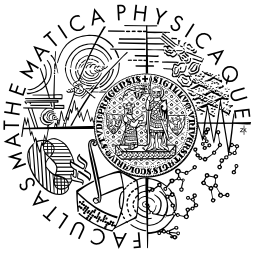

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Dominik Matula

Modely pro data s nadbytečnými nulami

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	doc. Mgr. Michal Kulich, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie

Praha 2016

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

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

V Praze dne 21. července 2016 Dominik Matula

Název práce: Modely pro data s nadbytečnými nulami

Autor: Dominik Matula

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Mgr. Michal Kulich, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Cílem práce je podat ucelený přehled hlavních přístupů k modelování dat zatížených nadbytečnými nulami. Autor se věnuje třem podtřídám *modelů s upraveným počtem nul* (ZMM), a sice *modelům s nadbytečnými nulami*, jimž je věnována stěžejní část práce, *modelům bez nulové odezvy* a *hradbovým modelům*. Modely každé podtřídy vždy nejprve řádně definuje, posléze se zabývá konstrukcí maximálně věrohodných odhadů regresních koeficientů. V rámci modelů ZMM se setkáváme především s modely založenými na Poissonově či negativně binomickém rozdělení typu 2 (NB2). V této práci jsou provedena zobecnění na modely ZMM vycházející obecně z diskrétních rozdělení exponenciálního typu. Odvozen je i postup, jímž lze v těchto modelech získat maximálně věrohodné odhady regresních koeficientů. Dosavadní práce se téměř nevěnovaly modelům ZIM založeným na negativně binomickém rozdělení typu 1 (NB1). Toto rozdělení není exponenciálního typu, nelze proto použít standardní přístup ke konstrukci odhadů regresních koeficientů. Autor však navrhuje modifikaci tohoto přístupu pro modely ZIM založené na NB1 využívaje metodu *kvazi-věrohodnosti*. Práci uzavírají dvě simulační studie.

Klíčová slova: modely s nadbytečnými nulami, poissonovské modely s nadbytečnými nulami, modely bez nulové složky, hradbové modely, EM algoritmus

Title: Models for zero-inflated data

Author: Dominik Matula

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Michal Kulich, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The aim of this thesis is to provide a comprehensive overview of the main approaches to modeling data loaded with redundant zeros. There are three main subclasses of *zero modified models* (ZMM) described here – *zero inflated models* (the main focus lies on models of this subclass), *zero truncated models and hurdle models*. Models of each subclass are defined and then a construction of maximum likelihood estimates of regression coefficients is described. ZMM models are mostly based on Poisson or negative binomial type 2 distribution (NB2). In this work, author has extended the theory to ZIM models generally based on any discrete distributions of exponential type. There is described a construction of MLE of regression coefficients of theese models, too. Just few of present works are interested in ZIM models based on negative binomial type 1 distribution (NB1). This distribution is not of exponential type therefore a common method of MLE construction in ZIM models cannot be used here. In this work provides modification of this method using quasi-likelihood method. There are two simulation studies concluding the work.

Keywords: zero inflated models, zero inflated Poisson models, zero truncated models, hurdle models, EM algorithm

Na prvním místě Bohu, jenž mne stvořil. Pak rodičům, kteří mne vychovali a dopřáli mi vzdělání. Svému školiteli doc. Kulichovi, který mne vedl při psaní této práce a byl mi nápomocen mnoha užitečnými radami. Všem přátelům, kteří trpělivě naslouchali mému vyprávění o modelech a nadbytečných nulách. A hlavně své milé manželce Markétě, která vedle mne stála a která mne podporovala, kdykoliv jsem klesal na mysli.

Těm všem a mnohým dalším bych chtěl na tomto místě poděkovat, neboť bez nich bych se na konec své diplomové práce nikdy nedostal.

Obsah

Seznam použitého značení

Obory čísel a funkce

$\mathcal{N}$	obor přirozených čísel	
$\mathcal{N}_0$	obor přirozených čísel včetně nuly	
$\mathcal{R}$	obor reálných čísel	
$f(x)$	reálná funkce reálné proměnné	
$f'(x)$	derivace funkce $f(x)$	
$f(x,y,z)$	reálná funkce více reálných proměnných	
$\frac{\partial f(x,y,z)}{\partial x}$	parciální derivace funkce $f(x,y,z)$ podle x	
$\frac{\partial^2 f(x,y,z)}{\partial x^2}$	druhá parciální derivace funkce $f(x,y,z)$ dle x	
$\frac{\partial^2 f(x,y,z)}{\partial x \partial y}$	druhá parciální derivace funkce $f(x,y,z)$ dle x a y	

Maticový počet

A	skalár	$A \in \mathcal{R}$
$\mathbf{A}$	reálný vektor	$\mathbf{A} \in \mathcal{R}^j$
$\mathbb{A}$	reálná matice	$\mathbb{A} \in \mathcal{R}^{j \times l}$
$\mathbf{A}^\mathsf{T}, \mathbb{A}^\mathsf{T}$	transpozice vektoru (matice)	
$\mathbf{A}^\mathsf{T}\mathbf{B}$, $\mathbb{A}^\mathsf{T}\mathbb{B}$	skalární součin vektorů (matic)	

Regresní modely

$Y_i, \mathbf{Y}$	odezva (vektor odezvy)	
$X_{ij}, \mathbf{X}_i, \mathbb{X}$	regresor (vektor regresorů, matice regresorů)	
k	počet složek vektoru regresorů $\mathbf{X}_i$	$k \in \mathcal{N}$
n	počet pozorování	$n \in \mathcal{N}$
i	index pozorovnání	$1 \leq i \leq n$
$V_{ij}, \mathbf{V}_i, \mathbb{V}$	regresor (vektor regresorů, matice regresorů) pro parametry podkladového rozdělení modelu	
$\beta_i, \boldsymbol{\beta}$	regresní koeficient (vektor regresních koeficientů) pro parametry podkladového rozdělení modelu	
v	počet složek vektoru regresorů $\mathbf{V}_i$ a vektoru $\boldsymbol{\beta}$	$v \in \mathcal{N}$
$W_{ij}, \mathbf{W}_i, \mathbb{W}$	regresor (vektor regresorů, matice regresorů) pro pravděpodobnost *perfektního* stavu	
$\gamma_i, \boldsymbol{\gamma}$	regresní koeficient (vektor regresních koeficientů) pro pravděpodobnost *perfektního* stavu	
w	počet složek vektoru regresorů $\mathbf{W}_i$ a vektoru $\boldsymbol{\gamma}$	$w \in \mathcal{N}$
$Z_i, \mathbf{Z}$	latentní veličina (vektor latentních veličin)	

Metoda maximální věrohodnosti

$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma})$	věrohodnost pozorovaných dat o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	
$\ell(\boldsymbol{\beta},\boldsymbol{\gamma})$	logaritmická věrohodnost pozorovaných dat o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	
$\mathbf{U}(\boldsymbol{\beta},\boldsymbol{\gamma})$	skórová funkce pozorovaných dat o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	
$\mathbb{I}(\boldsymbol{\beta},\boldsymbol{\gamma})$	očekávaná Fisherova informační matice o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	
$\mathcal{L}_c(\boldsymbol{\beta},\boldsymbol{\gamma})$	věrohodnost úplných dat $\left(Y_i,\mathbf{X}_i^\intercal,Z_i\right)$, $i=1,\ldots,n$ o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	
$\ell_c(\boldsymbol{\beta},\boldsymbol{\gamma})$	logaritmická věrohodnost úplných dat $\left(Y_i,\mathbf{X}_i^\intercal,Z_i\right)$, $i=1,\ldots,n$ o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	
$\mathbf{U}_c(\boldsymbol{\beta},\boldsymbol{\gamma})$	skórová funkce úplných dat $\left(Y_i,\mathbf{X}_i^\intercal,Z_i\right)$, $i=1,\ldots,n$ o regresních koeficientech $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$	

EM algoritmus

j	index iterace EM algoritmu	$j \in \mathcal{N}$
$\boldsymbol{\beta}^{(j)}$	přiblížení maximálně věrohodného odhadu $\boldsymbol{\beta}$ v j-tém kroku EM algoritmu	$\boldsymbol{\beta}^{(j)} \in \mathcal{R}^v$
$\boldsymbol{\gamma}^{(j)}$	přiblížení maximálně věrohodného odhadu $\boldsymbol{\gamma}$ v j-tém kroku EM algoritmu	$\boldsymbol{\gamma}^{(j)} \in \mathcal{R}^w$
$Z_i^{(j)}, \mathbf{Z}^{(j)}$	aposteriorní střední hodnota náhodné veličiny Z_i (náh. vektoru $\mathbf{Z}$) v j-tém kroku EM algoritmu	
$\mathcal{Q}\left(\theta,\theta^{(j)}\right)$	střední hodnota logaritmické věrohodnosti pozorovaných dat vzhledem k podmíněnému rozdělení $\mathbf{Z}\|\mathbb{X}$, přičemž za neznámé parametry je voleno jejich aktuální přiblížení	

Rozdělení náhodných veličin

$Poiss(\lambda)$	Poissonovo rozdělení s parametrem λ	$\lambda > 0$
$Alt(p)$	Alternativní rozdělení s parametrem p	$p \in [0,1]$
$NB(\lambda,\alpha)$	Negativně binomické rozdělení s parametry λ a α	$\lambda > 0$, $\alpha > 0$
$\Gamma(\alpha,\delta)$	Gamma rozdělení s prametry α a δ	$\alpha > 0, \delta > 0$
$\mathcal{G}\boldsymbol{\theta}$	rozdělení exponenciálního typu s parametrem $\boldsymbol{\theta}$	
$\mathcal{N}_G$	nosič rozdělení G	$\mathcal{N}_G \subset \mathcal{N}_0$
$g(x)$	hustota rozdělení G	$x \in \mathcal{N}_G$
$g_T(x)$	hustota rozdělení G zbaveného nulové složky	$x \in \mathcal{N}_G \setminus \{0\}$

Konvergence posloupností

$\to$	konvergence
$\xrightarrow{\mathcal{D}}$	konvergence v distribuci

1. Úvod

1.1 Názvosloví regresních modelů pro četnosti

K modelování počtů událostí či výskytů nějakého jevu se běžně používá poissonovských regresních modelů. Příkladem může být situace, v níž sledujeme počty chyb v nějakém výrobním procesu. Je-li však stroj seřízen správně, je pravděpodobnost chyby zanedbatelná a odezva je tedy skoro jistě nulová. Právě popsaná situace je jedním z typických uplatnění *modelů s nadbytečnými nulami* (*zero inflated models*, ZIM). Není tedy divu, že se i mezi ZIM často setkáváme s modely, v nichž se odezva řídí vhodně upraveným Poissonovým rozdělením.

Modely s nadbytečnými nulami, jimiž se v této práci budeme primárně zabývat, patří do obecnější třídy regresních modelů, mezi tzv. *modely s upraveným počtem nul* (*zero modified models*, ZMM). Spadají sem i modely, v nichž je nulová složka odezvy potlačena, tzv. *modely se sníženým počtem nul* (*zero deflated models*, ZDM). Těmi se zde ale zabývat nebudeme, vyjma speciálního případu, tzv. *modelů bez nulové odezvy* (*zero truncated models*, ZTM) – s jejich pomocí je totiž možné konstruovat tzv. *hradbové modely* (*hurdle models*, HM), které, jak uvidíme v kapitole 6, mohou představovat alternativu právě k *modelům s nadbytečnými nulami*.

V kapitole 3 se podrobněji zaměříme na *poissonovský model s nadbytečným množstvím nul* (*zero inflated Poisson model*, ZIP), a v kapitolách 2.3 a 6.4 na další modely využívající Poissonovo rozdělení ve snaze modelovat data s upraveným počtem nul, a sice po řadě *poissonovský model bez nulové odezvy* (*zero truncated Poisson model*, ZTP) a *poissonovský hradbový model* (*Poisson hurdle model*, PHM).

Poissonovo rozdělení můžeme při konstrukci *modelů pro data s upraveným počtem nul* samozřejmě nahradit jinými rozděleními. V rámci této práci jsme se pokusili zobecnit přístup použitý v článku Lambert (1992) v případě modelů ZIP na modely ZMM založené obecně na *rozděleních exponenciálního typu*, neboť právě do této třídy Poissonovo rozdělení přirozeně spadá. Vzhledem k aplikacím ZMM, v nichž modelujeme především počty událostí a jevů, jsme se přitom omezili pouze na diskrétní rozdělení spadající do této třídy. Takto zvolenými modely ZMM se budeme zabývat v kapitole 4, v kapitole 2 pak *modely bez nulové odezvy* (*zero truncated models*, ZTM) a v kapitole 6 *hradbovými modely* (*hurdle models*, HM), v nichž se odezva řídí takto zvoleným rozdělením.

Jsou-li modelovaná data zatížena nadměrnou dispersí, používáme namísto Poissonova rozdělení zpravidla rozdělení negativně binomické. Stejný přístup lze aplikovat i v případě modelů ZMM. V kapitole 5 se tedy budeme věnovat *negativně binomickému modelu s nadbytečnými nulami* (*zero inflated negative binomial model*, ZINB), v kapitolách 2.4 a 6.5 pak po řadě *negativně binomickému modelu bez nulové odezvy* (*zero truncated negative binomial model*, ZINB) a *negativně binomickému hradbovému modelu* (*negative binomial hurdle model*, NBHM).

V literatuře se ovšem setkáváme hned s několika různými rozděleními, jenž bývají označována jako *negativně binomická* (*NB*). Tato nekonzistence se samozřejmě projevuje i mezi modely ZINB (resp. ZTNB a NBHM). Většina autorů přitom vychází z rozdělení též označovaného jako *negativně binomické rozděle-*

Tabulka 1.1: Přehled použitých zkratek tříd regresních modelů, jimiž se v této práci budeme zabývat. Spolu s běžně používaným anglickým názvem je vždy uveden i návrh českého označení, které budeme používat v rámci této práci. Zkratka NB zde nahrazuje *negativně binomický* (resp. *negative binomial*).

Zkr.	**Anglický název**	**Český název**
ZMM	Zero modified model	Model s upraveným počtem nul
ZTM	Zero truncated model	Model bez nulové odezvy
ZTP	Zero truncated Poisson model	Poisson. model bez nulové odezvy
ZTNB	Zero truncated NB model	NB model bez nulové odezvy
ZIM	Zero inflated model	Model s nadbytečnými nulami
ZIP	Zero inflated Poisson model	Poiss. model s nadbytečnými nulami
ZINB	Zero inflated NB model	NB model s nadbytečnými nulami
HM	Hurdle model	Hradbový model
PHM	Poisson hurdle model	Poissonovský hradbový model
NBHM	NB hurdle model	NB hradbový model

ní typu 2 (NB2). Jak ukážeme v podkapitole 1.2, při pevné volbě parametru dispersního parametru je toto rozdělení rozdělením exponenciálního typu. Je tedy snadné získat výsledky o tomto speciálním případu *negativně binomických modelů s upraveným množstvím nul*, a sice prostou aplikací výsledků obecnějších kapitol. Toho v této práci využijeme i my.

Dalším *negativně binomickým rozdělením*, s nímž se v literatuře setkáme, je tzv. *negativně binomické rozdělení typu 1* (NB1). Jak uvedeme v podkapitole 1.2, nejedná se o rozdělení exponenciálního typu a to ani v případě, že je dispersní parametr známý. V kapitole 5.2.2 však ukážeme, že obdobných výsledků jako v případě modelů ZIM lze dosáhnout u modelů ZINB založených na rozdělení NB1 vhodnou modifikací standardního postupu. V žádné práci věnované modelům ZINB jsme se s tímto přístupem nesetkali. Poznamenejme ještě, že obdobný postup je zřejmě možný i v případě *modelů bez nulové odezvy* a *hradbových modelů* založených na rozdělení NB1, pro omezený rámec této práce se tím zde ale zabývat nebudeme.

Přehled tříd modelů, jimiž se v této práci zabýváme, dává tabulka 1.1. Ač jsou v literatuře anglické názvy jednotlivých tříd již poměrně ustálené, jejich české ekvivalenty často úplně chybí. V tabulce jsou proto uvedeny návrhy českých názvů, jak je budeme v rámci této práce používat.

1.2 Zobecněné lineární modely

Jak jsme zmínili v předchozí podkapitole, cílem této práce je představit modely pro počty událostí, jevů atp. v případě, kdy jsou data zatížena nadbytečným množstvím nul. Jak uvidíme v dalších kapitolách, budeme o pozorovaných datech předpokládat, že pocházejí z jednoho ze dvou stavů – první z nich, v dalším označován jako *perfektní stav*, umožňuje pouze nulovou odezvu, druhý pak představuje

původní stav, v němž události generující proces není zatížen nadbytečnými nulami. *Perfektní stav* je přitom zřejmě spjat s rozdělením degenerovaným v nule, druhý stav pak s nějakým známým rozdělením, které v dalším budeme nazývat *podkladové rozdělení*. Je tedy důležité rozhodnout, jaká rozdělení budeme uvažovat za *podkladová*.

