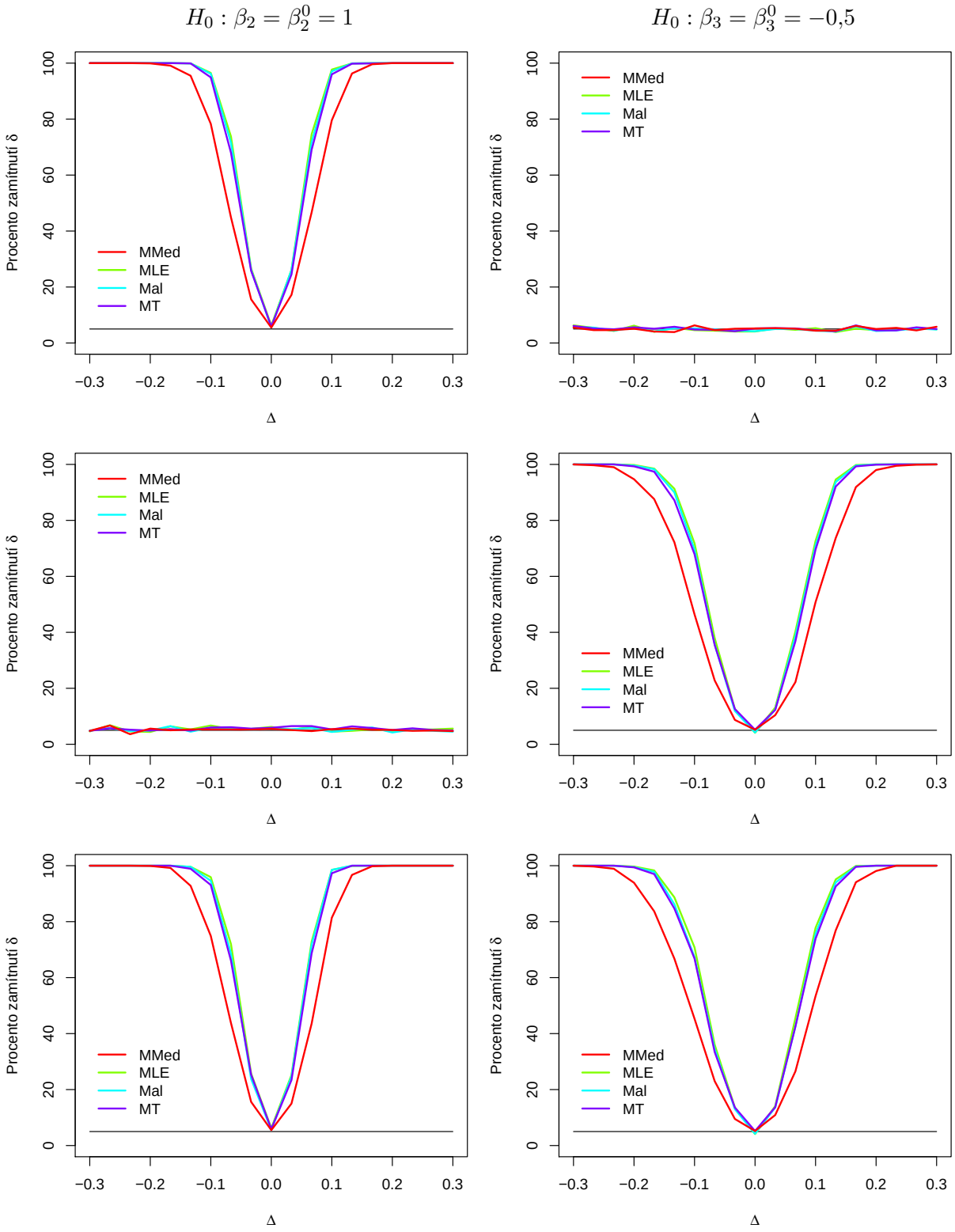
Výsledky experimentu

Výsledky experimentu si můžeme prohlédnout na obrázcích 5.4 a 5.5. Pokud testujeme hypotézu o parametru, se kterým nehýbeme (prostřední graf v levém sloupci a horní graf v pravém sloupci), tak se všechny uvažované metody odhadu drží kolem hranice zamítání 5 %. Odhadnutá pravděpodobnost chyby 1. druhu tedy odpovídá nastavené hladině významnosti testu v kódech a všechny metody odhadu se chovají dle očekávání.

Přesuňme se ke grafům, které znázorňují sílu testu. U všech metod odhadu má křivka znázorňující sílu testu tvar "U", což je očekávaný tvar. Vrchol je v nule, což odpovídá odhadu pravděpodobnosti chyby 1. druhu, protože v nule jsou testované hypotézy pravdivé. Čím více se vzdalujeme od původní hodnoty parametru, tím jsou si metody jistější, že testovaná hypotéza není pravdivá, a zamítají ji ve větším procentu případů až se postupně dostanou k zamítání ve 100 % případů. Vidíme, že tři metody odhadu vykazují téměř totožné chování. Jedná se o maximálně věrohodný odhad, Mallowsův odhad a M-odhad založený na transformaci odezvy. Zobecněný mediánový odhad má širší křivku než zmiňované odhady, což znamená, že má menší sílu testu a pro zamítání nulové hypotézy ve 100 % případů potřebuje větší vychýlení od původní hodnoty parametru.



Obrázek 5.4: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 250$ a čistá data. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.5: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 500$ a čistá data. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.

U testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ je šířka křivky znázorňující sílu testu menší než v případě testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$. Za zmínku také stojí, že grafy znázorňující sílu testu vypadají pro dané testy hypotéz stejně nezávisle na tom, zda se hýbalo pouze s testovanou složkou parametru nebo se hýbalo s oběma složkami parametru.

Pokud se zaměříme na vliv počtu pozorování na sílu testu, je vidět, že křivky jsou užší pro případ $n = 500$ než pro případ $n = 250$. To znamená, že pokud má test k dispozici více dat, je si jistější v zamítání nulové hypotézy a vykazuje tedy větší sílu testu.

5.6 Vliv odlehlých pozorování

V této části budeme zkoumat vliv přítomnosti odlehlých pozorování na testování hypotéz. Nastavení experimentu je stejné jako v části 5.5 s tím rozdílem, že znečistíme vysvětlované proměnné. Vektor $\boldsymbol{\beta}$ nám tedy určuje vztah (5.13) a pro λ_i platí $\lambda_i = \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}$, $i = 1, 2, \ldots, n$.

Znečištění probíhá následovně. Vždy když generujeme vysvětlované proměnné, tak je nejdříve nagenerujeme pomocí vztahu $Y_i \sim \mathrm{Po}(\lambda_i)$, $i = 1, 2, \ldots, n$. Poté jednoduchým náhodným výběrem bez opakování zvolíme $\lfloor \varepsilon n \rfloor$ vysvětlovaných proměnných a ty modifikujeme pomocí vztahu

$$\tilde{Y} = 2(Y + 1). \tag{5.16}$$

Díky tomuto předpisu máme jistotu, že hodnotu zvolené vysvětlované proměnné opravdu změníme. Dále pak počítáme již s takto upravenými vysvětlovanými proměnnými. Hodnoty vysvětlujících veličin neměníme.

V rámci experimentu budeme uvažovat dvě úrovně znečištění. Pro ε tedy platí $\varepsilon \in \{0{,}05; 0{,}1\}$.