Již jsme zmínili, že v situaci, kdy data nejsou kontaminována nadbytečnými nulami, běžně k modelování počtů událostí používáme *poissonovský regresní model*. Ten přitom spadá do třídy *zobecněných lineárních modelů* (*generalized linear models*, GLM), v nichž se odezva řídí rozdělením tzv. *exponenciálního typu*. V této práci zobecníme teorii představenou ve článku Lambert (1992) pro modely ZIP na případ, kdy je podkladové rozdělení *diskrétní rozdělení exponenciálního typu*. Věnujme proto tuto podkapitolu krátkému připomenutí teorie *zobecněných lineárních modelů* a konstrukci maximálně věrohodných odhadů v této třídě regresních modelů.

Předpokládejme tedy, že máme k dipozici nezávislá, stejně rozdělená pozorování $\left(Y_i, \mathbf{X}_i^{\mathsf{T}}\right)$, $i = 1, 2, \ldots, n$ splňující:

1. Rozdělení odezvy Y_i závisí na regresorech $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})^{\mathsf{T}}$ skrze vektor regresních koeficientů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathsf{T}}$. Podmíněnou střední hodnotu odezvy označíme $\mu_i = \mathsf{E}[Y_i|\mathbf{X}_i]$.

2. Hustota $Y_i|\mathbf{X}_i$ patří do třídy hustot *exponenciálního typu* a můžeme ji tedy zapsat jako

$$g(y; \psi_{\mathrm{i}}, \varphi) = \exp\left(\frac{y\,\psi_{\mathrm{i}} - b(\psi_{\mathrm{i}})}{\varphi} + c(y, \varphi)\right), \tag{1.1}$$

   kde $b(.)$ je známá dvakrát spojitě diferencovatelná funkce, ψ_{i} je funkcí regresorů $\mathbf{X}_i$ a vektoru regresních koeficientů $\boldsymbol{\beta}$ a φ je známá anebo neznámá konstanta.

3. ψ_{i} závisí na $\mathbf{X}_i$ a $\boldsymbol{\beta}$ skrze tzv. *lineární prediktor* $\eta_i = \mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}$.

4. Existuje známá ryze monotonní, dvakrát spojitě diferencovatelná funkce $h(.)$, tzv. linková funkce, splňující $h(\mu_i) = \eta_i$. Pokud platí $h(\mu_i) = \psi_{\mathrm{i}}$, nazýváme linkovou funkci *kanonickou*.

Pozorování $\left(Y_i, \mathbf{X}_i^{\mathsf{T}}\right)$, $i = 1, 2, \ldots, n$ pak splňují *zobecněný lineární model*.

V dalším budeme označovat $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathsf{T}}$ vektor odezvy a $\mathbb{X}$ matici regresorů, pro níž platí $\mathbb{X}^{\mathsf{T}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Předpokládáme, že má matice $\mathbb{X}$ plnou sloupcovou hodnost, tedy $\mathrm{h}(\mathbb{X}) = k$. Dále připomeňme, že pro podmíněnou střední hodnotu a podmíněný rozptyl odezvy platí

$$\mu_i = \mathsf{E}[Y_i|\mathbf{X}_i] = b'(\psi_{\mathrm{i}}), \qquad \mathrm{var}[Y_i|\mathbf{X}_i] = \varphi\, V(\mu_i),$$

kde $V(.)$ je tzv. *rozptylová funkce*, která je jednoznačně určena funkcí $b(.)$, neboť platí vztah

$$b''(\psi_{\mathrm{i}}) = V(b'(\psi_{\mathrm{i}})).$$

Rozptylová funkce charakterizuje dané rozdělení v rámci třídy rozdělení *exponenciálního typu*.

Předpokládejme, že skutečná hodnota dispersního parametru je φ_0 a skutečná hodnota vektoru regresních koeficientů je $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0k})^{\mathsf{T}}$. Věrohodností pozorovaných dat budeme rozumět výraz

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} g(Y_i|\mathbf{X}_i, \boldsymbol{\beta}, \varphi_0) = \prod_{i=1}^{n} \exp\left(\frac{Y_i\,\psi_{\mathrm{i}} - b(\psi_{\mathrm{i}})}{\varphi_0} + c(Y_i, \varphi_0)\right)$$

a logaritmickou věrohodností pozorovaných dat výraz

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log(g(Y_i|\mathbf{X}_i, \boldsymbol{\beta}, \varphi_0)) = \sum_{i=1}^{n} \left[\frac{Y_i\,\psi_{\mathrm{i}} - b(\psi_{\mathrm{i}})}{\varphi_0} + c(Y_i, \varphi_0)\right].$$

Označíme-li dále

$$w(\mu_i) = \frac{1}{V(\mu_i)[h'(\mu_i)]^2},$$

kde dle zavedení zobecněných lineárních modelů platí $h(\mu_i) = \mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}$ a $\mu_i = b'(\psi_{\mathrm{i}})$, můžeme předpis tzv. skórové funkce zapsat jako

$$\mathbf{U}(\boldsymbol{\beta}\,|Y_i) = \frac{1}{\varphi_0} w(\mu_i) h'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i.$$

Pro tzv. skórovou statistiku pak platí

$$\mathbf{U}_n(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial\, \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{\beta}\,|Y_i). \tag{1.2}$$

Maximálně věrohodný odhad $\hat{\boldsymbol{\beta}}_n$ vektoru regresních koeficientů $\boldsymbol{\beta}$ získáme jako řešení soustavy rovnic $\mathbf{U}_n(\boldsymbol{\beta}) = 0$, označované též jako *odhadovací rovnice*

$$\sum_{i=1}^{n} w(\hat{\mu}_i) h'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)\mathbf{X}_i = 0, \qquad \text{kde} \qquad \hat{\mu}_i = h^{-1}\Big(\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}_n\Big).$$

Předpokládejme, že matice $\mathsf{E}[w(\mu_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}]$ má plnou hodnost a všechny její prvky jsou konečné. Označme

$$\begin{aligned}
\mathbb{J}_i &= \left[w'(\mu_i) + w(\mu_i)\frac{h''(\mu_i)}{h'(\mu_i)}\right](Y_i - \mu_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}, \\
\mathbb{I}(\boldsymbol{\beta}\,|Y_i) &= \frac{1}{\varphi}\Big[w(\mu_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}} - \mathbb{J}_i\Big].
\end{aligned}$$

Pak pozorovanou Fisherovu informační matici můžeme zapsat jako

$$\mathbb{I}_n(\boldsymbol{\beta}\,|\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\boldsymbol{\beta}\,|Y_i).$$

Střední hodnotou $\mathbb{I}_n(\boldsymbol{\beta}\,|\mathbf{Y})$ je tzv. očekávaná Fisherova informační matice, jejíž předpisem je

$$\mathbb{I}(\boldsymbol{\beta}_0) = \mathsf{E}[\mathbb{I}_n(\boldsymbol{\beta}\,|\mathbf{Y})] = \mathsf{E}[\mathbb{I}(\boldsymbol{\beta}\,|Y_i)] = \frac{1}{\varphi_0}\,\mathsf{E}[w(\mu_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}]. \tag{1.3}$$

Uveďme ještě asymptotické vlastnosti maximálně věrohodných odhadů. Platí následující tvrzení, jejichž důkaz můžeme nalézt např. v knize Anděl (2011).

Věta 1. *i) Mají-li odhadovací rovnice* $\mathbf{U}_n(\boldsymbol{\beta}) = 0$ *jediné řešení, je maximálně věrohodný odhad* $\hat{\boldsymbol{\beta}}_n$ *konzistentním odhadem* $\boldsymbol{\beta}_0$*,*

ii)

$$\frac{1}{\sqrt{n}}\mathbf{U}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_k(\mathbf{0},\mathbb{I}(\boldsymbol{\beta}_0)),$$

iii)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_k(\mathbf{0},\mathbb{I}^{-1}(\boldsymbol{\beta}_0)).$$

Z právě uvedeného již snadno získáme testové statistiky hypotéz

$$\mathsf{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \qquad \mathsf{H}_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0\,. \tag{1.4}$$

Věta 2. *Buď* $\hat{\mathbb{I}}_n$ *konzistentní odhad očekávané Fisherovy informační matice. Pak za nulové hypotézy uvedené v* (1.4) *platí*

i) (Statistika Raova skórového testu)

$$R_n = \frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta}_0)^{\mathsf{T}}\hat{\mathbb{I}}_n^{-1}\mathbf{U}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \chi^2_k,$$

ii) (Statistika Waldova testu)

$$W_n = n\Big(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\Big)^{\mathsf{T}}\hat{\mathbb{I}}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \chi^2_k.$$

Důkaz této věty je rovněž uveden např. v knize Anděl (2011).

Řešení soustavy $\mathbf{U}_n(\boldsymbol{\beta}) = 0$ můžeme získat například pomocí tzv. *iterativního algoritmu vážených nejmenších čtverců* (*iterative weighted least squares*, IWLS), který je popsán například v článku Green (1984). V krátkosti shrneme kroky tohoto algoritmu aplikovaného na zobecněný lineární model, jak jsou uvedeny například v práci Sen a da Motta Singer (1993). Nejprve však připomeňme, že o maximálně věrohodném odhadu $\hat{\boldsymbol{\beta}}_n$ lze ukázat, že řeší soustavu rovnic

$$\hat{\boldsymbol{\beta}}_n = \Big(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X}\Big)^{-1}\Big(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\hat{\mathbf{Z}}\Big), \tag{1.5}$$

kde jsme použili matici $\hat{\mathbb{W}} = \mathrm{diag}\Big(w(\hat{\mu}_1),\ldots,w(\hat{\mu}_n)\Big)$ a vektor $\hat{\mathbf{Z}} = (\hat{Z}_1,\ldots,\hat{Z}_n)^{\mathsf{T}}$, pro jehož složky platí

$$\hat{Z}_i = \mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}} + (Y_i - \hat{\mu}_i)g'(\hat{\mu}_i).$$

Iterativní algoritmus vážených nejmenších čtverců pak spočívá v následujících krocích:

krok 0: Buď $j = 0$ a vezměme $\hat{\mu}_i^{(0)} = Y_i$. V případě problémů, např. pokud $Y_i = 0$ a linkovou funkcí je logaritmus, můžeme přičíst malou konstantu d_i.

krok 1: Zvýšíme j o 1 a spočítáme matici $\hat{\mathbb{W}}^{(j)}$ a vektor $\hat{\mathbf{Z}}^{(j)}$, kde

$$\hat{\mathbb{W}}^{(j)} = \mathrm{diag}\Big(w\Big(\hat{\mu}_1^{(j)}\Big),\ldots,w\Big(\hat{\mu}_n^{(j)}\Big)\Big),$$
$$\hat{Z}_i^{(j)} = \mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(j)} + (Y_i - \hat{\mu}_i^{(j)})g'(\hat{\mu}_i^{(j)}).$$

krok 2: Spočítáme nové přiblížení odhadu $\hat{\boldsymbol{\beta}}_n$ dle rovnice (1.5), tedy

$$\hat{\boldsymbol{\beta}}_n^{(j+1)} = \left(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}^{(j)}\mathbb{X}\right)^{-1}\left(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}^{(j)}\hat{\mathbf{Z}}^{(j)}\right),$$

a přepočítáme

$$\hat{\mu}_i^{(j+1)} = g^{-1}\left(\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}_n^{(j+1)}\right).$$

Kroky 1 a 2 opakujeme, dokud není relativní změna v odhadu $\hat{\boldsymbol{\beta}}_n^{(j)}$ menší než předem specifikovaná mez. Výsledný vektor je hledaným maximálně věrohodným odhadem, jak je uvedeno např. v práci Sen a da Motta Singer (1993).

Na závěr podkapitoly ukážeme, že modely, z nichž budeme při konstrukci ZMM vycházet, jsou skutečně zobecněnými lineárními modely a že tedy jsou příslušná podkladová rozdělení exponenciálního typu.

Poissonovský regresní model $(Y_i|\mathbf{X}_i \sim Poiss(\lambda_i))$

Hustotu $Y_i|\mathbf{X}_i$ můžeme zapsat jako

$$f(y;\lambda_i) = \frac{\mathrm{e}^{-\lambda_i}\,\lambda_i^y}{y!} = \exp(y\log(\lambda_i) - \lambda_i - \log(y!))$$

a tedy

$$\begin{aligned}
&\psi_{\mathrm{i}} = \log(\lambda_i), \qquad b(\psi_{\mathrm{i}}) = \mathrm{e}^{\psi_{\mathrm{i}}}, \qquad c(Y_i,\varphi) = -\log(Y_i!),\\
&\varphi = 1, \qquad \mu_i = b'(\psi_{\mathrm{i}}) = \mathrm{e}^{\psi_{\mathrm{i}}} = \lambda_i, \qquad V(\mu_i) = \mu_i.
\end{aligned}$$

V celé práci se omezíme na linkovou funkci

$$\log(\lambda_i) = \mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}.$$

Logaritmická věrohodnost, skórová statistika a pozorovaná Fisherova informační matice mají následující předpisy

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}[Y_i\log(\lambda_i) - \lambda_i - \log(Y_i!)]\\
\mathbf{U}_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(Y_i - \lambda_i)\mathbf{X}_i,\\
\mathbb{I}_n(\boldsymbol{\beta}\,|\mathbf{Y}) &= \frac{1}{n}\sum_{i=1}^{n}\lambda_i\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}.
\end{aligned}$$

Negativně binomický regresní model $(Y_i|\mathbf{X}_i \sim NB(\lambda_i,\alpha),\ \alpha > 0$ dané)

Tak jako v případě poissonovského regresního modelu zapišme hustotu $Y_i|\mathbf{X}_i$ ve tvaru, z něhož je zřejmé, že je exponenciálního typu, a sice

$$\begin{aligned}
f(y;\lambda_i) &= \frac{\Gamma\left(y+\frac{1}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{\alpha}\right)}(\alpha\lambda_i)^y(1+\alpha\lambda_i)^{-\left(y+\frac{1}{\alpha}\right)}\\
&= \exp\left[y\log(\alpha\lambda_i) - y\log(1+\alpha\lambda_i) + \frac{1}{\alpha}\log(1+\alpha\lambda_i) + c_\alpha(y,1)\right]\\
&= \exp\left[y\log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right) - \frac{1}{\alpha}\log(1+\alpha\lambda_i) + c_\alpha(y,1)\right]
\end{aligned}$$

a tedy

$$\psi_{\mathrm{i}} = \log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right), \quad b(\psi_{\mathrm{i}}) = -\frac{1}{\alpha}\log\left(1-\mathrm{e}^{\psi_{\mathrm{i}}}\right), \quad V(\mu_i) = \mu_i + \alpha\mu_i^2,$$
$$c_\alpha(Y_i,\varphi) = \log\left(\frac{\Gamma\left(y+\frac{1}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{\alpha}\right)}\right), \quad \varphi = 1, \quad \mu_i = b'(\psi_{\mathrm{i}}) = \frac{1}{\alpha}\frac{\mathrm{e}^{\psi_{\mathrm{i}}}}{1-\mathrm{e}^{\psi_{\mathrm{i}}}} = \lambda_i.$$

Opět se zde omezíme na jedinou volbu linkové funkce, a sice

$$\log(\lambda_i) = \mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}.$$

Logaritmická věrohodnost, skórová statistika a pozorovaná Fisherova informační matice mají následující předpisy

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[Y_i\log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right) - \frac{1}{\alpha}\log(1+\alpha\lambda_i) + c_\alpha(Y_i,1)\right],$$
$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{Y_i-\lambda_i}{1+\alpha\lambda_i}\mathbf{X}_i,$$
$$\mathbb{I}_n(\boldsymbol{\beta}\,|\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i(1+\alpha Y_i)}{(1+\alpha\lambda_i)^2}\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}.$$

Vše dostaneme dosazením do předpisů logaritmické věrohodnosti, skórové statistiky a Fisherovy informační matice, jež jsme obecně uvedli výše na straně 8.

Avšak jak jsme již předeslali, setkáváme se v literatuře s několika rozděleními, která bývají označována jako *negativně binomická*. Předpis hustoty dvou z nich, jenž dle článku Ridout et al. (2001) patří k nejrozšířenějším, získáme z předpisu

$$f(y;\lambda_i,\alpha) = \frac{\Gamma\left(y+\frac{\lambda_i^{2-c}}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{\lambda_i^{2-c}}{\alpha}\right)}(1+\alpha\lambda_i^{c-1})^{-\frac{\lambda_i^{2-c}}{\alpha}}\left(1+\frac{\lambda_i^{1-c}}{\alpha}\right)^{-y} \tag{1.6}$$

volbou parametru $c \in \{1,2\}$. Předpis hustoty výše popsané negativně binomické rozdělení, které bývá někdy označováno též jako *negativně binomické rozdělení typu 2* (NB2), získáme z (1.6) volbou $c = 2$. Dosazením $c = 1$ do (1.6) pak dostaneme předpis hustoty *negativně binomického rozdělení typu 1* (NB1). Toto rozdělení však už není rozdělením exponenciálního typu. O obou rozděleních je pojednáno v knize McCullagh a Nelder (1998, s. 199, 373). Kromě předpisů jejich rozdělení jsou popsány příklady náhodných veličin, které se řídí rozdělením NB1 či NB2.

Pokusme se nyní v krátkosti ukázat odlišnost obou rozdělení. Předpis rozdělení NB2 přepíšeme do následujícího tvaru

$$f(y;\lambda_i,\alpha) = \frac{\Gamma\left(y+\frac{1}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y}. \tag{1.7}$$

Zavedeme-li substituce

$$r = \frac{1}{\alpha} > 0, \qquad p = \frac{1}{1+\alpha\lambda} \in (0,1),$$

můžeme (1.7) dále přepsat.

$$f(y;p_i,r) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p_i^r (1-p_i)^y.$$

Nechť navíc platí $r \in \mathcal{N}_0$ a mějme posloupnost nezávislých pokusů, kde úspěch nastane s pravděpodobností p_i. Pak se náhodná veličina Y udávající počet neúspěchů před r-tým úspěchem řídí rozdělením $NB2(r,p_i)$.

V případě rozdělení NB1 můžeme vztah (1.6) po dosazení $c = 1$ a úpravách přepsat jako

$$f(y;\lambda_i,\alpha) = \frac{\Gamma\left(y + \frac{\lambda_i}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{\lambda_i}{\alpha}\right)} \left(\frac{1}{1+\alpha}\right)^{\frac{\lambda_i}{\alpha}} \left(\frac{\alpha}{1+\alpha}\right)^y. \tag{1.8}$$

Zvolme nyní substituce

$$r_i = \frac{\lambda_i}{\alpha}, \qquad p = \frac{1}{1+\alpha}.$$

Stejně jako v případě NB2 ihned vidíme, že $r_i > 0$ a $p \in (0,1)$. Předpis rozdělení NB1 pak můžeme přepsat jako

$$f(y;p,r_i) = \frac{\Gamma(y+r_i)}{\Gamma(y+1)\Gamma(r_i)} p^{r_i} (1-p)^y.$$

Za dodatečné podmínky $r_i \in \mathcal{N}$ můžeme i rozdělením $NB1(r_i,p)$ modelovat počet neúspěchů před r_i-tým úspěchem v sérii nezávislých pokusů, které skončí nezdarem s pravděpodobností $1-p$. Na rozdíl od rozdělení NB2 však nyní na regresorech může záviset počet úspěšných pokusů r_i, zatímco pravděpodobnost úspěchu p zůstává konstantní.

1.3 Kvazi-věrohodnost

V některých pracích věnovaných *negativně binomickým modelům s upraveným počtem nul*, například ve článcích Grogger a Carson (1991), Welsh et al. (1996) či Deng a Paul (2005), se autoři omezují na případ, kdy je podkladovým rozdělením rozdělení NB2 a parametr $\alpha > 0$ (resp. obdobný v jiné parametrizaci) je navíc pevně daný. Podkladové rozdělení je pak exponenciálního typu, což umožňuje poměrně snadnou aplikaci poznatků o ZMM. Nicméně fixace parametru značně omezuje praktické využití těchto regresních modelů. Je-li však tento předpoklad odstraněn, nespadá už negativně binomický regresní model do třídy *zobecněných lineárních modelů* a nemůžeme tak použít právě uvedené poznatky o maximálně věrohodných odhadech.

V literatuře jsme nenarazili na žádnou práci, která by se zabývala např. konstrukcí odhadů regresních koeficientů v případě *negativně binomických modelů s nadbytečnými nulami* založenými na rozdělení NB1. V kapitole 5 nicméně ukážeme, že je v jejich případě konstrukce poměrně snadná a to i tehdy, je-li parametr α neznámý. Využijeme přitom metodu *kvazi-věrohodnosti* (*quasi-likelihood*), představenou v práci Wedderburn (1974). Nyní tuto metodu v krátkosti popíšeme.

Mějme tedy k dispozici nezávislá pozorování $\left(Y_i, \mathbf{X}_i^\mathsf{T}\right)$, $i = 1,2,\ldots,n$. Opět označme střední hodnotu odezvy $\mu_i = \mathsf{E}[Y_i|\mathbf{X}_i]$ a lineární prediktor $\eta_i = \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}$, kde $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)^\mathsf{T}$ je vektor regresních koeficientů. Nechť dále platí:

1. Existuje ryze monotonní, dvakrát spojitě diferencovatelná funkce $h(.)$ splňující $h(\mu_i) = \eta_i$.

2. Buď $\varphi > 0$ dispersní parametr a $V(.)$ nezáporná, dvakrát spojitě diferencovatelná funkce, kterou budeme opět nazývat *varianční funkcí*. Pro rozptyl odezvy nechť platí $\mathrm{var}(Y_i|\mathbf{X}_i) = \varphi\, V(\mu_i)$.

Kvazi-věrohodnost pozorovaných dat pak definujeme předpisem

$$\mathcal{K}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{K}_i(\boldsymbol{\beta}), \qquad \text{kde} \qquad \mathcal{K}_i(\boldsymbol{\beta}) = \int_{Y_i}^{\mu_i} \frac{Y_i - u}{\varphi\, V(u)} \mathrm{d}u.$$

Vektor $\hat{\boldsymbol{\beta}}_n$, jenž maximalizuje *kvazi-věrohodnost* $\mathcal{K}(.)$, nazýváme *maximálně kvazi-věrohodným odhadem* $\boldsymbol{\beta}$.

Derivací výrazu $\mathcal{K}(\boldsymbol{\beta})$ podle vektorového parametru $\boldsymbol{\beta}$ dostaneme tzv. *kvazi-skórovou statistiku*, jejímž předpisem je

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}\,|Y_i),$$

přičemž pro jednotlivé sčítance platí

$$\mathbf{U}_i(\boldsymbol{\beta}\,|Y_i) = \frac{\partial \mathcal{K}_i(\boldsymbol{\beta})}{\partial\, \boldsymbol{\beta}} = \frac{1}{\varphi_0} w(\mu_i) h'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i.$$

Předpis *kvazi-skórové statistiky* tedy odpovídá předpisu skórové statistiky (1.2). Řešením soustavy rovnic $\mathbf{U}_n(\boldsymbol{\beta}) = 0$ dostaneme *maximálně kvazi-věrohodný odhad* $\hat{\boldsymbol{\beta}}_n$.

Označme $\boldsymbol{\beta}_0$ skutečnou hodnotu parametru $\boldsymbol{\beta}$ a buď $\mathbb{I}(\boldsymbol{\beta}_0)$ definován vztahem (1.3) uvedeným na straně 8. Platí následující tvrzení, jehož důkaz je uveden např. v práci McCullagh a Nelder (1998):

Věta 3. *i)* $E(\mathbf{U}_n(\boldsymbol{\beta}_0)) = \mathbf{0}$, $\quad var(\mathbf{U}_n(\boldsymbol{\beta}_0)) = \mathbb{I}(\boldsymbol{\beta}_0)$,

ii) Existuje posloupnost $\hat{\boldsymbol{\beta}}_n$ kvazi-věrohodných odhadů, pro niž platí

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathcal{P}} \boldsymbol{\beta}_0\,,$$

iii)

$$\frac{1}{\sqrt{n}} \mathbf{U}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_k(\mathbf{0}, \mathbb{I}(\boldsymbol{\beta}_0)),$$

iv)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_k(\mathbf{0}, \mathbb{I}^{-1}(\boldsymbol{\beta}_0)).$$

Stejně jako v případě zobecněných lineárních modelů můžeme konstruovat testové statistiky Waldova a Raova skórového testu.

Vidíme tedy, že pokud jsou jednotlivá pozorování nezávislá a rozptylová funkce je správně definována, platí teorie odhadů regresních koeficientů představená v předchozí podkapitole i v případech, kdy není rozdělení odezvy exponenciálního typu. O *kvazi-věrohodnosti* $\mathcal{K}(.)$ lze ukázat, že má mnohé vlastnosti logaritmické

věrohodnosti. Navíc pokud je rozdělení odezvy exponenciálního typu, je $\mathcal{K}(.)$ dokonce přímo logaritmickou věrohodností, přičemž důkaz je uveden např. v práci Wedderburn (1974).

Na závěr podkapitoly ukážeme, že negativně binomický regresní model založený na rozdělení NB1 ($Y_i|\mathbf{X}_i \sim \text{NB1}(\lambda_i,\alpha)$, $\alpha > 0$ známý či neznámý parametr) splňuje výše uvedené předpoklady. Připomeňme, že hustota rozdělení NB1 má dle (1.8) následující předpis

$$f(y;\lambda_i,\alpha) = \frac{\Gamma\left(y + \frac{\lambda_i}{\alpha}\right)}{\Gamma(y+1)\Gamma\left(\frac{\lambda_i}{\alpha}\right)}\left(\frac{1}{1+\alpha}\right)^{\frac{\lambda_i}{\alpha}}\left(\frac{\alpha}{1+\alpha}\right)^{y}..$$

Varianční funkcí tohoto rozdělení je

$$V(\mu_i) = (1+\alpha)\mu_i,$$

kde stejně jako v případě NB2 platí

$$\mu_i = \mathsf{E}[Y_i|\mathbf{X}_i] = \lambda_i.$$

Rovněž za linkovou funkci vezměme stejně jako výše v případě NB2 funkci logaritmus. Dostáváme tak:

1. Zvolená funkce $\log(\mu_i) = \mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}$ je zřejmě ryze monotonní a dvakrát spojitě diferencovatelná.

2. Dříve použitá varianční funkce obsahuje parametr α a nemůžeme ji proto použít (je sice nezáporná a diferencovatelná, její předpis ale není známý). Namísto toho proto uvažujme následující funkci

   $$V(\mu_i) = \mu_i.$$

   Položíme-li dále dispersní parametr $\varphi = 1 + \alpha$, bude zřejmě splněn předpoklad

   $$\text{var}(Y_i|\mathbf{X}_i) = \varphi\, V(\mu_i) = \mu_i(1+\alpha).$$

Maximálně kvazi-věrohodný odhad tak získáme řešením rovnice

$$\mathbf{U}_n(\boldsymbol{\beta}) = 0.$$

Vraťme se ještě krátce k rozdělení NB2. Připomeňme, že předpis varianční funkce tohoto rozdělení je $V(\mu_i) = \mu_i + \alpha\mu_i^2$. Nemůžeme ho tedy vyjádřit jako součin dispersního parametru a známé nezáporné diferencovatelné funkce μ_i. V případě rozdělení NB2 tedy nemůžeme použít metodu *kvazi-věrohodnosti* a jsme nuceni alternativní přístupy ke konstrukci maximálně věrohodných odhadů. V této práci se jimi ale zabývat nebudeme.

1.4 EM algoritmus

Nyní se věnujme *EM algoritmu* (*Expectation-Maximization algorithm*, EM), který byl představen v práci Dempster et al. (1977). Tato iterativní metoda se hojně užívá při hledání maximálně věrohodných odhadů nejen v kontextu *modelů s nadbytečnými nulami* či obecněji *směsných modelů* a v dalším ji budeme opakovaně používat. Poznamenejme, že se při jejím představování omezíme pouze na případ maximálně věrohodných odhadů, nicméně teorii lze aplikovat i na kvazi-věrohodné odhady, jak je ukázáno např. v článku Heyde a Morton (1996).

Předpokládejme, že máme k dispozici nezávislá, stejně rozdělená pozorování $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$ z rozdělení s hustotou $f(.;\boldsymbol{\theta})$. Neznámý parametr $\boldsymbol{\theta}$ chceme odhadnout metodou maximální věrohodnosti. Ne vždy však lze postupovat standardní cestou, přímý výpočet odhadu $\hat{\boldsymbol{\theta}}_n$ bývá v některých případech příliš obtížný. Někdy však lze předpokládat, že existují nepozorovaná data $\mathbf{Z} = (Z_1, \ldots, Z_n)^\mathsf{T}$, po jejichž zahrnutí se výpočet značně zjednoduší. Tak je tomu i v případě *modelů s upraveným počtem nul*. Jak uvidíme v dalších kapitolách, tyto latentní veličiny ponesou informaci o tom, zda bylo pozorování Y_i vygenerováno z *perfektního stavu* či nikoli. Pokračujme však v představování EM algoritmu.

Máme tedy k dispozici nezávislá, stejně rozdělená pozorování $(X_i, Z_i)^\mathsf{T}$, $i = 1, \ldots, n$, tzv. *úplná data*, se sdruženou hustotou $f_c(.;\boldsymbol{\theta})$. Obdobně jako věrohodnost a logaritmickou věrohodnost pozorovaných dat $\mathbf{X}$ definujme *věrohodnost* a *logaritmickou věrohodnost úplných dat* po řadě předpisy

$$\mathcal{L}_c(\boldsymbol{\theta}) = \mathcal{L}_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} f_c(X_i, Z_i; \boldsymbol{\theta}),$$
$$\ell_c(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} \log(f_c(X_i, Z_i; \boldsymbol{\theta})).$$

EM algoritmus pak postupuje následovně – předpokládejme, že z dosavadního průběhu algoritumu již máme k dispozici $\boldsymbol{\theta}^{(j)}$ (kde $j \in \mathcal{N}$) a hledáme další přiblížení $\boldsymbol{\theta}^{(j+1)}$.

E krok Spočítáme střední hodnotu logaritmické věrohodnosti pozorovaných dat vzhledem k rozdělení $\mathbf{Z}|\mathbf{X}$, přičemž neznámý parametr $\boldsymbol{\theta}$ nahradíme aktuálním přiblížení $\boldsymbol{\theta}^{(j)}$, a sice

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}\right) = \mathsf{E}_{\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(j)}}\left[\ell_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) \middle| \mathbf{X}, \boldsymbol{\theta}^{(j)}\right].$$

M krok Zvolíme další iteraci $\boldsymbol{\theta}^{(j+1)}$ odhadu $\hat{\boldsymbol{\theta}}_n$ tak, aby maximalizovala střední hodnotu spočtenou v předchozím kroku, tedy

$$\boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}\right).$$

Kroky E a M opakujeme do konvergence. Poznamenejme ještě, že v následujících kapitolách budeme předpokládat, že jsou veličiny Z_i alternativně rozdělené. Díky tomu se výpočty prováděné v jednotlivých krocích EM algoritmu značně zjednoduší.

Na závěr uvedeme tři důležitá tvrzení o konvergenci EM algoritmu, jejichž důkaz je možné nalézt například v práci McLachlan a Krishnan (2008).

Věta 4. *Nechť $\boldsymbol{\theta}^{(j)}$ je j-tá iterace EM algoritmu. Potom pro logaritmickou věrohodnost pozorovaných dat platí*

$$\ell\Big(\boldsymbol{\theta}^{(j+1)}\Big) \geq \ell\Big(\boldsymbol{\theta}^{(j)}\Big).$$

Věta 5. *Nechť je funkce $\mathcal{Q}\Big(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}\Big)$ spojitá v obou parametrech $\boldsymbol{\theta}$ a $\tilde{\boldsymbol{\theta}}$. Potom*

i) Všechny limitní body EM algoritmu jsou stacionárními body logaritmické věrohodnosti pozorovaných dat $\ell(.)$, tj. existuje-li derivace funkce $\ell(.)$ v těchto bodech, pak je jistě nulová.

ii) Posloupnost

$$\Big\{\ell\Big(\boldsymbol{\theta}^{(j)}\Big)\Big\}_{j=1}^{\infty}$$

konverguje k $\ell(\boldsymbol{\theta}^*)$, kde $\boldsymbol{\theta}^*$ je stacionárním bodem funkce $\ell(.)$.

Věta 6. *Nechť má logaritmická věrohodnost pozorovaných dat $\ell(.)$ jednoznačné maximum $\boldsymbol{\theta}^*$, které je zároveň jediným stacionárním bodem této funkce. Nechť je funkce $\mathcal{Q}\Big(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}\Big)$ spojitě diferencovatelná. Potom každá posloupnost $\Big\{\,\boldsymbol{\theta}^{(j)}\,\Big\}_{j=1}^{\infty}$ generovaná EM algoritmem konverguje k $\boldsymbol{\theta}^*$.*

Věta 4 zajišťuje monotonii posloupnosti generované EM algoritmem, věta 5 pak to, že tato posloupnost konverguje ke stacionárnímu bodu pozorované věrohodnosti $\ell(.)$. Konečně věta 6 zaručuje, že při splnění dodatečných předpokladů konverguje každá posloupnost generovaná EM algoritmem k bodu, jenž je hledaným maximálně věrohodným odhadem parametrů modelu.

1.5 Struktura dat a značení

Na závěr kapitoly se věnujme předpokládané struktuře dat, kterou budeme uvažovat v kapitolách věnovaných *modelům s nadbytečnými nulami* a také v kapitolách zabývajících se *hradbovými modely*. Tento předpoklad označme písmenem (A), abychom na něj v dalším mohli snadno odkazovat.

Předpoklad (A):
Mějme k dispozici nezávislá, stejně rozdělená pozorování $(Y_i,\mathbf{X}_i^{\mathsf{T}})$, $i = 1,\ldots,n$, kde Y_i je odezva a $\mathbf{X}_i = (X_{i1},\ldots,X_{ik})^{\mathsf{T}}$ vektor regresorů. O rozdělení $Y_i|\mathbf{X}_i$ budeme předpokládat, že je směsí dvou rozdělení, a sice rozdělení $\mathcal{G}$ (*podkladové rozdělení*) s hustotou $g(.)$ a rozdělení degenerovaného v nule (*perfektní stav*). Proporce p_i rozdělení degenerovaného v nule ve výsledné směsi přitom může záviset na regresorech.

Předpokládejme dále, že existuje vektor latentních veličin $\mathbf{Z} = (Z_1,\ldots,Z_n)^{\mathsf{T}}$, kde veličina Z_i indikuje, zda odezva Y_i pochází z *perfektního stavu* ($Z_i = 1$) či z *podkladového rozdělení* ($Z_i = 0$). Předpokládáme, že jsou veličiny Z_i nezávislé a alternativně rozdělené s parametrem p_i.