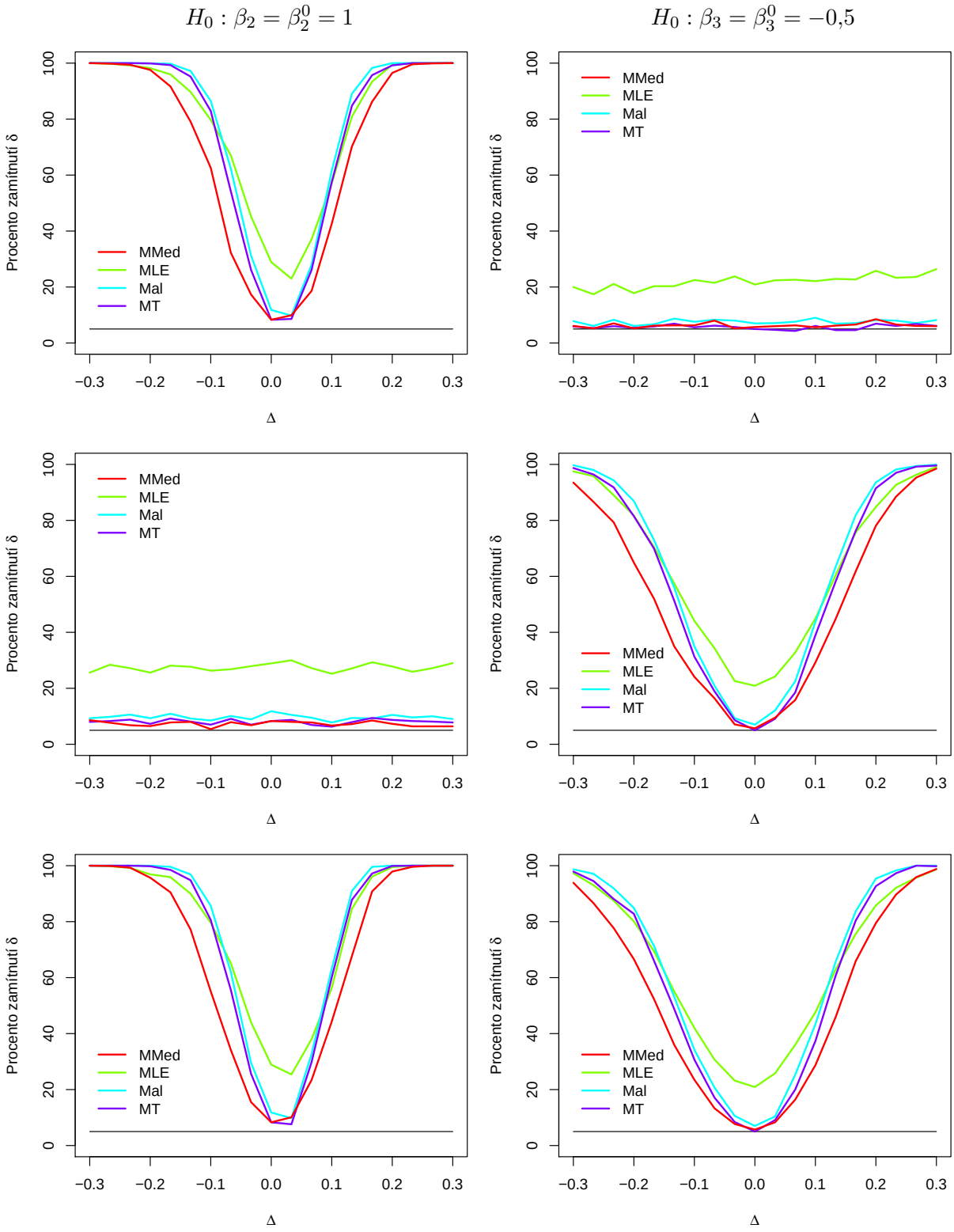
Výsledky experimentu

Zaměřme se nejdříve na výsledky pro variantu $n = 250$, které si můžeme prohlédnout na obrázku 5.6 pro úroveň znečištění $\varepsilon = 0{,}05$ a na obrázku 5.7 pro úroveň znečištění $\varepsilon = 0{,}1$. V obou případech je patrné, že nejvíce citlivý na přítomnost znečištění je maximálně věrohodný odhad.

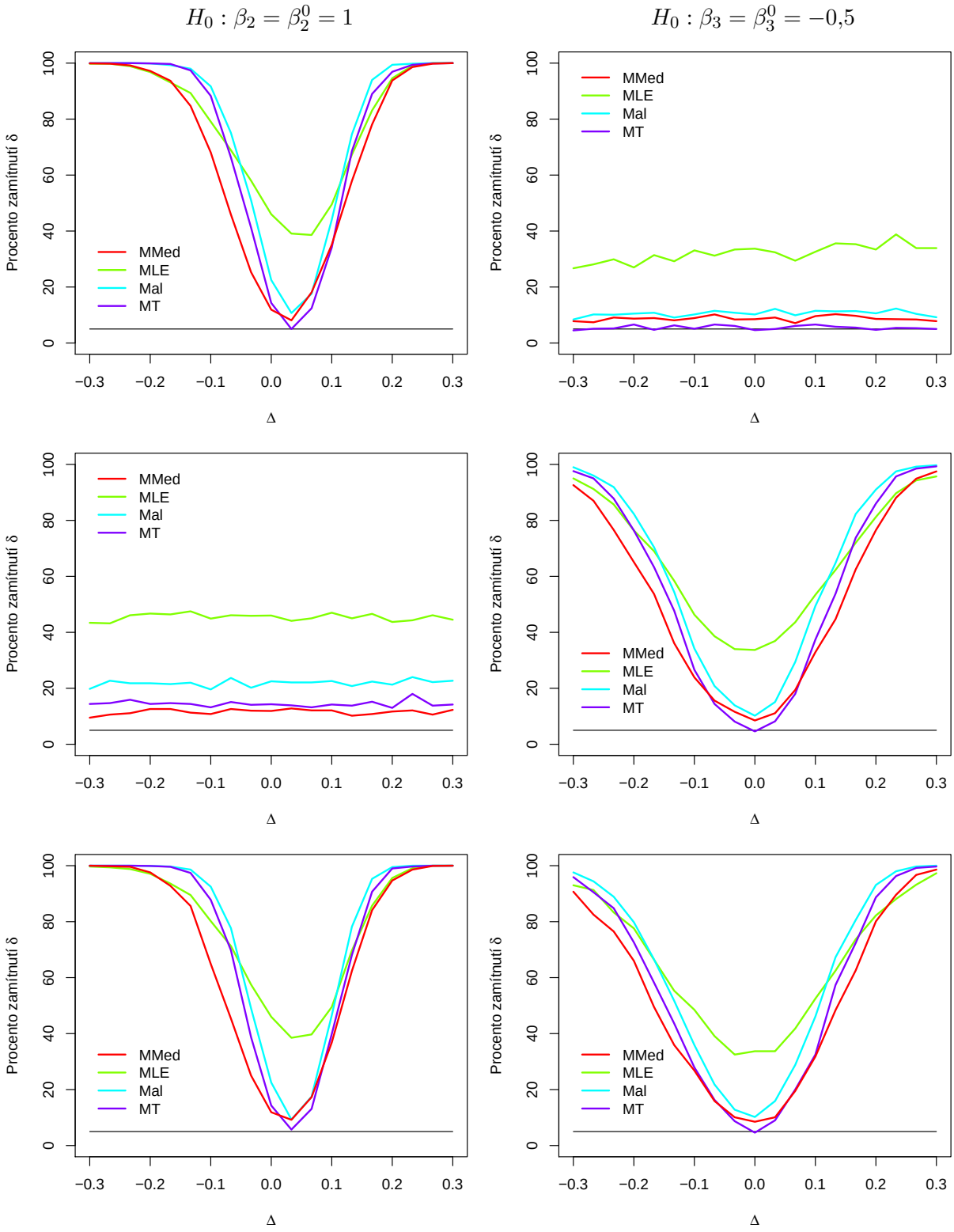
V případě nižší úrovně znečištění a odhadu pravděpodobnosti chyby 1. druhu se zbylé tři metody odhadu chovají velmi podobně a vykazují vyšší vychýlení od hranice 5 % při testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$. Pokud se přesuneme k síle testu, tak vidíme, že Mallowsův odhad a M-odhad založený na transformaci odezvy mají podobné křivky a zobecněný mediánový odhad má, stejně jako v případě čistých dat, širší křivku a tudíž i menší sílu testu. Všechny křivky ale zůstávají přibližně centrované.