Dále označme μ_i střední hodnotu odezvy v *podkladovém rozdělení* a předpokládejme, že existují funkce dvě $h_1(.)$ a $h_2(.)$ a dva vektory regresních koeficientů $\boldsymbol{\xi}_1 = (\xi_{11},\ldots,\xi_{1k})^{\mathsf{T}}$ a $\boldsymbol{\xi}_2 = (\xi_{21},\ldots,\xi_{2k})^{\mathsf{T}}$, pro něž platí

$$\mu_i = h_1^{-1}\Big(\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\xi}_1\Big), \qquad p_i = h_2^{-1}\Big(\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\xi}_2\Big).$$

Ve shodě s pracemi věnovanými *modelům pro data s upraveným počtem nul*, zmiňme například Lambert (1992) či Welsh et al. (1996), budeme funkce $h_1(.)$ a $h_2(.)$ nazývat *linkovými funkcemi*.

Ze složek vektoru regresorů $\mathbf{X}_i$ dále vytvoříme vektory $\mathbf{V}_i = (V_{i1}, \ldots, V_{iv})^\mathsf{T}$ a $\mathbf{W}_i = (W_{i1}, \ldots, W_{iw})^\mathsf{T}$, tedy

$$\begin{aligned} \underset{v\in\{1,\ldots,k\}}{\exists}\ \underset{j\in\{1,\ldots,v\}}{\forall}\ \underset{l_j\in\{1,\ldots,k\}}{\exists}\ \underset{i\in\mathcal{N}}{\forall} \quad V_{ij} &= X_{il_j}, \\ \underset{w\in\{1,\ldots,k\}}{\exists}\ \underset{j\in\{1,\ldots,w\}}{\forall}\ \underset{l_j\in\{1,\ldots,k\}}{\exists}\ \underset{i\in\mathcal{N}}{\forall} \quad W_{ij} &= X_{il_j}, \end{aligned}$$

přičemž požadujeme, aby platilo

$$\mu_i = h_1^{-1}\big(\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}\big), \qquad p_i = h_2^{-1}\big(\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}\big),$$

kde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_v)^\mathsf{T}$ a $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_w)^\mathsf{T}$ jsou vektory regresních koeficientů. Některé či dokonce všechny složky vektoru $\mathbf{X}_i$ se mohou vyskytovat v obou nově definovaných vektorech.

Poznamenejme, že zřejmě platí

$$Z_i|\mathbf{W}_i \sim Alt(p_i), \qquad \text{a tedy} \qquad \mathsf{E}[Z_i|\mathbf{W}_i] = p_i.$$

V mnoha pracích zabývajících se *modely s nadbytečnými nulami* (např. v článcích Lambert (1992), Ridout et al. (2001) či Hall (2000)) je za funkci $h_2(.)$ zvolna funkce logit.. Tato poměrně přirozená volba odpovídá logistickému regresnímu modelu, v němž je $\mathbf{W}_i$ vektor regresorů a odezvou Z_i, indikátor, že veličina Y_i pochází z *perfektního stavu*.

Označme ještě $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\mathsf{T}$ matici všech regresorů a obdobně také $\mathbb{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_n)^\mathsf{T}$ a $\mathbb{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)^\mathsf{T}$ matice vybraných regresorů.

2. Modely bez nulové odezvy

První pokusy vypořádat se s nadměrným množstvím nul ve vysvětlované veličině vedly přes *modely bez nulové odezvy* (*zero truncated models*, ZTM). Z odezvy stačilo oddělit nenulovou část a tu pak modelovat zvlášť pomocí vhodného modelu ZTM. S tímto postupem se setkáváme například v pracích Rider (1953) či Cohen (1960). Někteří autoři se posléze pokoušeli použít ZTM při budování *modelů s nadbytečnými nulami*. Vzniklé modely budeme v této práci nazývat *hradbové modely* (*hurdle models*, HM) a více se jim budeme věnovat v kapitole 6.

Zde si nejprve krátce představíme myšlenku modelů ZTM. Tak jako v případě dalších tříd *modelů s upraveným počtem nul* se zaměříme na model, v němž je *podkladové rozdělení* obecně diskrétní rozdělení exponenciálního typu. Pak se budeme věnovat speciálním případům modelů ZTM, a sice tzv. *poissonovskému modelu bez nulové odezvy* (*zero truncated Poisson model*, ZTP) v podkapitole 2.3 a *negativně binomickému modelu bez nulové odezvy* (*zero truncatedy negative binomial model*, ZTNB) v podkapitole 2.4, přičemž se omezíme na modely vycházející z rozdělení NB2 se známým parametrem α. Právě těmto dvěma případům se věnovala celá řada autorů, např. v článku Grogger a Carson (1991) či v již zmiňované práci Cohen (1960).

Pro každý z modelů uvedeme nejprve předpis rozdělení odezvy, její střední hodnoty a rozptylu. Dále se budeme zabývat konstrukcí maximálně věrohodných odhadů regresních koeficientů. Uvedených poznatků později použijeme v kapitole 6 při konstrukci příslušných *hradbových modelů*.

2.1 Definice modelu

Vyjděme z definice zobecněného lineárního modelu, který jsme představili v úvodní kapitole na straně 6. Předpokládáme tedy, že pozorujeme nezávislá a stejně rozdělená data $\left(Y_i^*, \mathbf{X}_i^{*\mathsf{T}}\right)$, $i = 1, 2, \ldots, n$, kde Y_i^* je modelovaná veličina a $\mathbf{X}_i^* = (X_{i1}^*, \ldots, X_{ik}^*)^{\mathsf{T}}$ je vektor regresorů. Rozdělení odezvy $Y_i^*|\mathbf{X}_i^*$ označme $\mathcal{G}$ a předpokládejme o něm, že je diskrétním rozdělením exponenciálního typu. V dalším tedy *hustotou* budeme rozumět hustotu vůči čítací míře. Označme dále $\mu_i^* = \mathsf{E}[Y_i^*|\mathbf{X}_i^*]$ a $h(.)$ vhodnou linkovou funkci.

Hustotu rozdělení $\mathcal{G}$ označíme $g(.)$ a budeme o ní dále předpokládat, že splňuje

$$g(0|\mathbf{X}_i^*) = \mathsf{P}(Y_i^* = 0|\mathbf{X}_i^*) > 0. \tag{2.1}$$

Obdobně jako v následujících kapitolách věnovaných *modelům s nadbytečnými nulami* a *hradbovým modelům*, budeme tuto hustotu nazývat *podkladovou hustotou* a rozdělení $\mathcal{G}$ *podkladovým rozdělením*.

O rozdělení $\mathcal{G}$ budeme dále předpokládat, že jeho nosičem jsou přirozená čísla doplněná o nulu, resp. jejich podmnožina, která však dle (2.1) nulu obsahuje. Tento předpoklad je při běžných aplikacích ZTM splněn a případné zobecnění je vesměs přímočaré.

Předpokládejme, že pro nějaké $m \in \{1, \ldots, n\}$ jsou pozorování $Y_{i_1}^*, \ldots, Y_{i_m}^*$ nenulová a zbylá pozorování $Y_{i_{m+1}}^*, \ldots, Y_{i_n}^*$ jsou vesměs nulová. Proveďme následující přeznačení

$$Y_{i_k}^* = Y_k, \qquad \mathbf{X}_{i_k}^* = \mathbf{X}_k \quad \text{a} \quad \mu_{i_k}^* = \mu_k, \qquad k \in \{1, \ldots, n\}.$$

Případné další parametry závislé na regresorech přeznačíme obdobně.

Jelikož je naším cílem modelovat data bez nulové odezvy, omezíme se v dalším pouze na pozorování $\left(Y_i, \mathbf{X}_i^\intercal\right)$, $i = 1, 2, \ldots, m$. Hustotu $g_T(.)$ takto oříznuté odezvy pak můžeme dle věty o podmíněné hustotě (viz např. Anděl (2011)) zapsat jako

$$g_T(y|\mathbf{X}_k) = g\Big(y\Big|\mathbf{X}^*_{i_k}, Y^*_{i_k} > 0\Big), \qquad y \neq 0,$$

přičemž uvedený vztah můžeme dále rozepsat

$$g_T(y|\mathbf{X}_k) = \frac{g\Big(y|\mathbf{X}^*_{i_k}\Big)}{\mathsf{P}\Big(Y^*_{i_k} > 0\Big|\mathbf{X}^*_{i_k}\Big)}, \qquad y \neq 0. \tag{2.2}$$

Uplatnili jsme tedy požadavek (2.1) na hustotu $g(.)$. Poznamenejme, že hustota $g_T(.|\mathbf{X}_k)$ již obecně není hustotou exponenciálního typu.

2.2 Obecný model bez nulové odezvy

Zabývejme se *modely bez nulové odezvy*, v nichž je rozdělení odezvy odvozeno z nějakého diskrétního rozdělení exponenciálního typu. V literatuře jsme se s obecným přístupem k ZTM nesetkali, většina autorů se zabývá *poissonovským* či jejich jiným speciálním případem. Jsme proto nuceni níže uvedené poznatky sami odvodit. Budeme přitom vycházet právě z článků věnovaných speciálním případům ZTM, např. z práce Grogger a Carson (1991).

Z předpisu hustoty rozdělení exponenciálního typu (viz (1.1) na straně 7) a ze vztahu (2.2) snadno získáme předpis hustoty rozdělení odezvy v ZTM, a sice

$$\begin{aligned} g_T(j;\psi_k,\varphi) &= \mathsf{P}(Y_k = j|\mathbf{X}_k) \\ &= \mathsf{P}\Big(Y^*_{i_k} = j|\mathbf{X}^*_{i_k}, Y^*_{i_k} > 0\Big) = \frac{g\Big(j|\psi^*_{i_k},\varphi\Big)}{1 - \mathsf{P}\Big(Y^*_{i_k} = 0|\mathbf{X}^*_{i_k}\Big)} \\ &= \frac{g\Big(j|\psi^*_{i_k},\varphi\Big)}{1 - g\Big(0;\psi^*_{i_k},\varphi\Big)} = \frac{g(j|\psi_k,\varphi)}{1 - g(0;\psi_k,\varphi)} \\ &= \frac{\exp\left(\dfrac{j\psi_k}{\varphi} + c(j,\varphi) - c(0,\varphi)\right)}{\exp\left(\dfrac{b(\psi_k)}{\varphi} - c(0,\varphi)\right) - 1}, \qquad j \in \mathcal{N}. \end{aligned} \tag{2.3}$$

Připomeňme, že použitý vztah $\psi^*_{i_k} = \psi_k$ vychází z výše zvoleného přeznačení.

Odvoďme ještě předpis střední hodnoty a rozptylu odezvy v ZTM. Postupnými úpravami dostaneme

$$\begin{aligned}\mathsf{E}[Y_k|\mathbf{X}_k] &= \sum_{j\in\mathcal{N}} j\cdot g_T(j;\psi_k,\varphi) = \sum_{j\in\mathcal{N}} \frac{j\cdot g\Big(j;\psi^*_{i_k},\varphi\Big)}{1-g\Big(0;\psi^*_{i_k},\varphi\Big)}\\ &= \frac{1}{1-g\Big(0;\psi^*_{i_k},\varphi\Big)}\cdot\sum_{j\in\mathcal{N}_0} j\cdot g\Big(j;\psi^*_{i_k},\varphi\Big)\\ &= \frac{\mathsf{E}\Big[Y^*_{i_k}|\mathbf{X}^*_{i_k}\Big]}{1-g\Big(0;\psi^*_{i_k},\varphi\Big)}.\end{aligned}$$

Obdobně získáme i druhý moment tohoto rozdělení

$$\mathsf{E}\Big[Y_k^2\Big|\mathbf{X}_k\Big] = \frac{\mathsf{E}\Big[(Y^*_{i_k})^2\Big|\mathbf{X}^*_{i_k}\Big]}{1-g\Big(0;\psi^*_{i_k},\varphi\Big)}$$

a z obou momentů pak již snadno spočítáme předpis rozptylu odezvy, a sice

$$\begin{aligned}\mathrm{var}[Y_k|\mathbf{X}_k] &= \mathsf{E}\Big[Y_k^2\Big|\mathbf{X}_k\Big] - \Big(\mathsf{E}[Y_k|\mathbf{X}_k]\Big)^2\\ &= \frac{\mathsf{E}\Big[(Y^*_{i_k})^2\Big|\mathbf{X}^*_{i_k}\Big] - g\Big(0;\psi^*_{i_k},\varphi\Big)\,\mathsf{E}\Big[(Y^*_{i_k})^2\Big|\mathbf{X}^*_{i_k}\Big] - \Big(\,\mathsf{E}\Big[Y^*_{i_k}|\mathbf{X}^*_{i_k}\Big]\Big)^2}{\Big(1-g\Big(0;\psi^*_{i_k},\varphi\Big)\Big)^2}\\ &= \frac{\mathrm{var}\Big[Y^*_{i_k}|\mathbf{X}^*_{i_k}\Big] - g\Big(0;\psi^*_{i_k},\varphi\Big)\,\mathsf{E}\Big[(Y^*_{i_k})^2\Big|\mathbf{X}^*_{i_k}\Big]}{\Big(1-g\Big(0;\psi^*_{i_k},\varphi\Big)\Big)^2}.\end{aligned}$$

Nyní pokračujme v odvozování maximálně věrohodných odhadů regresních koeficientů v ZTM. V dalším tedy budeme pracovat pouze s daty bez nulové odezvy, a sice $\Big(Y_i,\mathbf{X}_i^{\mathsf{T}}\Big)$, $i=1,2,\ldots,m$. Z předpisu (2.3) dostaneme standardním způsobem logaritmickou věrohodnost pozorovaných dat

$$\begin{aligned}\ell_m(\boldsymbol{\beta}) = &\sum_{i=1}^{m}\left[\frac{Y_i\,\psi_{\mathrm{i}}}{\varphi} - \log\left(\exp\left(\frac{b(\psi_{\mathrm{i}})}{\varphi} - c(0,\varphi)\right) - 1\right)\right]\\ &+\sum_{i=1}^{m} c(Y_i,\varphi) - m\cdot c(0,\varphi).\end{aligned}$$

Derivací výrazu $\ell_m(\boldsymbol{\beta})$ získáme skórovou statistiku $\mathbf{U}_m(\boldsymbol{\beta})$. Při jejím výpočtu můžeme využít tzv. *řetízkového pravidla*, a sice

$$\mathbf{U}_m(\boldsymbol{\beta}) = \frac{\partial\ell_m(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}} = \frac{\partial\ell_m(\boldsymbol{\beta})}{\partial\,\psi_{\mathrm{i}}}\cdot\frac{\partial\,\psi_{\mathrm{i}}}{\partial\mu_i}\cdot\frac{\partial\mu_i}{\partial\eta_i}\cdot\frac{\partial\eta_i}{\partial\,\boldsymbol{\beta}}.$$

Po úpravách dostaneme

$$\mathbf{U}_m(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^{m}\left[\frac{Y_i - \frac{\mu_i}{1-g(0;\psi_{\mathrm{i}},\varphi)}}{V(\mu_i)h'(\mu_i)}\right]\mathbf{X}_i. \tag{2.4}$$

Připomeňme, že $h(.)$ je linková funkce, μ_i střední hodnota odezvy, $V(\mu_i)$ varianční funkce a η_i lineární prediktor v podkladovém rozdělení $\mathcal{G}$.

Položíme-li $\mathbf{U}_m(\boldsymbol{\beta}) = 0$, dostaneme odhadovací rovnici, jejímž řešením je hledaný maximálně věrohodný odhad vektoru regresních koeficientů $\boldsymbol{\beta}$.

2.3 Poissonovský model bez nulové odezvy

Poissonovský model bez nulové odezvy (*zero truncated Poisson model*, ZTP) je v literatuře dobře podchycen, zmiňme například práce Grogger a Carson (1991), Singh (1978) či Rider (1953). Jak již bylo předesláno, modely ZTP představují speciální případ modelů ZTM, přičemž za podkladové rozdělení je zvoleno rozdělení Poissonovo. Postup konstrukce odhadů je tak obdobou výše uvedeného.

Připomeňme, že předpis podkladové hustoty jsme uvedli na straně 10. Ze znalosti podkladového rozdělení pak ihned plyne

$$\mathsf{P}(Y_i^* = 0|\mathbf{X}_i^*) = \exp(-\lambda_i^*).$$

Dosadíme-li do rovnice (2.2), dostaneme předpis hustoty rozdělení odezvy v ZTP, a sice

$$\begin{aligned} g_T(j|\lambda_i^*) &= \mathsf{P}(Y_i^* = j|\mathbf{X}_i^*, Y_i^* > 0) \\ &= \frac{1}{1-\exp\left(-\lambda_i^*\right)} \cdot \frac{\exp\left(-\lambda_i^*\right)(\lambda_i^*)^j}{j!} \\ &= \frac{(\lambda_i^*)^j}{(\exp(\lambda_i^*)-1)j!}, \qquad\qquad j \in \mathcal{N}. \end{aligned}$$

Linkovou funkcí udávající vztah mezi vektorem regresorů a λ_i^*, střední hodnotou odezvy v podkladovém rozdělení, buď logaritmus, tedy

$$\lambda_i^* = \log(\mathbf{X}_i^{*\mathsf{T}}\boldsymbol{\beta}).$$

Připomeňme jen, že logaritmus je kanonickou linkovou funkcí *poissonovských regresních modelů*.