Pokud zvýšíme úroveň znečištění na $\varepsilon = 0{,}1$, dojde k oddělení křivek Mallowsova odhadu a M-odhadu založeného na transformaci odezvy a v případě testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ také k vychýlení všech křivek znázorňujících sílu testu do kladné části osy x. V případě odhadu pravděpodobnosti chyby 1. druhu u testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ se kromě maximálně věrohodného odhadu od 5% hranice odchýlí i zbylé metody odhadu. Nejblíže této hranici je zobecněný mediánový odhad, následuje ho M-odhad založený na transformaci odezvy a nejvíce vychýlený ze tří uvažovaných robustních metod je Mallowsův odhad. V případě testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ se M-odhad založený na transformaci odezvy drží hranice 5 % a zobecněný mediánový odhad a Mallowsův odhad jsou od ní vychýleny, ale méně než v případě testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$.

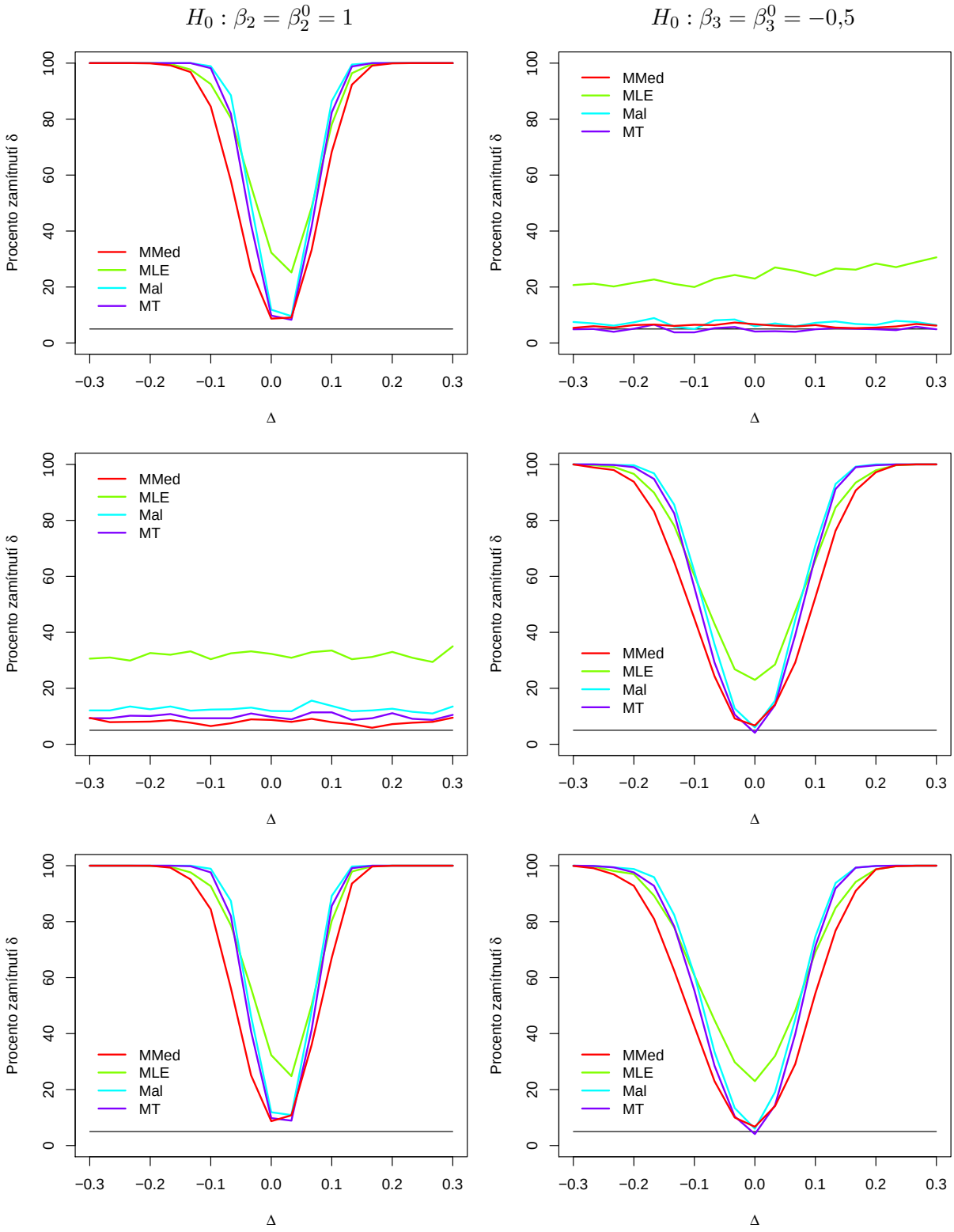
Přesuňme se nyní k variantě $n = 500$, pro kterou si můžeme prohlédnout výsledky na obrázcích 5.8 a 5.9. Pro nižší úroveň znečištění a odhad pravděpodobnosti chyby 1. druhu vyšly pro



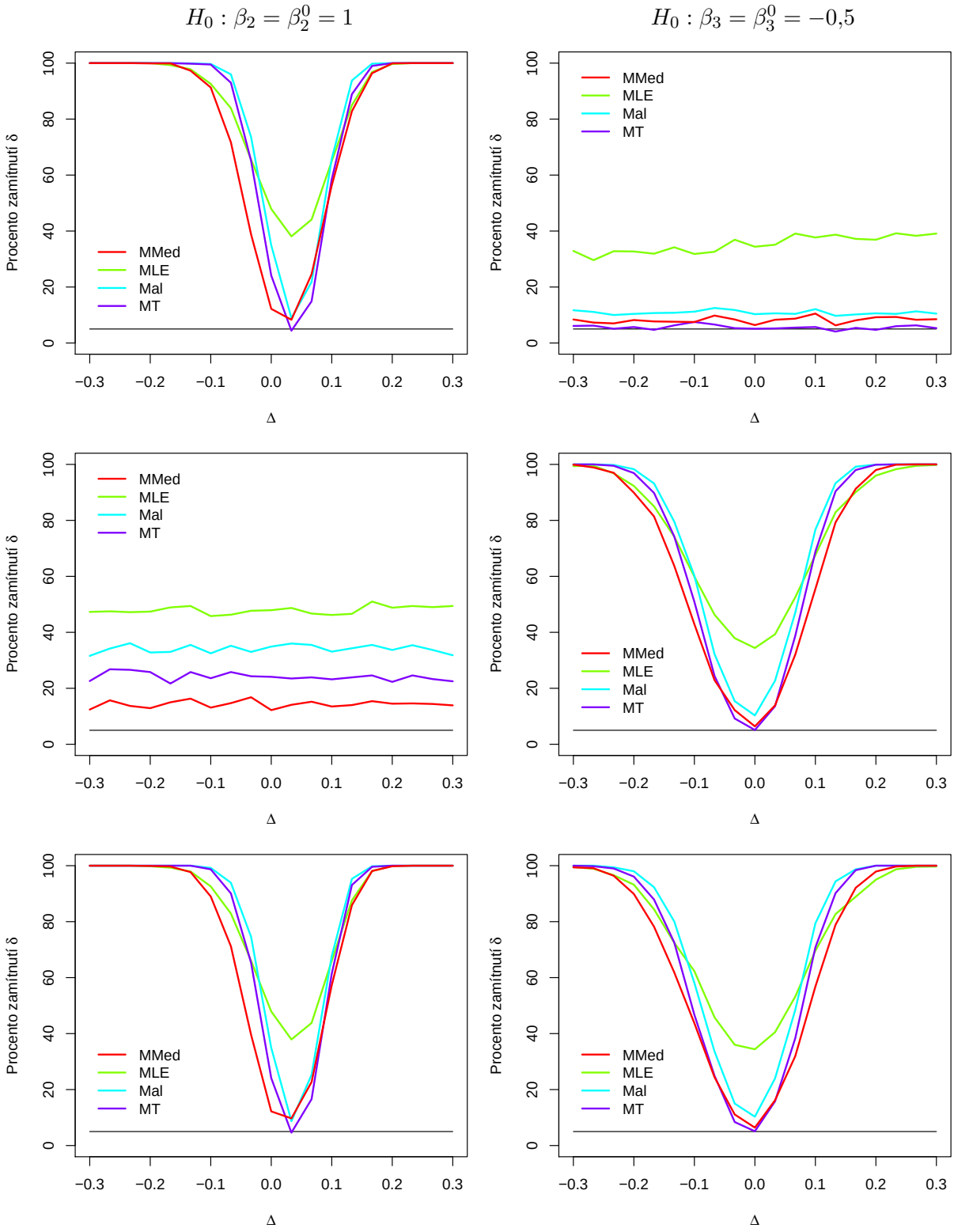
Obrázek 5.6: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 250$ a úroveň znečištění odlehlými pozorováními $\varepsilon = 0{,}05$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.7: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 250$ a úroveň znečištění odlehlými pozorováními $\varepsilon = 0{,}1$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.8: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 500$ a úroveň znečištění odlehlými pozorováními $\varepsilon = 0{,}05$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.9: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 500$ a úroveň znečištění odlehlými pozorováními $\varepsilon = 0{,}1$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.

test hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ velmi podobné grafy jako pro $n = 250$, u testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ došlo k oddělení křivek pro jednotlivé robustní odhady. V případě síly testu došlo díky zvýšení počtu pozorování k zúžení všech křivek, a tedy ke zlepšení síly testu.

U varianty úrovně znečištění $\varepsilon = 0{,}1$ došlo k významným změnám u odhadu pravděpodobnosti chyby 1. druhu u testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$. Mezi jednotlivými metodami odhadu vznikl výrazný rozdíl s tím, že zobecněný mediánový odhad zůstal nejblíže hranici 5 %. U síly testu došlo pro tuto hypotézu zvýšením úrovně znečištění, stejně jako v případě $n = 250$, k vychýlení křivek znázorňujících sílu testu do kladné části osy x. Celkově jsou všechny křivky znázorňující sílu testu podobné těm pro stejnou úroveň znečištění a menší počet pozorování. Zvýšením počtu pozorování pouze došlo k zúžení těchto křivek.

Simulační experiment nám ukázal, že přítomnost znečištění odlehlými pozorováními má vliv na pravděpodobnost chyby 1. druhu i na sílu testu. Pokud bychom neznali úroveň znečištění, tak pro test hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ je nejvhodnější zobecněný mediánový odhad a pro test hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ se jako nejvhodnější jeví M-odhad založený na transformaci odezvy.

5.7 Vliv pákových bodů

Druhý typ znečištění, který budeme uvažovat, je znečištění pákovými body. Nyní nebudeme upravovat hodnoty vysvětlované veličiny, nýbrž hodnoty vysvětlujících veličin.

Postup pro vytváření pákových bodů jsme převzali z práce [3] a vypadá následovně. Pro všechny vektory $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, určíme hodnotu $\lambda_i = \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}$, kde parametr $\boldsymbol{\beta}$ je dán vztahem (5.13). Vznikne nám vektor $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^{\mathrm{T}}$, jehož složky seřadíme vzestupně podle velikosti. To znamená, že najdeme permutaci $\pi : \{1, 2, \ldots, n\} \mapsto \{1, 2, \ldots, n\}$, která převede vektor $\boldsymbol{\lambda}$ na vektor $\tilde{\boldsymbol{\lambda}} = (\lambda_{\pi(1)}, \lambda_{\pi(2)}, \ldots, \lambda_{\pi(n)})^{\mathrm{T}}$, pro který platí $\lambda_{\pi(1)} \leq \lambda_{\pi(2)} \leq \ldots \leq \lambda_{\pi(n)}$.

Vysvětlující proměnné s pákovými body nyní vytvoříme pomocí vztahu

$$\tilde{\boldsymbol{x}}_{\pi(j)} = \begin{cases} \boldsymbol{x}_{\pi(n-j+1)}, & \text{je-li } j \leq \lfloor \varepsilon n \rfloor, \\ \boldsymbol{x}_{\pi(j)}, & \text{jinak.} \end{cases} \tag{5.17}$$

Tento předpis znamená, že $\lfloor \varepsilon n \rfloor$ vektorů $\boldsymbol{x}_i$, pro něž nabývá λ_i nejmenších hodnot, zaměníme za hodnotu $\lfloor \varepsilon n \rfloor$ vektorů $\boldsymbol{x}_i$, pro něž nabývá λ_i největších hodnot. Ostatní vektory $\boldsymbol{x}_i$ ponecháme beze změny.

Jelikož máme v rámci experimentů pevné vysvětlující veličiny, tvoříme vysvětlující veličiny s pákovými body pro konkrétní nastavení experimentu také pouze jednou. Dále při odhadování pak využíváme tyto znečištěné vysvětlující veličiny. Vysvětlované veličiny jsou však generovány pomocí původních, tedy neznečištěných, vysvětlujících veličin $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, a platí pro ně tedy $Y_i \sim \mathrm{Po}(\lambda_i)$, $i = 1, 2, \ldots, n$.

V rámci tohoto experimentu budeme uvažovat dvě úrovně znečištění, konkrétně $\varepsilon \in \{0{,}02; 0{,}04\}$. Zbylé nastavení experimentu je stejné jako v části 5.5.

Výsledky experimentu

Nejprve si představíme výsledky pro $n = 250$, které jsou znázorněny na obrázcích 5.10 a 5.11. Stejně jako v případě znečištění odlehlými pozorováními je nejcitlivější na přítomnost znečištění pákovými body maximálně věrohodný odhad. Nejméně citlivá metoda na přítomnost znečištění pákovými body je M-odhad založený na transformaci odezvy. U tohoto odhadu mají křivky

v podstatě stejný tvar jako v případě čistých dat. M-odhad založený na trasformaci odezvy tedy vůbec nezaznamená přítomnost znečištění pákovými body.