V dalším se opět omezíme na nenulová pozorování. Budeme tak uvažovat pozorování $\left(Y_i, \mathbf{X}_i^{\mathsf{T}}\right)$, $i = 1, 2, \ldots, m$ pro vhodné $m \in \{1, \ldots, n\}$ a to po přečíslování představeném v podkapitole 2.1. Než se budeme věnovat odhadům regresních koeficientů, zmiňme předpis střední hodnoty a rozptylu odezvy v ZTP, jak jsou uvedeny např. v práci Grogger a Carson (1991). Jejich předpisy jsou

$$\begin{aligned} \mathsf{E}[Y_i|\mathbf{X}_i] &= \frac{\lambda_i}{1-\exp(-\lambda_i)}, \\ \mathrm{var}[Y_i|\mathbf{X}_i] &= \mathsf{E}[Y_i|\mathbf{X}_i]\Big(1-\exp(-\lambda_i)\,\mathsf{E}[Y_i|\mathbf{X}_i]\Big). \end{aligned}$$

Logaritmickou věrohodností pozorovaných dat v modelech ZTP je funkce

$$\ell_m(\boldsymbol{\beta}) = \sum_{i=1}^{m}[Y_i \log(\lambda_i) - \log(\exp(\lambda_i) - 1) - \log(Y_i!)],$$

její předpis snadno dostaneme z předpisu hustoty $g_T(.)$. Derivováním $\ell_m(\boldsymbol{\beta})$ a dalšími úpravami dostaneme odhadovací rovnici

$$\mathbf{U}_m(\boldsymbol{\beta}) = \sum_{i=1}^{m}\left[Y_i - \frac{\lambda_i}{1-\mathrm{e}^{-\lambda_i}}\right]\mathbf{X}_i = 0,$$

jejímž řešením získáme $\hat{\boldsymbol{\beta}}_m$, maximálně věrohodný odhad regresních koeficientů $\boldsymbol{\beta}$. V práci Grogger a Carson (1991) je k řešení této odhadovací rovnice použit Newtonův-Raphsonův algoritmus.

2.4 Negativně binomický model bez nulové odezvy

Negativně binomickými modely bez nulové odezvy (*zero truncated negative binomial models*, ZTNB) se zabývala řada autorů – zmiňme například práci Welsh et al. (1996), kde jsou ZTNB použity k modelování výskytů živočichů na určitém území a tento přístup je srovnáván s dalšími modely, či článek Lee et al. (2003), kde jsou ZTNB využity při modelování lékařských dat.

Většina prací věnovaných modelům ZTNB přitom vychází z rozdělení NB2 (v dalším tyto modely budeme označovat též ZTNB2). Již v úvodní kapitole jsme ukázali, že je-li parametr α pevně daný, je rozdělení NB2 exponenciálního typu. Konstrukce odhadů regresních koeficientů je tak poměrně snadná, jedná se o přímou aplikaci výsledků podkapitoly 2.2 věnované ZTM obecně. V krátkosti se budeme tomuto speciálnímu případu věnovat v následující podkapitole.

Tak jako u dalších tříd modelů ZMM založených na negativně binomických rozděleních, ani v případě ZTM jsme se v literatuře příliš nesetkali s modely, v nichž by bylo podkladovým rozdělením rozdělení NB1 (v dalším jej označujme též ZTNB1). Není divu – při pohledu na předpis hustoty (viz např. rovnice (1.6) na straně 11) vidíme, že se vektor regresorů $\boldsymbol{\beta}$ vyskytuje v argumentu funkce $\Gamma(.)$. Je tedy patrné, že konstruovat maximálně věrohodné odhady je v tomto případě dosti nepříjemné.

S obdobnou situací se setkáme i v kapitole 5 v případě *negativně binomických modelů s nadbytečnými nulami* (*zero inflated negative binomial models*, ZINB). V části 5.2.2 však navrhneme postup, s jehož pomocí je možné získat odhady regresních koeficientů i v případě modelů ZINB vycházejících z rozdělení NB1. Zdá se, že by bylo možné postupovat obdobně i v případě modelů ZTNB1. Pro omezený rozsah této práce se zde tímto problémem zabývat nebudeme, nicméně jedná se o zajímavé téma hodné další pozornosti.

2.5 Modely ZTNB2 se známým parametrem α

Jelikož v podkladovém rozdělení NB2 platí

$$\mathsf{P}(Y_i^* = 0|\mathbf{X}_i^*) = g(0;\lambda_i^*) = \left(\frac{1}{1+\alpha\lambda_i^*}\right)^{\frac{1}{\alpha}},$$

dle postupu uvedeného v podkapitole 2.1 či přímo dosazením do (2.3) ze strany 19 dostaneme předpis rozdělení *NB2 bez nulové odezvy*, a sice

$$g_T(j|\lambda_k) = \frac{\exp\left[j\log\left(\frac{\alpha\lambda_k}{1+\alpha\lambda_k}\right) + c_\alpha(j,1)\right]}{\exp\left[\frac{1}{\alpha}\log(1+\alpha\lambda_k)\right]-1}, \qquad j \in \mathcal{N}, \tag{2.5}$$

kde jsme stejně jako v úvodní kapitole na straně 10 označili

$$c_\alpha(j,1) = \log\left(\frac{\Gamma\left(j+\frac{1}{\alpha}\right)}{\Gamma(j+1)\Gamma\left(\frac{1}{\alpha}\right)}\right).$$

Linkovou funkcí buď stejně jako v případě ZTP logaritmus, tedy

$$\lambda_k = \log(\mathbf{X}_k{}^{\mathsf{T}}\boldsymbol{\beta}).$$

Předpisy střední hodnoty a rozptylu odezvy, uvedené například v článku Grogger a Carson (1991), jsou

$$\begin{aligned}
\mathsf{E}[Y_k|\mathbf{X}_k] &= \frac{\lambda_k}{1-g(0;\lambda_k)} = \frac{\lambda_k}{1-(1+\alpha\lambda_k)^{-\frac{1}{\alpha}}}, \\
\operatorname{var}[Y_k|\mathbf{X}_k] &= \frac{\mathsf{E}[Y_k|\mathbf{X}_k]}{g(0;\lambda_k)^{\alpha}}\Big(1-g(0;\lambda_k)^{1+\alpha}\cdot\mathsf{E}[Y_k|\mathbf{X}_k]\Big) \\
&= \frac{\lambda_k\Big[1+\alpha\lambda_k-(1+\alpha\lambda_k)^{-\frac{1}{\alpha}}[1+(\alpha+1)\lambda_k]\Big]}{\Big[1-(1+\alpha\lambda_k)^{-\frac{1}{\alpha}}\Big]^2}.
\end{aligned}$$

Z rovnice (2.5) dostaneme logaritmickou věrohodnost pozorovaných dat

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}\left[Y_i\log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right) - \log\left(\exp\left(\frac{1}{\alpha}\log(1+\alpha\lambda_i)\right)-1\right)\right] + \sum_{i=1}^{m} c_\alpha(Y_i,1).$$

Derivováním $\ell(\boldsymbol{\beta})$ dle $\boldsymbol{\beta}$ anebo přímo dosazením výrazů ze strany 10 do (2.4) dostaneme odhadovací rovnici

$$\mathbf{U}_m(\boldsymbol{\beta}) = \sum_{i=1}^{m}\left[\frac{Y_i - \frac{\lambda_i}{1-g(0;\lambda_i)}}{1+\alpha\lambda_i}\right]\mathbf{X}_i,$$

jejímž řešením získáme kýžený maximálně věrohodný odhad regresních koeficientů $\boldsymbol{\beta}$.

3. Poissonovské modely s nadbytečnými nulami

Nyní se konečně začneme zabývat *modely s nadbytečnými nulami* (*zero inflated models*, ZIM). Na rozdíl od kapitoly 2, v níž jsme nejprve představili obecnou myšlenku *modelů bez nulové odezvy* a až posléze jsme se věnovali jejich speciálním případům, nyní budeme postupovat obráceně. V této kapitole popíšeme vlastnosti *poissonovských modelů s nadbytečnými nulami* (*zero inflated Poisson models*, ZIP), které byly poprvé představeny v práci Lambert (1992). *Modelům s nadbytečnými nulami* se poněkud obecněji budeme věnovat v následující kapitole.

Zde nejprve na motivačním příkladě demonstrujeme použití modelů ZIP. Následně modely ZIP řádně definujeme a dále se budeme zabývat maximálně věrohodnými odhady regresních koeficientů v této třídě modelů. V závěru kapitoly ještě zmíníme alternativní definici *poissonovského modelu s nadbytečnými nulami*, tzv. ZIP(τ), jenž byl rovněž představen v článku Lambert (1992).

3.1 Motivační příklad

V praxi se často setkáváme s úkolem modelovat počty událostí či jiné celočíselné veličiny. V takovém případě standardně používáme modely poissonovské regrese. Připomeňme, že jsme se poissonovskému regresnímu modelu krátce věnovali v úvodní kapitole na straně 10. Více se touto třídou modelů se zabývá např. práce McCullagh a Nelder (1998) a další.

V některých aplikacích však narážíme na problém, že data obsahují příliš mnoho nul na to, aby pocházela z Poissonova rozdělení. Uvažujme následující motivační příklad, který je zmíněn v článku Lambert (1992). Máme za úkol modelovat počty vad vzniklých při strojovém sváření v laboratořích Bell. Počet defektů přitom závisí na tom, zda se stroj nacházel v *perfektním* či neperfektním stavu. Byl-li seřízen správně, byly chyby sváření extrémně vzácné, v opačném případě k chybám s jistou pravděpodobností docházelo. V práci Lambert (1992) je nejprve ukázáno, že taková data nelze dobře modelovat běžně používanou poissonovskou regresí, a je proto navržena alternativa, tzv. *poissonovský model s nadbytečnými nulami* (ZIP).

Využití těchto modelů je však mnohem širší – byly publikovány články, v nichž byly modely ZIP použity mj. k modelování výskytu živočichů ohroženého druhu (Welsh et al. (1996)), ukazatele zdravotní kondice chrupu dětské populace (tzv. DMFT index; Dietz a Böhning (2000)), počtu dopravních nehod na sledovaných silničních úsecích (Shankar et al. (1997)) atp.

3.2 Definice modelu

Nechť je splněn předpoklad (A) ze strany 16, přičemž podkladové rozdělení $\mathcal{G}$ nechť je rozdělení Poissonovo s hustotou (vůči čítací míře)

$$g(j;\lambda_i) = \frac{\lambda_i^j \,\mathrm{e}^{-\lambda_i}}{j!}, \qquad j \in \mathcal{N}_0.$$

Linkové funkce $h_1(.)$ a $h_2(.)$ udávající vliv regresorů na odezvu Y_i a veličinu Z_i nechť jsou

$$\log(\lambda_i) = \mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}, \qquad \mathrm{logit}(p_i) = \mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}, \tag{3.1}$$

kde $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_v)^{\mathsf{T}}$ a $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_w)^{\mathsf{T}}$ jsou vektory regresních koeficientů. Je samozřejmě možné zvolit i jiné linkové funkce; protože se věnujeme modelu ZIP tak, jak byl představen v článku Lambert (1992), další volby pomineme.

Shrneme-li předpoklady, vidíme, že má odezva Y_i s pravděpodobností p_i degenerované rozdělení v 0 (*perfektní stav*) a s pravděpodobností $1-p_i$ má Poissonovo rozdělení s parametrem λ_i (*podkladové rozdělení*). Snadno tedy můžeme zapsat předpis rozdělení odezvy, a sice

$$\begin{aligned} \mathsf{P}(Y_i = 0|\mathbf{X}_i) &= f(0|\mathbf{X}_i) = p_i + (1-p_i)\,\mathrm{e}^{-\lambda_i}\,, \\ \mathsf{P}(Y_i = j|\mathbf{X}_i) &= f(j|\mathbf{X}_i) = (1-p_i)\frac{\mathrm{e}^{-\lambda_i}\,\lambda_i^j}{j!}, \qquad j \in \mathcal{N}. \end{aligned} \tag{3.2}$$

Předpis střední hodnoty a rozptylu odezvy Y_i je uvedený například v článku Chin a Quddus (2003), a sice

$$\begin{aligned} \mathsf{E}[Y_i|\mathbf{X}_i] &= (1-p_i)\,\lambda_i, \\ \mathrm{var}(Y_i|\mathbf{X}_i) &= \mathsf{E}[Y_i|\mathbf{X}_i]\left(1 + \frac{p_i}{1-p_i}\,\mathsf{E}[Y_i|\mathbf{X}_i]\right), \end{aligned}$$

není však obtížné oba výrazy z právě uvedeného dopočítat.

3.3 Maximálně věrohodné odhady v ZIP

Pokusme se nyní odhadnout všech $v+w$ regresních koeficientů metodou maximální věrohodnosti. V porovnání se standardním modelem poissonovské regrese (viz též strana 10), kde je třeba odhadnout k regresních koeficientů, zůstal počet odhadovaných parametrů stejný nebo se dokonce zvětšil (to pokud existují regresory X_{il}, $l = 1,\ldots,k$, vyskytující se v obou nově definovaných vektorech regresorů $\mathbf{V}_i$ i $\mathbf{W}_i$). V práci Lambert (1992) je proto navržena redukce počtu regresorů tím, že se předpokládá funkční závislost mezi parametry λ_i a p_i. Krátce se o tomto přístupu, který autorka označuje jako modely ZIP(τ), zmíníme v podkapitole 3.4. Nyní se věnujme maximálně věrohodným odhadům regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ v modelu ZIP, jak jsme jej výše zavedli.

Vyjdeme-li z rozdělení odezvy (3.2), ihned dostaneme věrohodnost pozorovaných dat v ZIP, a sice

$$\mathcal{L}(\boldsymbol{\gamma}\,,\boldsymbol{\beta}) = \prod_{i:Y_i=0}\left[p_i + (1-p_i)\,\mathrm{e}^{-\lambda_i}\right] \cdot \prod_{i:Y_i>0}\left[(1-p_i)\frac{\mathrm{e}^{-\lambda_i}\,\lambda_i^{Y_i}}{Y_i!}\right].$$

Po dosazením z (3.1) za p_i a λ_i a zlogaritmování celého výrazu obdržíme předpis logaritmické věrohodnosti pozorovaných dat

$$\begin{aligned}\ell(\boldsymbol{\gamma}\,,\boldsymbol{\beta}) = &\sum_{i:Y_i=0} \log\Big(\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}} + \exp\Big(-\mathrm{e}^{\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}}\Big)\Big) \\ &+ \sum_{i:Y_i>0} \Big(Y_i\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta} - \mathrm{e}^{\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}} - \log(Y_i!)\Big) \\ &- \sum_{i=1}^{n} \log\Big(1 + \mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}\Big).\end{aligned} \tag{3.3}$$

Standardně bychom se nyní snažili získat odhadovací rovnici, jejímž řešením je kýžený maximálně věrohodný odhad regresních koeficientů. Nicméně v našem případě je tento postup značně komplikovaný, neboť struktura logaritmické věrohodnosti neumožňuje oddělit $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ a hledat tak jejich maxima zvlášť. Namísto toho v článku Lambert (1992) navrhuje autorka postupovat pomocí EM algoritmu. Ten nabízí poměrně přímočarý přístup ke konstrukci těchto odhadů. Připomeňme ještě, že jsme myšlenku EM algoritmu představili v úvodní kapitole na straně 15.

Využijeme přitom vektor latentních veličin $\mathbf{Z} = (Z_1, \ldots, Z_n)^\mathsf{T}$, pro jehož složky platí

$$Z_i|\mathbf{W}_i \sim Alt(p_i). \tag{3.4}$$

Rozšíříme data o vektor $\mathbf{Z}$. Logaritmus věrohodnosti úplných dat $\big(Y_i, \mathbf{X}_i^\mathsf{T}, Z_i\big)$, $i = 1, \ldots, n$, pak můžeme zapsat obecně jako

$$\ell_c(\boldsymbol{\beta}\,,\boldsymbol{\gamma}) = \sum_{i=1}^{n} \log(f_Y(Y_i|\mathbf{X}_i, Z_i)) + \sum_{i=1}^{n} \log(f_Z(Z_i|\mathbf{X}_i)). \tag{3.5}$$

Hustota $f_Y(Y_i|\mathbf{X}_i, Z_i)$ je pro $Z_i = 1$ hustotou degenerovaného rozdělení v 0 a pro $Z_i = 0$ hustotou Poissonova rozdělení s parametrem λ_i. Po úpravách tak můžeme sčítance v první sumě (3.5) vyjádřit jako

$$\log\Big(f_Y(Y_i|\mathbf{X}_i, Z_i)\Big) = (1 - Z_i)[Y_i \log(\lambda_i) - \lambda_i - \log(Y_i!)].$$

Sčítance v druhé sumě logaritmické věrohodnosti úplných dat (3.5) jsou dle (3.4) po úpravách rovny

$$\begin{aligned}\log\Big(f_Z(Z_i|\mathbf{X}_i)\Big) &= \log\Big(p_i^{Z_i}(1 - p_i)^{1-Z_i}\Big) \\ &= Z_i \operatorname{logit}(p_i) + \log(1 - p_i).\end{aligned}$$

Celkem tedy dostáváme následující předpis pro logaritmickou věrohodnost úplných dat, a sice

$$\ell_c(\boldsymbol{\gamma}\,,\boldsymbol{\beta}) = \ell_c(\boldsymbol{\gamma}) + \ell_c(\boldsymbol{\beta}) - \sum_{i=1}^{n}(1 - Z_i)\log(Y_i!),$$

kde jsme označili

$$\begin{aligned}\ell_c(\boldsymbol{\beta}) = \ell_c(\boldsymbol{\beta}; \mathbf{Y}, \mathbb{V}, \mathbf{Z}) &= \sum_{i=1}^{n}(1 - Z_i)[Y_i \log(\lambda_i) - \lambda_i], \\ \ell_c(\boldsymbol{\gamma}) = \ell_c(\boldsymbol{\gamma}; \mathbf{Y}, \mathbb{W}, \mathbf{Z}) &= \sum_{i=1}^{n}[Z_i \operatorname{logit}(p_i) + \log(1 - p_i)].\end{aligned} \tag{3.6}$$

Maximalizovat logaritmus věrohodnosti úplných dat $\ell_c(\boldsymbol{\gamma}, \boldsymbol{\beta})$ je nyní snadné, neboť můžeme jednotlivé části $\ell_c(\boldsymbol{\gamma})$ a $\ell_c(\boldsymbol{\beta})$ maximalizovat zvlášť.