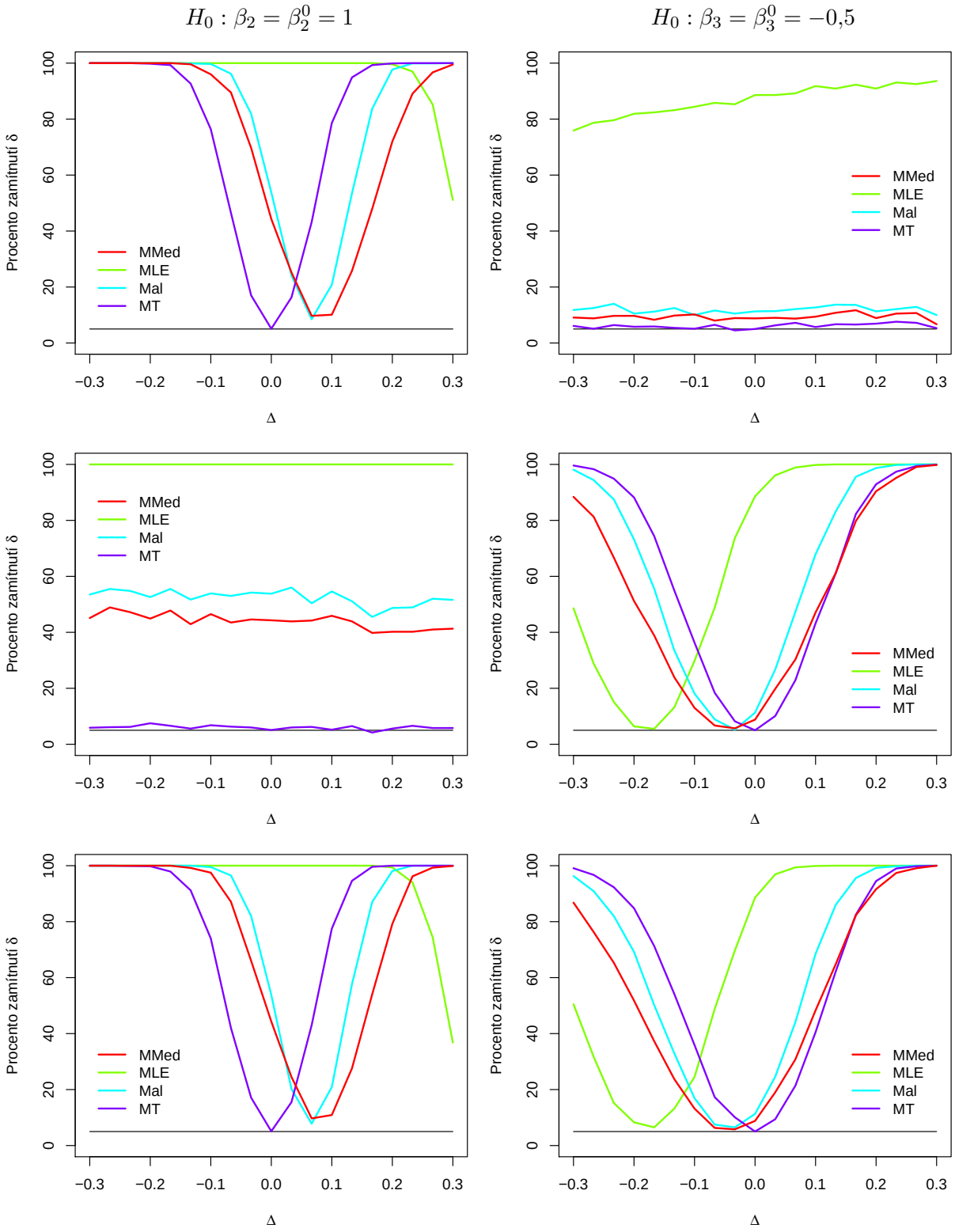
Pokud se zaměříme na nižší úroveň znečištění $\varepsilon = 0{,}02$, zjistíme, že v případě chyby 1. druhu jsou zbylé metody citlivější u testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$. U této hypotézy zamítá maximálně věrohodný odhad ve 100 % případů. Zobecněný mediánový odhad dosahuje nižšího odhadu pravděpodobnosti chyby 1. druhu než Mallowsův odhad, ale oba odhady jsou už velmi vzdáleny od hranice 5 %. U testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ se maximálně věrohodný odhad blíží k hranici zamítání ve 100 % případů. Zobecněný mediánový odhad a Mallowsův odhad jsou vychýleny od 5% hranice, ale jejich vychýlení není tak výrazné jako u testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$. V případě síly testu dochází u maximálně věrohodného odhadu, zobecněného mediánového odhadu a Mallowsova odhadu k vychýlení křivek do kladné části osy x pro test hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ a do záporné části osy x pro test hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$. Vychýlení je nejvýraznější pro maximálně věrohodný odhad.

Pokud jsou data znečištěna na úrovni $\varepsilon = 0{,}04$, tak se zvýrazní všechny výše popsané jevy. U odhadu pravděpodobnosti chyby 1. druhu se maximálně věrohodný odhad drží u hranice zamítání ve 100 % případů a zobecněný mediánový odhad a Mallowsův odhad se vzdálily ještě více od hranice 5 %. Dokonce se u těchto dvou odhadů vyskytl skok v křivce znázorňující odhad pravděpodobnosti chyby 1. druhu u testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$. Zvýšením úrovně znečištění se zvětšilo také vychýlení do kladné, případně záporné, části osy x u křivek znázorňujících sílu testu. U maximálně věrohodného odhadu a testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ dokonce nevidíme ani začátek křivky ve tvaru "U", v námi uvažovaném intervalu odhad zamítá nulovou hypotézu ve 100 % případů.

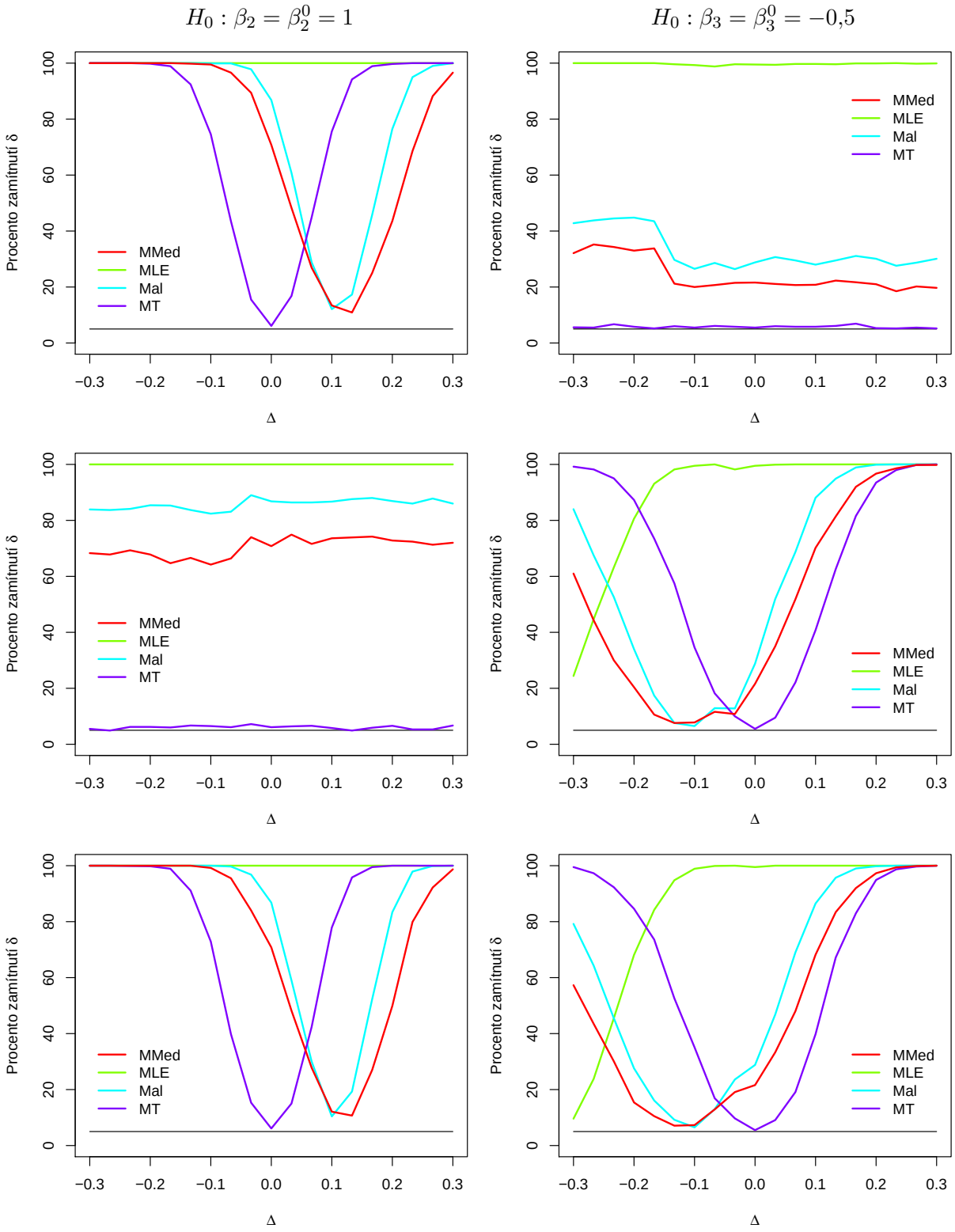
Výsledky simulačního experimentu pro počet pozorování $n = 500$ si můžeme prohlédnout na obrázcích 5.12 a 5.13. I v tomto případě je nejcitlivější maximálně věrohodný odhad a M-odhad založený na transformaci odezvy má opět křivky výrazně podobné těm pro čistá data.

Uvažujme nyní nižší úroveň znečištění $\varepsilon = 0{,}02$. U testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ zamítá maximálně věrohodný odhad až na jednu výjimku ve 100 % případů nezávisle na tom, zda zkoumáme chybu 1. druhu nebo sílu testu. V případě testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ a odhadu pravděpodobnosti chyby 1. druhu se u maximálně věrohodného odhadu objevuje výrazný skok v grafu pro případ $\Delta = -0{,}1$ a u síly testu můžeme pozorovat také výrazný výkyv v křivce. Skok v pravděpodobnosti chyby 1. druhu u testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ pozorujeme i u zobecněného mediánového odhadu a Mallowsova odhadu, ale není tak výrazný jako v případě maximálně věrohodného odhadu. Oba dva odhady jsou také vychýlené od 5% hranice. Ještě větší vzdálení od 5% hranice můžeme pozorovat u odhadu pravděpodobnosti chyby 1. druhu a testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$. U této hypotézy také pozorujeme vychýlení křivek znázorňujících sílu testu směrem do kladné části osy x. U testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ a síly testu jsou maximálně věrohodný odhad, zobecněný mediánový odhad a Mallowsův odhad vychýleny do záporné části osy x. Nejvýraznější vychýlení vykazuje maximálně věrohodný odhad. Zvýšením počtu pozorování se zúžily křivky znázorňující sílu testu.

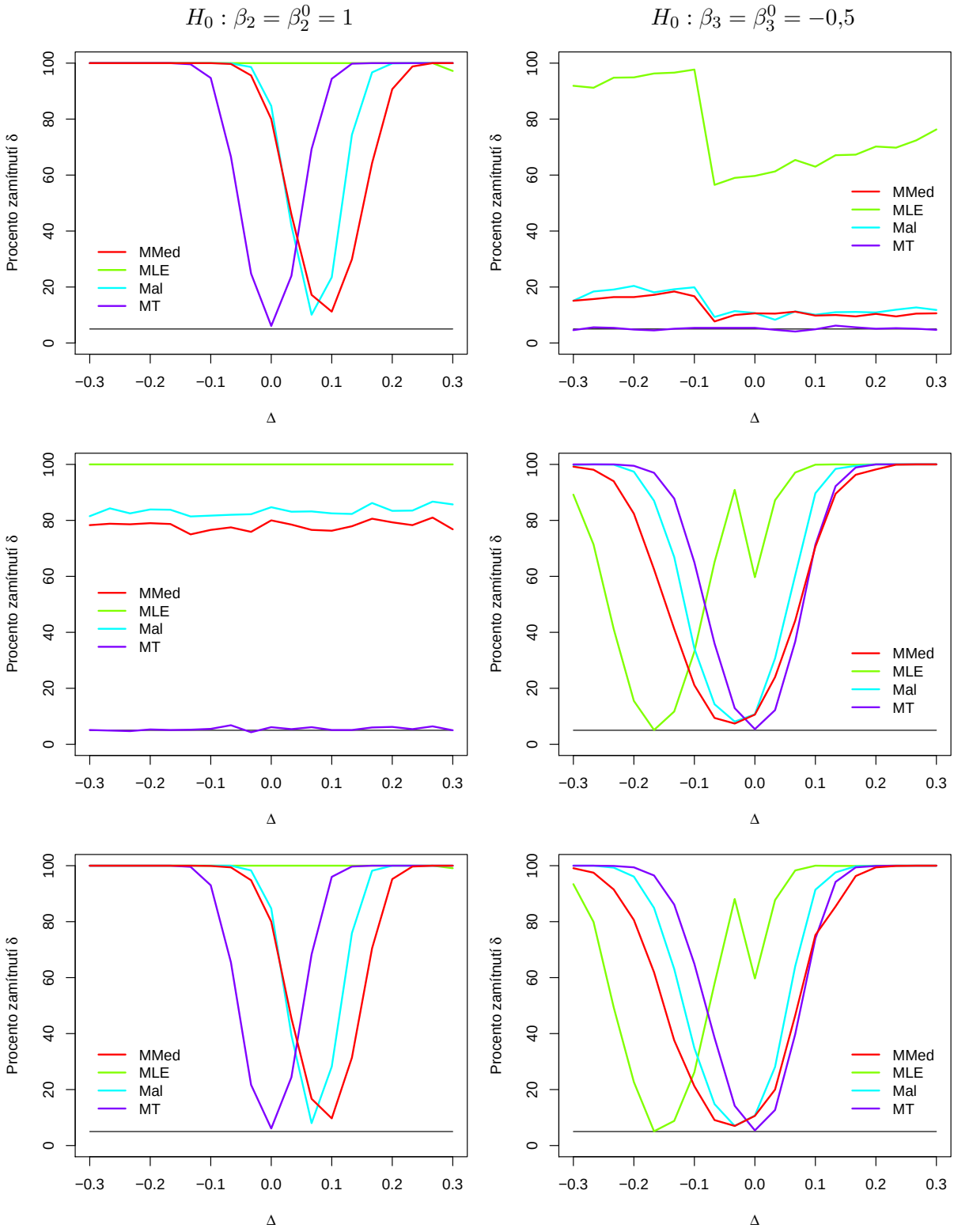
U vyšší úrovně znečištění $\varepsilon = 0{,}04$ už žádný odhad nevykazuje výrazné výkyvy v grafu, ale všechny ostatní výše jmenované problémy se ještě zvýraznily. U odhadu pravděpodobnosti chyby 1. druhu a testu hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ se Mallowsův odhad přiblížil k hranici zamítání ve 100 % případů a zobecněný mediánový odhad už je této hranici taktéž blízko. V případě testu hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ a odhadu pravděpodobnosti chyby 1. druhu se maximálně věrohodný odhad pohybuje kolem hranice 100 % a zobecněný mediánový odhad a Mallowsův odhad se zhoršily oproti variantě $\varepsilon = 0{,}02$. U síly testu se všechny křivky kromě M-odhadu založeného na transformaci odezvy ještě více vychýlily do kladné, případně záporné, části osy x.