K nalezení maximálně věrohodných odhadů $\hat{\boldsymbol{\beta}}_n$ a $\hat{\boldsymbol{\gamma}}_n$ použijeme EM algoritmus. Předpokládejme, že se algoritmus nachází v j-té iteraci a máme tak k dispozici $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$ a $Z_i^{(j-1)}$. Další iterace algoritmu spočívá v následujících dvou krocích:

E krok Jak jsme uvedli na straně 15, naším úkolem je spočítat $\mathcal{Q}\big(\boldsymbol{\theta}\,, \boldsymbol{\theta}^{(j)}\big)$, střední hodnotu logaritmické věrohodnosti vzhledem k rozdělení $\mathbf{Z}|\mathbb{X}$, a to za použití aktuálního přiblížení $\boldsymbol{\theta}^{(j)}$ namísto neznámého parametru $\boldsymbol{\theta}$. V našem případě je přitom $\boldsymbol{\theta}^\mathsf{T} = \big(\boldsymbol{\beta}^\mathsf{T}\,, \boldsymbol{\gamma}^\mathsf{T}\big)$.

Nejprve spočítejme předpis rozdělení, jež mají latentní veličiny Z_i v aktuálním kroku iterace. Protože podle (3.4) má Z_i alternativní rozdělení s parametrem p_i, vezměme za odhad $Z_i^{(j)}$ její aposteriorní střední hodnotu, tedy

$$\begin{aligned} Z_i^{(j)} &= \mathsf{P}\Big(Z_i = 1\Big|Y_i, \mathbf{X}_i, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\Big) \\ &= \frac{f\Big(Y_i\Big|\mathbf{X}_i, Z_i = 1, \boldsymbol{\beta}^{(j)}\Big)\mathsf{P}\Big(Z_i = 1\Big|\mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big)}{\sum\limits_{\nu\in\{0,1\}} f\Big(Y_i\Big|\mathbf{X}_i, Z_i = \nu, \boldsymbol{\beta}^{(j)}\Big)\mathsf{P}\Big(Z_i = \nu\Big|\mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big)} \\ &= \begin{cases} \dfrac{p_i^{(j)}}{p_i^{(j)} + (1 - p_i^{(j)})\,\mathrm{e}^{-\lambda_i^{(j)}}}, & \text{pro } Y_i = 0, \\ 0, & \text{pro } Y_i \in \mathcal{N}, \end{cases} \end{aligned}$$

kde jsme parametry $p_i^{(j)}$ a $\lambda_i^{(j)}$ získali z rovnic (3.1) na straně 25 dosazením aktuálních přiblížení $\boldsymbol{\beta}^{(j)}$ a $\boldsymbol{\gamma}^{(j)}$. Platí totiž

$$\begin{aligned} f\Big(Y_i\Big|\mathbf{X}_i, Z_i = 1, \boldsymbol{\beta}^{(j)}\Big) &= \mathbb{1}_{[Y_i=0]}, & \mathsf{P}\Big(Z_i = 1|\mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big) &= p_i^{(j)}, \\ f\Big(Y_i\Big|\mathbf{X}_i, Z_i = 0, \boldsymbol{\beta}^{(j)}\Big) &= g\big(Y_i; \lambda_i^{(j)}\big), & \mathsf{P}\Big(Z_i = 0|\mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big) &= 1 - p_i^{(j)}. \end{aligned}$$

Poněkud nestandardní označení pravděpodobnosti $Z_i^{(j)}$ jsme zvolili proto, že v dalším budeme veličiny Z_i nahrazovat právě pravděpodobnostmi $Z_i^{(j)}$. Toto označení nám tedy, doufejme, zpřehlední zápis.

Nyní již snadno dostaneme předpis funkce $\mathcal{Q}(.)$. Podle (3.6) můžeme rozdělit logaritmickou věrohodnost úplných dat na součet $\ell_c(\boldsymbol{\beta})$ a $\ell_c(\boldsymbol{\gamma})$ a konstantu nezávisející ani na jednom z regresních koeficientů. Konstantní člen v dalším již nebudeme uvažovat, neboť při maximalizaci $\mathcal{Q}(.)$ v následujícím kroku EM algoritmu nehraje žádnou roli. Díky linearitě podmíněné střední hodnoty můžeme vyjádřit výraz $\mathcal{Q}(.)$ jako

$$\mathcal{Q}\big(\boldsymbol{\theta}\,, \boldsymbol{\theta}^{(j)}\big) = \mathsf{E}_{Z_i|\mathbf{Y},\mathbb{X},\boldsymbol{\theta}^{(j)}}\Big[\ell_c(\boldsymbol{\beta})\Big|\mathbf{Y},\mathbb{V},\boldsymbol{\theta}^{(j)}\Big] + \mathsf{E}_{Z_i|\mathbf{Y},\mathbb{X},\boldsymbol{\theta}^{(j)}}\Big[\ell_c(\boldsymbol{\gamma})\Big|\mathbf{Y},\mathbb{W},\boldsymbol{\theta}^{(j)}\Big].$$

Zbývá si uvědomit, že jak v $\ell_c(\boldsymbol{\beta})$, tak i v $\ell_c(\boldsymbol{\gamma})$ je náhodná veličina Z_i zastoupena lineárně. Oba dva sčítance proto snadno získáme z výrazů (3.6) prostým nahrazením alternativně rozdělené veličiny Z_i odhadem pravděpodobnosti $Z_i^{(j)}$, že je tato veličina rovna jedné. Platí tedy

$$\mathcal{Q}\big(\boldsymbol{\theta}\,, \boldsymbol{\theta}^{(j)}\big) = \ell_c\big(\boldsymbol{\beta}; \boldsymbol{Y}, \mathbb{V}, \mathbf{Z}^{(j)}\big) + \ell_c\big(\boldsymbol{\gamma}; \boldsymbol{Y}, \mathbb{W}, \mathbf{Z}^{(j)}\big).$$

M krok Nalezneme $\boldsymbol{\beta}^{(j+1)}$ a $\boldsymbol{\gamma}^{(j+1)}$ jako

$$\text{a)}\ \ \boldsymbol{\beta}^{(j+1)} = \arg\max_{\boldsymbol{\beta}} \ell_c\Big(\boldsymbol{\beta}; \boldsymbol{Y}, \mathbb{V}, \mathbf{Z}^{(j)}\Big),$$
$$\text{b)}\ \ \boldsymbol{\gamma}^{(j+1)} = \arg\max_{\boldsymbol{\gamma}} \ell_c\Big(\boldsymbol{\gamma}; \boldsymbol{Y}, \mathbb{W}, \mathbf{Z}^{(j)}\Big).$$

Můžeme postupovat takto:

a) Jak je uvedeno v článku Lambert (1992), výraz $\ell_c\Big(\boldsymbol{\beta}; \mathbf{Y},\mathbb{V}, \mathbf{Z}^{(j)}\Big)$ je vlastně logaritmickou věrohodností váženého poissonovského regresního modelu s kanonickou linkovou funkcí, v němž je vektorem odezvy $\mathbf{Y}$, maticí regresorů $\mathbb{V}$ a vektorem vah $\Big(1 - Z_1^{(j)}, \ldots, 1 - Z_n^{(j)}\Big)^\mathsf{T}$.

Zderivujeme-li $\ell_c\Big(\boldsymbol{\beta}; \mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big)$, dostaneme skórovou statistiku

$$\mathbf{U}_c\Big(\boldsymbol{\beta}; \mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big) = \sum_{i=1}^{n}\Big(1 - Z_i^{(j)}\Big)(Y_i - \lambda_i)\mathbf{V}_i.$$

Řešením odhadovací rovnice $\mathbf{U}_c\Big(\boldsymbol{\beta}; \mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big) = 0$ např. iterativní metodou vážených nejmenších čtverců je hledané $\boldsymbol{\beta}^{(j+1)}$, další přiblížení maximálně věrohodného odhadu $\hat{\boldsymbol{\beta}}_n$.

b) Obdobně jako v předchozím případě zderivujeme logaritmickou věrohodnost $\ell_c(\boldsymbol{\gamma})$, abychom tak získali skórovou statistiku

$$\mathbf{U}_c(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\left[\ Z_i^{(j)} - \frac{1}{1 + \mathrm{e}^{-\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}}\right]\mathbf{W}_i.$$

Uvědomme si, že platí

$$\frac{1}{1 + \mathrm{e}^{-\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}} = \frac{\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}}{1 + \mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}} = \mathrm{logit}^{-1}(\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}) = p_i.$$

Odhadovací rovnice tak má následující předpis

$$\sum_{i=1}^{n}\Big[\ Z_i^{(j)} - p_i\Big]\mathbf{W}_i = 0.$$

Jejím řešením např. iterativní metodou vážených nejmenších čtverců je další přiblížení $\boldsymbol{\gamma}^{(j+1)}$ maximálně věrohodného odhadu parametru $\boldsymbol{\gamma}$.

Oba kroky iterujeme do konvergence. Odhady $\boldsymbol{\beta}^{(j)}$ a $\boldsymbol{\gamma}^{(j)}$ z poslední iterace jsou hledanými maximálně věrohodnými odhady, jak vyplývá z vět 4, 5 a 6 uvedených na straně 16.

Vhodnou volbou počátečních hodnot, tedy $\boldsymbol{\beta}^{(0)}$ a $\boldsymbol{\gamma}^{(0)}$, docílíme relativně rychlé konvergence algoritmu. Za $\boldsymbol{\beta}^{(0)}$ je v práci Lambert (1992) doporučeno vzít maximálně věrohodný odhad $\boldsymbol{\beta}$ z modelu ZTP spočteného na vyšetřovaných datech zbavených nulové odezvy (víc viz kapitola 2.3), za $\boldsymbol{\gamma}^{(0)}$ pak vektor nulový až na koeficient absolutního členu $\gamma_1^{(0)}$. Pro něj je doporučeno zvolit odhad pravděpodobnosti, že nulové pozorování pochází z *perfektního* stavu, tedy

$$\gamma_1^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbb{1}_{[Y_i=0]} - \mathrm{e}^{\exp(\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}^{(0)})}\Big].$$

3.4 Modely ZIP(τ) – alternativní definice ZIP

Jak jsme již zmínili v podkapitole 3.2, oproti klasickému modelu poissonovské regrese má model ZIP stejné či větší množství parametrů, které je třeba odhadnout. Na závěr této kapitoly se budeme zabývat alternativním zavedením modelů ZIP, tzv. modelům ZIP(τ), jež byly rovněž poprvé představeny v práci Lambert (1992). Přínosem přístupu ZIP(τ) oproti ZIP tkví právě v redukci počtu odhadovaných parametrů. Existují i další možnosti, jak alternativně definovat modely ZIP, zde je však pomineme.

Tak jak byly modely ZIP výše představeny, nepředpokládají vztah mezi pravděpodobností *perfektního* stavu p_i a λ_i, střední hodnotou v *podkladovém rozdělení*. To však nemusí vždy odpovídat realitě – uvažujme opět situaci, na níž jsme v podkapitole 3.1 demonstrovali použití modelů ZIP, a sice modelování počtu výrobních chyb v závislosti na seřízení stroje. Je přirozené předpokládat, že existuje pomyslná *míra špatnosti seřízení stroje*. Dokud je tato *míra* pod jistou hranicí, je stroj nastaven dostatečně správně a chyby nenastávají. Jde tedy o *perfektní* stav. Je-li tato hranice překročena, začíná s jistou pravděpodobností k chybám docházet. Lze si představit, že výskyt chyby je se zvyšující se *mírou špatnosti seřízení* stále pravděpodobnější, až nakonec dospějeme do stavu, kdy je stroj seřízen natolik špatně, že prakticky není schopen produkovat nedefektní výrobky.

Jde tedy vlastně o připad, kdy tytéž regresory ovlivňují jak p_i tak λ_i. Lambert (1992) proto navrhuje předpokládat funkční vztah mezi těmito dvěma parametry a zavádí tzv. ZIP(τ) model modifikací předpisu linkových funkcí $h_1(.)$ a $h_2(.)$ původního modelu. Rovnice (3.1) uvedené na na straně 25 nahraďme následujícím způsobem

$$\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \qquad \operatorname{logit}(p_i) = -\tau\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \tag{3.7}$$

kde τ je reálný parametr. Vztah mezi parametry p_i a λ_i pak můžeme vyjádřit jako

$$p_i = \frac{1}{1+\lambda_i^\tau}.$$

Volbou jiné linkové funkce v (3.7) můžeme modelovat další funkční závislosti mezi p_i a λ_i. Aniž bychom se jimi v dalším zabývali, uveďme příklady zmiňované ve článcích Lambert (1992) a Heilbron (1994):

$$\begin{aligned}
&\log(-\log(p_i)) = \tau\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta} && \text{vedoucí na} && p_i = \exp(-\lambda_i^\tau),\\
&\log(-\log(1-p_i)) = -\tau\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta} && \text{vedoucí na} && p_i = 1-\exp\left(-\lambda_i^{-\tau}\right),\\
&\log(-\log(p_i)) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}+\log(\tau) && \text{vedoucí na} && p_i = \exp(-\tau\lambda_i),\\
&\log(-\log(p_i)) = \log(\alpha)+\tau\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta} && \text{vedoucí na} && p_i = \frac{\alpha}{\exp\left(\tau\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}\right)+\alpha}.
\end{aligned}$$

Vraťme se však k volbě (3.7). Zafixujeme-li parametr τ na nějaké pevné hodnotě, můžeme rozlišit tyto případy:

- Pro $\tau > 0$ se vzrůstající hodnotou λ_i (a tedy s očekávaným počtem událostí) klesá pravděpodobnost *perfektního* stavu.
- Pro $\tau < 0$ se naopak se vzrůstající hodnotou λ_i stává *perfektní* stav jistější.

- A konečně pro $\tau = 0$ je zřejmě pravděpodobnost *perfektního stavu* konstantně 0,5, není tedy ovlivněna regresory. Odezva Y_i pochází z *perfektního stavu* se stejnou pravděpodobností jako z podkladového rozdělení.

Logaritmickou věrohodnost modelu ZIP(τ) můžeme zapsat jako

$$\begin{aligned}\ell(\boldsymbol{\beta},\tau;\mathbf{Y},\mathbb{V}) = &\sum_{i:Y_i=0} \log\Big(\mathrm{e}^{-\tau\mathbf{V}_i^\intercal\boldsymbol{\beta}} + \exp(-\mathrm{e}^{\mathbf{V}_i^\intercal\boldsymbol{\beta}})\Big) \\ &+ \sum_{i:Y_i>0} \Big(Y_i\mathbf{V}_i^\intercal\boldsymbol{\beta} - \mathrm{e}^{\mathbf{V}_i^\intercal\boldsymbol{\beta}}\Big) \\ &- \sum_{i=1}^{n} \log\Big(1 + \mathrm{e}^{-\tau\mathbf{V}_i^\intercal\boldsymbol{\beta}}\Big).\end{aligned}$$

(srovnej s předpisem logaritmické věrohodnosti modelu ZIP, rovnice (3.3) na straně 26). Výše navrhovaným postupem nyní bohužel nelze parametry $\boldsymbol{\beta}$ a τ v předpisu $\ell(\boldsymbol{\beta},\tau)$ oddělit. Není proto namístě použít EM algoritmus navrhovaný v případě modelů ZIP, neboť bychom s jeho pomocí nezískali žádnou výhodu. V článku Lambert (1992) je navrženo aplikovat Newtonův-Raphsonův algoritmus s následující volbou počátečních přiblížení

$$\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}^{ZIP}, \qquad \tau^{(0)} = -\underset{j=1,\ldots,J}{\mathrm{median}}\left(\frac{\hat{\gamma}_j^{ZIP}}{\hat{\beta}_j^{ZIP}}\right),$$

kde $\hat{\boldsymbol{\beta}}^{ZIP}$ a $\hat{\boldsymbol{\gamma}}^{ZIP}$ jsou maximálně věrohodné odhady koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ získané ze standardního modelu ZIP a $J = \min(v,w)$ (připomeňme, že $\boldsymbol{\beta} \in \mathcal{R}^v$ a $\boldsymbol{\gamma} \in \mathcal{R}^w$). Pokud by tento přístup v konkrétním případě selhal, navrhuje autorka neuvažovat absolutní člen při výpočtu $\tau^{(0)}$ a za $\boldsymbol{\beta}^{(0)}$ přitom zvolit

$$\boldsymbol{\beta}^{(0)} = \arg\max_{\boldsymbol{\beta}} \ell_c\Big(\boldsymbol{\beta},\tau^{(0)};\mathbf{Y},\mathbb{V}\Big),$$

přičemž výraz $\ell_c\Big(\boldsymbol{\beta},\tau^{(0)};\mathbf{Y},\mathbb{V}\Big)$ získáme z rovnice (3.6) uvedené na straně 26 snadným dosazením z rovnic (3.7) ze strany 29.

3.5 Binomický model s nadbytečnými nulami jako alternativa ZIP

Na závěr kapitoly věnované *poissonovským modelům s nadbytečnými nulami* se ještě jednou zabývejme příklady jejich použití, které jsme uvedli v kapitole 3.1. V některých případech je ihned patrné, že volba Poissonova rozdělení za rozdělení podkladové není úplně optimální a mnohem příhodnější by byl *binomický model s nadbytečnými nulami* (*zero inflated binomial model*, ZIB). Ukazuje to i následující příklad převzatý z práce Dietz a Böhning (2000).

Jedním ze způsobů, jímž se v praxi charakterizuje stav chrupu dítěte, je tzv. DMFT index, který představuje součet zkažených, chybějících či plombou ošetřených zubů jedince. Vzhledem k velikosti dětského chrupu může DMFT index nabývat pouze hodnot z množiny $\{0,\ldots,20\}$. Použití Poissonova rozdělení je tak pouze hrubou aproximací. S obdobným problémem se setkáme i v původním článku Lambert (1992), v němž byly modely ZIP představeny. Na každém obráběném

předmětu bylo vždy provedeno 48 sváření a odezva zachycující počet neúspěšných pokusů tak jistě nabývá hodnot pouze z množiny $\{0,\ldots,48\}$. V tomto případě je však aproximace Poissonovým rozdělením přesnější.

Vidíme tedy, že v mnoha situacích není navrhovaný model ZIP nejvhodnějším a že mnohem lepší volbou by byl zmíněný *binomický model s nadbytečnými nulami*. Definici modelů ZIB můžeme nalézt např. v článku Hall (2000), jinak se s touto třídou modelů v literatuře příliš nesetkáme. Ani v této práci se modelům ZIB podrobněji věnovat nebudeme. Jedná se ale o snadnou aplikaci teorie uvedené v následující kapitole, v níž se budeme zabývat obecně *modely s nadbytečnými nulami* s diskrétním podkladovým rozdělením exponenciálního typu.

4. Modely s nadbytečnými nulami

Poté, co jsme se v předchozí kapitole věnovali jejich speciálnímu případu, tzv. *poissonovskému modelu s nadbytečnými nulami*, zabývejme se nyní obecně *modely s nadbytečnými nulami* (*zero inflated models*, ZIM). Budeme opět předpokládat, že se odezva řídí směsí rozdělení degenerovaného v nule (*perfektní stav*) a rozdělení $\mathcal{G}$ (*podkladové rozdělení*) s hustotou $g(.)$. Přitom rozdělení $\mathcal{G}$ nechť je nyní libovolné diskrétní rozdělení exponenciálního typu. Mezi tato rozdělení patří Poissonovo rozdělení, binomické rozdělení či negativně binomické rozdělení typu 2 s pevně zvoleným parametrem $\alpha > 0$, jak jsme ukázali v úvodní kapitole.

V dostupné literatuře jsme nenarazili na práce zabývající se tímto způsobem problematikou ZIM. V článku Deng a Paul (2005) jsou sice představeny modely ZIM, jejichž podkladové rozdělení je exponenciálního typu, avšak na rozdíl od článku Lambert (1992) nezávisí pravděpodobnost *perfektního stavu* na regresorech. S obdobným přístupem pro Poissonovo rozdělení se setkáme např. v práci van den Broek (1995) a pro některá další běžně užívaná diskrétní rozdělení v práci Rao a Sumathi (2009). Jelikož se jedná o speciální případ níže definovaného modelu ZIM, v podkapitole 4.2 tento model pouze zavedeme, jeho vlastnostmi se však v této práci zabývat nebudeme.

I v některých článcích věnovaných speciálním případům ZIM (ať už modelům ZIP či ZINB2) se sice setkáme s obecnou definicí modelů ZIM v souladu s původní myšlenkou uvedenou v práci Lambert (1992), např. v práci Ghosh et al. (2012), předmětem těchto prací však není konstrukce maximálně věrohodných odhadů regresních koeficientů, jíž se zde budeme zabývat. V podkapitole 4.3 proto odvodíme tyto odhady sami modifikací postupu, jenž byl pro modely ZIP představen v článku Lambert (1992) a jímž jsme se zabývali v předchozí kapitole.

4.1 Definice modelu

Opět mějme k dispozici datovou strukturu splňující předpoklad (A) uvedený na straně 16, přičemž podkladové rozdělení $\mathcal{G}$ nechť je diskrétní rozdělení exponenciálního typu s hustotou (vůči čítací míře)

$$g(j;\psi_{\mathrm{i}},\varphi) = \exp\left(\frac{j\,\psi_{\mathrm{i}} - b(\psi_{\mathrm{i}})}{\varphi} + c(j,\varphi)\right), \qquad j \in \mathcal{N}_0.$$

Připomeňme, že pro linkové funkce $h_1(.)$ a $h_2(.)$ z předpokladu (A) platí

$$p_i = h_1^{-1}\left(\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}\right), \qquad \mu_i = h_2^{-1}\left(\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}\right), \tag{4.1}$$

kde $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_v)^{\mathsf{T}}$ a $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_w)^{\mathsf{T}}$ jsou vektory regresních koeficientů. V dalším se omezíme pouze na následující volbu linkové funkce $h_1(.)$, a sice

$$h_1(.) = \mathrm{logit}(.).$$

Tato poměrně přirozená volba nám značně usnadní a zpřehlední zápis i jednotlivé úpravy. V případě jiné linkové funkce $h_1(.)$ bychom postupovali analogicky. Poznamenejme, že právě k této volbě $h_1(.)$ se uchýlili i autoři výše zmiňovaných

prací věnovaných *poissonovskému* i *negativně binomickému modelu s nadbytečnými nulami* (např. Lambert (1992), Ridout et al. (2001) či Welsh et al. (1996)).

Rozdělení odezvy $Y_i|\mathbf{X}_i$ nyní můžeme vyjádřit jako

$$\begin{aligned}
\mathsf{P}(Y_i = 0|\mathbf{X}_i) &= p_i + (1-p_i)g(0;\psi_{\mathrm{i}},\varphi) \\
&= p_i + (1-p_i)\exp\left(\frac{-b(\psi_{\mathrm{i}})}{\varphi} + c(0,\varphi)\right), \\
\mathsf{P}(Y_i = j|\mathbf{X}_i) &= (1-p_i)g(j;\psi_{\mathrm{i}},\varphi) \\
&= (1-p_i)\exp\left(\frac{j\,\psi_{\mathrm{i}} - b(\psi_{\mathrm{i}})}{\varphi} + c(j,\varphi)\right), \quad j \in \mathcal{N}.
\end{aligned} \tag{4.2}$$

Pro další úvahy označme $f_Y(.)$ hustotu tohoto rozdělení a $f_Z(.)$ hustotu latentních veličin $Z_i|\mathbf{W}_i$. Jedná se samozřejmě o hustoty vůči čítací míře.

Poznamenejme ještě, že předpis rozdělení má zjevně smysl i v případě, kdy podkladové rozdělení není exponenciálního typu. Toto poněkud umělé omezení jsme zvolili proto, že v dalším využijeme vlastností rozdělení náležících této třídě při konstrukci odhadů regresních koeficientů. Nic nám ovšem nebrání konstruovat i jiné ZMM modely, v nichž tento požadavek na podkladové rozdělení není splněn. Toho využijeme v následující kapitole při konstrukci některých *negativně binomických modelů s nadbytečnými nulami*.

Než se budeme zabývat maximálně věrohodnými odhady, uvedeme ještě předpisy střední hodnoty a rozptylu odezvy. Tak jako jsme v předpokladu (A) označili střední hodnotu podkladového rozdělení μ_i, označme ještě $\mu_i^{[2]}$ druhý moment a σ_i^2 rozptyl tohoto rozdělení. Pak snadnými úpravami dostaneme

$$\begin{aligned}
\mathsf{E}[Y_i|\mathbf{X}_i] &= \sum_{j\in\mathcal{N}_0} j\mathsf{P}(Y_i = j|\mathbf{X}_i) = \sum_{j\in\mathcal{N}} j\mathsf{P}(Y_i = j|\mathbf{X}_i) \\
&= \sum_{j\in\mathcal{N}} j(1-p_i)g(j;\psi_{\mathrm{i}},\varphi) \\
&= (1-p_i)\mu_i.
\end{aligned}$$

Obdobně dostaneme i druhý moment tohoto rozdělení, a sice

$$\mathsf{E}\left[Y_i^2\middle|\mathbf{X}_i\right] = (1-p_i)\mu_i^{[2]}.$$

Pro rozptyl odezvy tak platí

$$\begin{aligned}
\mathrm{var}[Y_i|\mathbf{X}_i] &= \mathsf{E}\left[Y_i^2\middle|\mathbf{X}_i\right] - (\mathsf{E}[Y_i|\mathbf{X}_i])^2 \\
&= (1-p_i)\mu_i^{[2]} - (1-p_i)^2\mu_i^2 \\
&= (1-p_i)\left(\mu_i^{[2]} - \mu_i^2 + p_i\mu_i^2\right) \\
&= (1-p_i)\left(\sigma_i^2 + p_i\mu_i^2\right).
\end{aligned}$$

4.2 Modely s pevnou proporcí nadbytečné nulové složky

Jak jsme zmínili v úvodu kapitoly, v literatuře se někdy setkáváme i s poněkud zjednodušenou verzí modelů ZIM, v níž pravděpodobnost *perfektního stavu* nezávisí na regresorech. Zmiňme například práce van den Broek (1995) či Min a Czado (2010), kde je podkladovým rozdělením rozdělení Poissonovo. Tuto zjednodušenou verzi modelů ZIM nyní v krátkosti představíme a uvedeme předpis rozdělení odezvy.

Protože pravděpodobnost *perfektního stavu* nezávisí na regresorech, můžeme proporci degenerovaného rozdělení ve směsném rozdělení odezvy vyjádřit jediným parametrem $p \in [0,1]$. Rovnici (4.2) definující rozdělení odezvy v modelech ZIM pak upravíme následovně

$$\begin{aligned}\mathsf{P}(Y_i = 0|p, \mathbf{X}_i) &= p + (1-p)g(0; \psi_{\mathrm{i}}, \varphi)\\ \mathsf{P}(Y_i = j|p, \mathbf{X}_i) &= (1-p)g(j; \psi_{\mathrm{i}}, \varphi)\end{aligned}$$

Předpokládáme tedy, že linková funkce $h_1(.)$ je konstantní a že tedy platí

$$h_1^{-1}\left(\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}\right) = p, \qquad \forall\, \mathbf{W}_i \in \mathcal{R}^w.$$

Přitom zřejmě $p = 1$ značí jistotu *perfektního stavu* a $p = 0$ naopak jeho nemožnost.

Vlastnostmi speciálně této podtřídy ZIM se zde zabývat nebudeme, lze je však získat snadnou modifikací postupu uvedeného dále pro modely ZIM. Poznamenejme však, že toto zjednodušení umožňuje např. poměrně snadné odvození testových statistik testů o regresních koeficientech, jejichž konstrukce pro původní ZIM je značně obtížná.

4.3 Maximálně věrohodné odhady

Postup pro nalezení maximálně věrohodných odhadů $\hat{\boldsymbol{\beta}}_n$ a $\hat{\boldsymbol{\gamma}}_n$ je zobecněním postupu uvedeného v podkapitole 3.3, kde jsme se zabývali hledáním těchto odhadů v *poissonovském modelu s nadbytečnými nulami* (ZIP). Zde však budou výpočty komplikovanější, neboť nám obecný tvar hustoty a linkové funkce $h_2(.)$ neumožní přílišné zjednodušení získaných výrazů.

Z rovnice (4.2) ihned máme věrohodnost pozorovaných dat, a sice

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \prod_{i=1}^{n} f_Y(Y_i|\mathbf{V}_i, \mathbf{W}_i)\\ &= \prod_{i:Y_i=0} [p_i + (1-p_i)g(0; \psi_{\mathrm{i}}, \varphi)] \cdot \prod_{i:Y_i\neq 0} (1-p_i)g(Y_i; \psi_{\mathrm{i}}, \varphi).\end{aligned}$$

Zlogaritmováním a dosazením za p_i a ψ_i z (4.1) dostaneme předpis logaritmické věrohodnosti

$$\begin{aligned}\ell(\boldsymbol{\beta}\,,\boldsymbol{\gamma}) &= \sum_{i:Y_i=0} \log\!\left(\frac{\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}}{1+\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}} + \frac{g(0;\psi_\mathrm{i}\,,\varphi)}{1+\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}}\right) + \sum_{i:Y_i\neq 0} \log\!\left(\frac{g(Y_i;\psi_\mathrm{i}\,,\varphi)}{1+\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}}\right)\\ &= \sum_{i:Y_i=0} \log\!\left(\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}} + \exp\left(-\frac{b\Big(h_2^{-1}\Big(\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}\Big)\Big)}{\varphi} + c(0,\varphi)\right)\right)\\ &+ \sum_{i:Y_i\neq 0}\left[\frac{Y_i h_2^{-1}\Big(\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}\Big) - b\Big(h_2^{-1}\Big(\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}\Big)\Big)}{\varphi} + c(Y_i,\varphi)\right]\\ &- \sum_{i=1}^{n}\log\Big(1+\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}\Big).\end{aligned} \tag{4.3}$$

Standardní postup výpočtu maximálně věrohodných odhadů je v tomto případě značně obtížný, pokusíme se proto postupovat obdobně jako při konstrukci maximálně věrohodných odhadů v ZIP.

Vezměme v potaz i latentní veličiny $\mathbf{Z} = (Z_1,\ldots,Z_n)^\mathsf{T}$, které máme dle předpokladu (A) k dispozici (viz strana 16). Již v předchozí kapitole jsme připomněli, že pro logaritmickou věrohodnost úplných dat platí

$$\ell_c(\boldsymbol{\beta}\,,\boldsymbol{\gamma}) = \sum_{i=1}^{n}\log\Big[f_Y(Y_i|\mathbf{V}_i,Z_i)\Big] + \sum_{i=1}^{n}\log\Big[f_Z(Z_i|\mathbf{W}_i)\Big]. \tag{4.4}$$

Protože pro $Z_i = 1$ je rozdělení Y_i degenerované v nule a pro $Z_i = 0$ se jedná o rozdělení s hustotou $g(.)$, můžeme sčítance z první sumy (4.4) přepsat jako

$$\begin{aligned}\log\Big[f_Y(Y_i|\mathbf{V}_i,Z_i)\Big] &= (1-Z_i)\log\Big(g(Y_i;\psi_\mathrm{i}\,,\varphi)\Big)\\ &= (1-Z_i)\left[\frac{Y_i\,\psi_\mathrm{i} - b(\psi_\mathrm{i})}{\varphi} + c(Y_i,\varphi)\right].\end{aligned}$$

Sčítance ve druhé sumě v předpisu (4.4) dostaneme ihned ze znalosti rozdělení latentních veličin $Z_i|\mathbf{W}_i$, které je dle předpokladu alternativní s parametrem p_i, a tedy

$$\begin{aligned}\log\Big(f_Z(Z_i|\mathbf{W}_i)\Big) &= Z_i\,\mathrm{logit}(p_i) + \log(1-p_i)\\ &= Z_i\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma} - \log\Big(1+\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}\Big).\end{aligned}$$

Označíme-li nyní

$$\begin{aligned}\ell_c(\boldsymbol{\beta}) = \ell_c(\boldsymbol{\beta}\,|\mathbf{Y},\mathbf{V}_i,\mathbf{Z}) &= \sum_{i=1}^{n}(1-Z_i)\left[\frac{Y_i\,\psi_\mathrm{i} - b(\psi_\mathrm{i})}{\varphi} + c(Y_i,\varphi)\right],\\ \ell_c(\boldsymbol{\gamma}) = \ell_c(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbf{W}_i,\mathbf{Z}) &= \sum_{i=1}^{n}\Big[Z_i\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma} - \log\Big(1+\mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}\Big)\Big],\end{aligned}$$

můžeme logaritmickou věrohodnost úplných dat (4.4) přepsat jako

$$\ell_c(\boldsymbol{\beta}\,,\boldsymbol{\gamma}) = \ell_c(\boldsymbol{\beta}) + \ell_c(\boldsymbol{\gamma}). \tag{4.5}$$

Hledání maximálně věrohodných odhadů regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ je nyní podstatně snazší, neboť jednotlivé části logaritmické věrohodnosti $\ell_c(.)$ můžeme maximalizovat zvlášť.