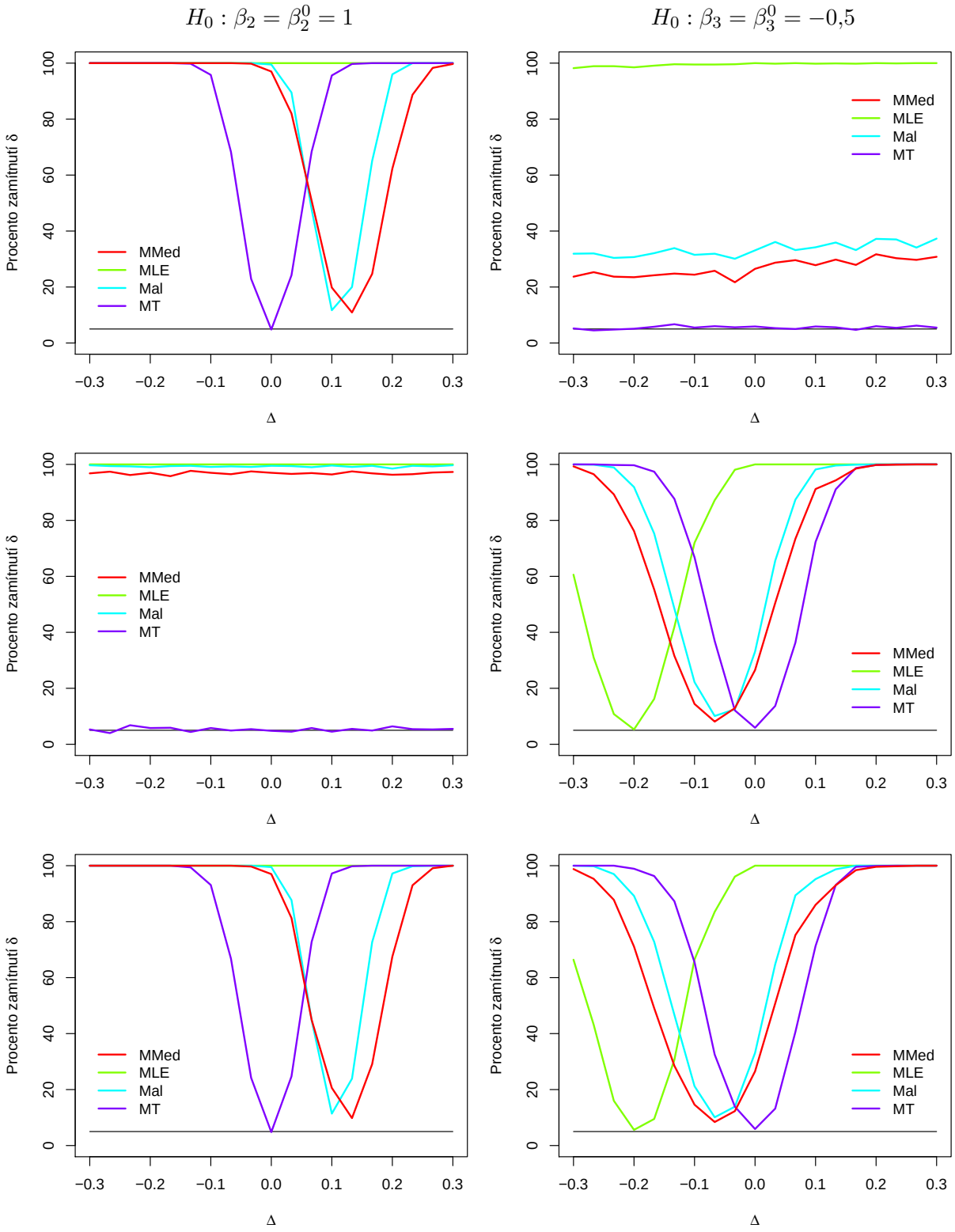
Obrázek 5.10: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 250$ a úroveň znečištění pákovými body $\varepsilon = 0{,}02$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.11: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 250$ a úroveň znečištění pákovými body $\varepsilon = 0{,}04$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.12: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 500$ a úroveň znečištění pákovými body $\varepsilon = 0{,}02$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.



Obrázek 5.13: Procento zamítnutí hypotéz $H_0 : \beta_2 = \beta_2^0 = 1$ (levý sloupec) a $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ (pravý sloupec) pro počet pozorování $n = 500$ a úroveň znečištění pákovými body $\varepsilon = 0{,}04$. V prvním řádku je $\boldsymbol{c} = (0; 0{,}5; 0)^{\mathrm{T}}$, ve druhém řádku je $\boldsymbol{c} = (0; 0; 1)^{\mathrm{T}}$ a ve třetím řádku je $\boldsymbol{c} = (0; 0{,}5; 1)^{\mathrm{T}}$. Černá čára značí 5% hranici.

Stejně jako v případě $\varepsilon = 0{,}02$ se zvýšením počtu pozorování zúžily křivky znázorňující sílu testu.

Pomocí simulačního experimentu jsme ukázali, že přítomnost znečištění pákovými body má vliv na pravděpodobnost chyby 1. druhu i sílu testu u maximálně věrohodného odhadu, zobecněného mediánového odhadu a Mallowsova odhadu. Oproti tomu M-odhad založený na transformaci odezvy není na tento typ znečištění citlivý a lze ho tedy doporučit pro data znečištěná pákovými body.

5.8 Poznámka k simulačním experimentům

Jak již bylo zmíněno, pro výpočet M-odhadu založeného na transformaci odezvy jsme využívali funkci `glmrob()` z balíčku robustbase, kde jsme nastavili `method = "MT"`. Během simulačních experimentů se objevil problém s touto funkcí, konkrétně nám někdy kromě číselné hodnoty vrátila i varování upozorňující na problém s konvergencí. Rozhodli jsme se takové odhady nevyřazovat a použili jsme hodnotu, kterou nám funkce vrátila. Výsledky námi provedených simulací ukazují, že to nemělo výrazný vliv na vlastnosti M-odhadu založeného na transformaci odezvy. Pro praxi je to ale nepříjemná vlastnost, protože nemáme jak ověřit, jak přesná je funkcí vrácená hodnota. U ostatních metod odhadu jsme problém s konvergencí nezaznamenali.

Pro představu o četnosti vracení varování uveďme konkrétní hodnoty pro některé simulace. U experimentu na chybu 1. druhu v závislosti na počtu pozorování jsme chybu obdrželi v 2,24 % případů. V případě počtu pozorování $n = 500$ a úrovně znečištění odlehlými pozorováními $\varepsilon = 0{,}1$ jsme chybu obdrželi v 3,05 % případů. Pokud jsme uvažovali počet pozorování $n = 500$ a úroveň znečištění pákovými body $\varepsilon = 0{,}04$, vrátila nám funkce varování v 3,36 % případů.

Závěr

Tato práce se věnovala robustnímu odhadování a testování v modelech poissonovské regrese. Seznámili jsme se s několika metodami odhadu, které lze použít při odhadování parametrů těchto modelů. Jako první jsme si představili maximálně věrohodný odhad, který se běžně používá a jehož nestabilita při odhadování ze znečištěných dat inspirovala vývoj robustních metod. Proto jsme dále uvedli několik zástupců již existujících robustních metod odhadů, jmenovitě mediánový odhad, Mallowsův odhad a M-odhad založený na transformaci odezvy. Odvodili jsme také nový robustní odhad vycházející z mediánového odhadu – zobecněný mediánový odhad.

Zbylá teoretická část se věnovala tomuto novému odhadu. Odvodili jsme asymptotické rozdělení zobecněného mediánového odhadu a diskutovali jsme konzistentní odhad jeho teoretické asymptotické kovarianční matice. Poté jsme se přesunuli k testování hypotéz, kde jsme nejprve navrhli testovací statistiku Waldova typu a určili jsme její asymptotické rozdělení. Získali jsme tak možnost testovat hypotézy o parametrech modelu poissonovské regrese.

Poslední kapitola byla věnována simulačním experimentům, jejichž cílem bylo ověřit fungování odvozených teoretických vlastností v praxi. Provedli jsme experiment ověřující konzistenci zobecněného mediánového odhadu, ze kterého nám vyšlo, že zatím není důvod se domnívat, že by zobecněný mediánový odhad nebyl konzistentním odhadem.

Výsledky simulací pro čistá data a chybu 1. druhu ukázaly dobrou stabilitu pro maximálně věrohodný odhad, zobecněný mediánový odhad a Mallowsův odhad. Oproti tomu M-odhad založený na transformaci odezvy je pro menší počet pozorování nestabilní. Více se to projevuje u testu hypotézy $H_0 : \beta_2^0 = 1$, kde dosáhl pro počet pozorování $n = 50$ odhad pravděpodobnosti chyby 1. druhu dokonce více než 14 %. U síly testu měly všechny metody odhadu očekávaný tvar křivky. Křivky maximálně věrohodného odhadu, Mallowsova odhadu a M-odhadu založeného na transformaci odezvy dokonce splývaly. Křivka zobecněného mediánového odhadu byla oproti ostatním širší, což odpovídá tomu, že zobecněný mediánový odhad potřebuje pro stejnou sílu testu odhadovat z více dat. Celkově ale všechny grafy získané pro zobecněný mediánový odhad odpovídají odvozené teorii.

U znečištění odlehlými pozorováními je patrný vliv na všechny metody odhadu. Dle očekávání byl nejcitlivější na přítomnost znečištění maximálně věrohodný odhad. Nový zobecněný mediánový odhad v konkurenci již existujících robustních metod odhadu obstál a pro test hypotézy $H_0 : \beta_2 = \beta_2^0 = 1$ a vyšší úroveň znečištění $\varepsilon = 0{,}1$ se jeví být dokonce nejlepší z porovnávaných metod. Pro nižší úroveň znečištění $\varepsilon = 0{,}05$ je jen mírně horší než M-odhad založený na transformaci odezvy. Pro test hypotézy $H_0 : \beta_3 = \beta_3^0 = -0{,}5$ se jeví jako nejvhodnější M-odhad založený na transformaci odezvy. Zobecněný mediánový odhad v tomto případě ale není výrazně horší.

U znečištění pákovými body vyšel jako nejcitlivější opět maximálně věrohodný odhad, oproti tomu M-odhad založený na transformaci odezvy přítomnost znečištění nezaznamenal a je tedy nejvhodnější pro použití při odhadování z dat, která jsou znečištěna pákovými body. Zobec-

něný mediánový odhad a Mallowsův odhad jsou citlivější na přítomnost pákových bodů než na přítomnost odlehlých pozorování.

Přínos této práce spočívá v zavedení nového robustního odhadu pro modely poissonovské regrese a vybudování potřebné teorie k tomu, aby se v praxi mohl používat. Simulační experimenty potvrdily, že dokáže konkurovat již existujícím robustním metodám a v některých případech dokáže být dokonce lepší.

Literatura

[1] J. Anděl, *Základy matematické statistiky*. MatfyzPress, Praha, 2005.

[2] A. M. Bianco, E. Martínez, *Robust testing in the logistic regression model*. Computational Statistics and Data Analysis 53, 2009, 4095–4105.

[3] P. Bohuslav, *Robustní odhady v zobecněných lineárních modelech*. Diplomová práce, FJFI ČVUT, Praha, 2016.

[4] E. Cantoni, E. Ronchetti, *Robust Inference for Generalized Linear Models*. Journal of the American Statistical Association 96 (455), 2001, 1022–1030.

[5] A. J. Dobson, A. G. Barnett, *An introduction to Generalized Linear Models*. CRC Press, Boca Raton, 2008.

[6] T. S. Ferguson, *A Course in Large Sample Theory*. Chapman & Hall, London, 1996.

[7] A. Ghosh, A. Mandal, N. Martín, L. Pardo, *Influence analysis of robust Wald-type tests*. Journal of Multivariate Analysis 147, 2016, 102–126.

[8] T. Hobza, N. Martín, L. Pardo, *A Wald-type test statistic based on robust modified median estimator in logistic regression models*. Journal of Statistical Computation and Simulation 87(12), 2017, 2309–2333.

[9] T. Hobza, L. Pardo, I. Vajda, *Robust median estimator in logistic regression*. Journal of Statistical Planning and Interference 138, 2008, 3822–3840.

[10] R. A. Maronna, R. D. Martin, V. J. Yohai, *Robust Statistics: Theory and Methods*. Wiley, Chichester, 2006.

[11] Ch. E. McCulloch, S. R. Searle, J. M. Neuhaus, *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New Jersey, 2008.

[12] J. Novotná, *Robustní odhady parametrů logistického regresního modelu*. Bakalářská práce, FJFI ČVUT, Praha, 2018.

[13] J. Shao, *Mathematical Statistics*. Springer-Verlag, New York, 1999.

[14] M. Valdora, V. J. Yohai, *Robust estimators for generalized linear models*. Journal of Statistical Planning and Interference 146, 2014, 31–48.

[15] R. W. M. Wedderburn, *Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method*. Biometrika 61, 1974, 439–447.

Příloha: Tabulka hodnot C_k

k	C_k	k	C_k	k	C_k	k	C_k
0	0,693147	25	25,667437	50	50,667057	75	75,666928
1	1,678347	26	26,667408	51	51,667049	76	76,666924
2	2,674060	27	27,667381	52	52,667042	77	77,666921
3	3,672061	28	28,667356	53	53,667035	78	78,666918
4	4,670909	29	29,667333	54	54,667028	79	79,666915
5	5,670161	30	30,667311	55	55,667022	80	80,666912
6	6,669637	31	31,667291	56	56,667015	81	81,666909
7	7,669249	32	32,667272	57	57,667009	82	82,666906
8	8,668951	33	33,667254	58	58,667004	83	83,666903
9	9,668715	34	34,667237	59	59,666998	84	84,666900
10	10,668522	35	35,667221	60	60,666992	85	85,666897
11	11,668363	36	36,667206	61	61,666987	86	86,666895
12	12,668229	37	37,667191	62	62,666982	87	87,666892
13	13,668115	38	38,667178	63	63,666977	88	88,666890
14	14,668016	39	39,667165	64	64,666972	89	89,666887
15	15,667930	40	40,667153	65	65,666968	90	90,666885
16	16,667854	41	41,667141	66	66,666963	91	91,666882
17	17,667786	42	42,667130	67	67,666959	92	92,666880
18	18,667726	43	43,667119	68	68,666954	93	93,666878
19	19,667672	44	44,667109	69	69,666950	94	94,666875
20	20,667624	45	45,667099	70	70,666946	95	95,666873
21	21,667579	46	46,667090	71	71,666942	96	96,666871
22	22,667539	47	47,667081	72	72,666939	97	97,666869
23	23,667502	48	48,667073	73	73,666935	98	98,666867
24	24,667468	49	49,667065	74	74,666931	99	99,666865

Tabulka: Hodnoty konstant C_k z Definice 1.