V dalším budeme postupovat opět podle EM algoritmu. Připomeňme, že jsme jej v krátkosti představili na straně 15, avšak mnohem názornější by pro čtenáře mohla být aplikace tohoto algoritmu v podkapitole 3.3 na straně 27 – níže uvedený postup je její obdobou.

Předpokládejme tedy, že se algoritmus nachází v j-té iteraci, $j \in \mathcal{N}$. Z dosavadního průběhu tak máme k dispozici $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$ a $\mathbf{Z}^{(j-1)}$ a při hledání dalšího přiblížení postupujeme takto:

E krok Nejprve spočítáme předpis rozdělení latentních veličin Z_i v právě vyšetřované iteraci EM algoritmu. Protože o veličinách Z_i předpokládáme, že jsou alternativně rozdělené s parametrem p_i, vezmeme za $Z_i^{(j)}$ její aposteriorní střední hodnotu, tedy

$$\begin{aligned} Z_i^{(j)} &= \mathsf{P}\Big(Z_i = 1 | Y_i, \mathbf{X}_i, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\Big) \\ &= \frac{f\Big(Y_i | Z_i = 1, \mathbf{X}_i, \boldsymbol{\beta}^{(j)}\Big) \mathsf{P}\Big(Z_i = 1 | \mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big)}{\sum\limits_{\nu \in \{0,1\}} f\Big(Y_i | Z_i = \nu, \mathbf{X}_i, \boldsymbol{\beta}^{(j)}\Big) \mathsf{P}(Z_i = \nu | \mathbf{X}_i, \boldsymbol{\gamma}^{(j)})}. \end{aligned}$$

Jednotlivé výrazy přitom snadno vyjádříme, a sice

$$\begin{aligned} f\Big(Y_i | Z_i = 1, \mathbf{X}_i, \boldsymbol{\beta}^{(j)}\Big) &= \mathbb{1}_{[Y_i = 0]}, \\ f\Big(Y_i | Z_i = 0, \mathbf{X}_i, \boldsymbol{\beta}^{(j)}\Big) &= g\Big(Y_i; {\psi_{\mathrm{i}}}^{(j)}, \varphi\Big), \\ \mathsf{P}\Big(Z_i = 1 | \mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big) &= p_i^{(j)}, \\ \mathsf{P}\Big(Z_i = 0 | \mathbf{X}_i, \boldsymbol{\gamma}^{(j)}\Big) &= 1 - p_i^{(j)}, \end{aligned}$$

kde jsme dle (4.1) označili

$${\psi_{\mathrm{i}}}^{(j)} = h_2^{-1}\Big(\mathbf{V}_i^{\mathsf{T}} \boldsymbol{\beta}^{(j)}\Big) \quad \text{a} \quad p_i^{(j)} = \mathrm{logit}^{-1}(\mathbf{W}_i^{\mathsf{T}} \boldsymbol{\gamma}^{(j)}).$$

Po dosazení a úpravách dostáváme vcelku kompaktní vyjádření $Z_i^{(j)}$, a sice

$$Z_i^{(j)} = \begin{cases} \dfrac{1}{1 + g\Big(0; {\psi_{\mathrm{i}}}^{(j)}, \varphi\Big) \exp\Big(-\mathbf{W}_i^{\mathsf{T}} \boldsymbol{\gamma}^{(j)}\Big)}, & \text{pokud } Y_i = 0 \\ 0, & \text{jinak.} \end{cases} \tag{4.6}$$

Nyní je již snadné vyjádřit $\mathcal{Q}\Big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}\Big)$, tedy střední hodnotu logaritmické věrohodnosti $\ell_c(.)$ vzhledem k $\mathbf{Z}|\mathbb{X}, \boldsymbol{\theta}^{(j)}$. Protože veličiny Z_i alternativně rozdělené a v $\mathcal{Q}\Big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}\Big)$ figurují lineárně, stačí je nahradit právě spočtenými pravděpodobnostmi $Z_i^{(j)}$. Dostaneme tak

$$\begin{aligned} \mathcal{Q}\Big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}\Big) &= \mathsf{E}_{\mathbf{Z}|\mathbf{Y}, \mathbb{X}, \boldsymbol{\theta}^{(j)}}\Big[\ell_c(\boldsymbol{\beta}) + \ell_c(\boldsymbol{\gamma}) \Big| \mathbf{Y}, \mathbb{X}, \boldsymbol{\theta}^{(j)}\Big] \\ &= \ell_c\Big(\boldsymbol{\beta} \,|\mathbf{Y}, \mathbb{V}, \mathbf{Z}^{(j)}\Big) + \ell_c\Big(\boldsymbol{\gamma} \,|\mathbf{Y}, \mathbb{W}, \mathbf{Z}^{(j)}\Big), \end{aligned}$$

přičemž předpisy těchto sčítanců jsou následující

$$\ell_c\Big(\boldsymbol{\beta}\,|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big) = \sum_{i=1}^{n}\Big(1 - Z_i^{(j)}\Big)\left[\frac{Y_i\,\psi_{\mathrm{i}} - b(\psi_{\mathrm{i}})}{\varphi} + c(Y_i,\varphi)\right],$$
$$\ell_c\Big(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\Big) = \sum_{i=1}^{n}\Big[\; Z_i^{(j)}\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma} - \log\Big(1 + \mathrm{e}^{\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}}\Big)\Big].$$

M krok Další iterace odhadů regresních koeficientů $\boldsymbol{\beta}^{(j+1)}$ a $\boldsymbol{\gamma}^{(j+1)}$ získáme maximalizací střední hodnoty logaritmické věrohodnosti $\mathcal{Q}(.)$. Vzhledem k separovatelnosti výrazů v logaritmické věrohodnosti úplných dat $\ell_c(.)$ na části, v nichž se vyskytuje pouze koeficient $\boldsymbol{\beta}$, resp. $\boldsymbol{\gamma}$, můžeme tuto maximalizaci provést pro každý vektorový parametr zvlášť. Hledáme tedy

$$\boldsymbol{\beta}^{(j+1)} = \arg\max_{\boldsymbol{\beta}} \ell_c\Big(\boldsymbol{\beta}\,|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big),$$
$$\boldsymbol{\gamma}^{(j+1)} = \arg\max_{\boldsymbol{\gamma}} \ell_c\Big(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\Big).$$

Popišme v krátkosti možný postup:

a) Zderivováním logaritmické věrohodnosti $\ell_c(\boldsymbol{\beta})$ podle $\boldsymbol{\beta}$ získáme skórovou funkci $\mathbf{U}_c(\boldsymbol{\beta})$, z níž ihned dostaneme následující odhadovací rovnici

$$\mathbf{U}_c(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^{n}\Big(1 - \;Z_i^{(j)}\Big)\frac{(Y_i - \mu_i)}{V(\mu_i)h_2'(\mu_i)}\mathbf{V}_i = 0. \tag{4.7}$$

Řešením této rovnice je další přiblížení $\boldsymbol{\beta}^{(j+1)}$.

b) Vzhledem k tomu, že jsme zvolili tutéž linkovou funkci jako v případě modelů ZIP, je konstrukce dalšího přiblížení $\boldsymbol{\gamma}^{(j+1)}$ shodná s postupem uvedeným v předchozí kapitole. Připomeňme proto pouze předpis skórové statistiky

$$\mathbf{U}_c(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\left[\; Z_i^{(j)} - \frac{1}{1 + \mathrm{e}^{-\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}}}\right]\mathbf{W}_i$$

a odhadovací rovnice

$$\sum_{i=1}^{n}\Big[\; Z_i^{(j)} - p_i\Big]\mathbf{W}_i = 0. \tag{4.8}$$

Řešením této rovnice je kýžený odhad $\boldsymbol{\gamma}^{(j+1)}$, jenž je dalším přiblížením maximálně věrohodného odhadu parametru $\boldsymbol{\gamma}$.

Oba kroky opakujeme do konvergence a výsledné odhady označíme $\hat{\boldsymbol{\beta}}$ a $\hat{\boldsymbol{\gamma}}$. Dle vět 4, 5 a 6 uvedených na straně 16 jsou tyto odhady maximálně věrohodnými odhady v ZIM.

Zbývá navrhnout vhodné počáteční přiblížení $\boldsymbol{\beta}^{(0)}$ a $\boldsymbol{\gamma}^{(0)}$. V literatuře jsme nenarazili na žádné doporučení, jak tuto volbu provést, budeme proto postupovat obdobně, jak je v článku Lambert (1992) doporučeno pro *poissonovské modely*

s nadbytečnými nulami. Jelikož jsme zvolili tutéž linkovou funkci $h_1(.) = \text{logit}(.)$, vezměme za $\boldsymbol{\gamma}^{(0)}$ opět vektor nulový až na první složku, která by měla být rovna

$$\gamma_1^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}_{[Y_i=0]} - \mathrm{e}^{\exp(\mathbf{V}_i^{\mathsf{T}} \boldsymbol{\beta}^{(0)})} \right).$$

Připomeňme, že jde o odhad pravděpodobnosti, že Y_i pochází z *perfektního* stavu. V případě $\boldsymbol{\beta}^{(0)}$ zvolme maximálně věrohodný odhad parametru $\boldsymbol{\beta}$ v *modelu bez nulové odezvy*, jehož podkladové rozdělení je taktéž $\mathcal{G}$. Konstrukci tohoto odhadu jsme představili v kapitole 2.2 na straně 19.

5. Negativně binomické modely s nadbytečnými nulami

Negativně binomické modely s nadbytečnými nulami (*zero inflated negative binomial models*, ZINB), jimiž se budeme v této kapitole zabývat, jsou poměrně širokou třídou regresních modelů. Jak již název napovídá, jedná se samozřejmě o *modely s nadbytečnými nulami*, v nichž je podkladovým rozdělením rozdělení negativně binomické.

Již v úvodní kapitole jsme uvedli, že pojmem *negativně binomické rozdělení* bývá označováno vícero různých rozdělení, přičemž nejčastějšími, jak uvádí např. McCullagh a Nelder (1998), jsou tzv. *negativně binomické rozdělení typu 1* (NB1) a *negativně binomické rozdělení typu 2* (NB2). Protože v následujících podkapitolách zkonstruujeme modely ZINB založené na obou z těchto rozdělení, poupravíme tradiční označení těchto modelů na ZINB1 (model ZINB, jehož podkladovým rozdělením je rozdělení NB1), resp. ZINB2 (model ZINB, jehož podkladovým rozdělením je rozdělení NB2). Zkratkou ZINB budeme nadále označovat *negativně binomické modely s nadbytečnými nulami* bez bližší specifikace typu podkladového negativně binomického rozdělení.

Modelům ZINB se věnují například práce Chin a Quddus (2003), Welsh et al. (1996) či Shankar et al. (1997), zabývá se jimi ale o poznání méně autorů než *poissonovskými modely s nadbytečnými nulami* (ZIP), které jsme představili v kapitole 3. Naprostá většina prací přitom předpokládá, že podkladovým rozdělením je rozdělení NB2 (například články Welsh et al. (1996) či Lim et al. (2013)). Někteří autoři (příkladem může být práce Zeileis et al. (2008)) dále předpokládají, že je parametr $\alpha > 0$ pevně daný a využívají tak faktu, že je takto zvolené rozdělení NB2 rozdělením exponenciálního typu. Při konstrukci odhadů regresních koeficientů tak můžeme postupovat obdobně jako v případě ZIP či ZMM. Tímto případem se budeme zabývat v podkapitole 5.1.2.

Konstrukcí odhadů regresních koeficientů v modelu ZINB2 v případě, že parametr α není známý, zmiňují články Welsh et al. (1996) a Cheung (2002), ani jeden nicméně neuvádí podrobnosti. Je samozřejmě možné postupovat standardními metodami maximální věrohodnosti, ty ovšem vedou na značně komplikované výrazy. (V literatuře jsme nenarazili na práci, která by se jimi zabývala. Pro představu však zmiňme například článek Lawless (1987), v němž se autor zabývá maximálně věrohodnými odhady ve standardním negativně binomickém regresním modelu vycházejícím z rozdělení NB2. I v tomto jednodušším případě jsou získávané výrazy značně obtížné.) Protože bychom v dalším chtěli především využít výsledků odvozených v předchozí kapitole, tento případ (tj. modely ZINB2 s neznámým parametrem α) v dalším při konstrukci odhadů regresních koeficientů pomineme.

Jednou z nemnoha prací zabývajících se modely ZINB1 je článek Ridout et al. (2001), v němž jsou závěry odvozeny obecně pro ZINB – konkrétní výrazy platné v případě ZINB1 či ZINB2 čtenář získá volbou parametru. V žádné práci věnované ZINB1 jsme nenarazili na konstrukci odhadů regresních koeficientů. Přitom, jak ukážeme v podkapite 5.2.2, není jejich konstrukce o mnoho obtížnější než v případě ZINB2 se známým parametrem α, a uvedený postup navíc

nevyžaduje znalost tohoto parametru. Při konstrukci odhadů využijeme metodu kvazi-věrohodnosti, kterou jsme představili v úvodní kapitole. Tento postup tak vzhledem ke struktuře varianční funkce rozdělení NB2 nelze v případě ZINB2 s neznámým parametrem α aplikovat.

5.1 Modely ZINB2

Jak jsme uvedli výše, je model ZINB2 se známým parametrem α speciálním případem ZIM. Konstrukce modelu a posléze i odhadů regresních koeficientů tak bude přímou aplikací poznatků uvedených v minulé kapitole. Díky tomu budeme postupovat podstatně rychleji.

Již v závěru podkapitoly 4.1, v níž jsme zavedli modely ZIM, jsme poznamenali, že pro konstrukci *modelů s nadbytečnými nulami* je omezení na podkladová rozdělení exponenciálního typu poněkud umělé a že při konstrukci těchto modelů z jiných rozdělení postupujeme obdobně. V následující podkapitole 5.1.1 proto zavedeme předpis rozdělení odezvy v modelech ZINB2 bez ohledu na to, zda je parametr α známý či nikoli. V podkapitole 5.1.2 se nicméně budeme zabývat konstrukcí odhadů regresních koeficientů pouze v případě, že je parametr α známý.

5.1.1 Definice modelu ZINB2

Mějme tedy k dispozici datovou strukturu, jak byla zavedena na straně 16. Podkladovým rozdělením, z něhož při konstrukci těchto modelů vycházíme, je *negativně binomické rozdělení typu 2* (NB2). Připomeňme, že náhodná veličina Y řídící se tímto rozdělením má hustotu (vůči čítací míře)

$$g(j;\lambda,\alpha) = \exp\left[j\log\left(\frac{\alpha\lambda}{1+\alpha\lambda}\right) - \frac{1}{\alpha}\log(1+\alpha\lambda) + c_\alpha(j,1)\right], \quad kde \quad \begin{array}{l} j \in \mathcal{N}, \\ \lambda > 0, \\ \alpha > 0. \end{array}$$

Tak jako v úvodní kapitole na straně 10 jsme i nyní označili

$$c_\alpha(j,1) = \log\left(\frac{\Gamma(j+\frac{1}{\alpha})}{\Gamma(j+1)\Gamma\left(\frac{1}{\alpha}\right)}\right).$$

Hustotu $g(.)$ jsme zapsali ve tvaru, z něhož je patrné, že jde o hustotu exponenciálního typu (při známém parametru α). Ihned nahlédneme, že dále platí

$$\psi_i = \log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right) \quad \text{a} \quad \varphi = 1.$$

Předpis rozdělení odezvy v modelu ZINB2 tedy je

$$\begin{aligned} \mathsf{P}(Y_i = 0|\mathbf{X}_i) &= p_i + (1-p_i)\exp\left[-\frac{1}{\alpha}\log(1+\alpha\lambda_i)\right] \\ &= p_i + (1-p_i)(1+\alpha\lambda_i)^{-\frac{1}{\alpha}}, \end{aligned}$$

$$\begin{aligned} \mathsf{P}(Y_i = j|\mathbf{X}_i) &= (1-p_i)\exp\left[j\log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right) - \frac{1}{\alpha}\log(1+\alpha\lambda_i) + c_\alpha(j,1)\right] \\ &= (1-p_i)\frac{\Gamma\left(j+\frac{1}{\alpha}\right)}{\Gamma(j+1)\Gamma\left(\frac{1}{\alpha}\right)}(\alpha\lambda_i)^j(1+\alpha\lambda_i)^{-\left(j+\frac{1}{\alpha}\right)}, \qquad j \in \mathcal{N}, \end{aligned} \tag{5.1}$$

přičemž linkovými funkcemi, které použijeme rovněž ve všech následujících podkapitolách věnovaných modelům ZINB, nechť jsou

$$\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \qquad \operatorname{logit}(p_i) = \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}, \tag{5.2}$$

kde $\boldsymbol{\beta} \in \mathcal{R}^v$ a $\boldsymbol{\gamma} \in \mathcal{R}^w$ jsou vektory regresních koeficientů. S touto volbou linkových funkcí, jenž je stejná jako v případě ZIP, se setkáme např. v článku Shankar et al. (1997) či Chin a Quddus (2003). Obě zmíněné práce uvádějí také předpis podmíněné střední hodnoty a podmíněného rozptylu odezvy, a sice

$$\begin{aligned}\mathsf{E}[Y_i|\mathbf{X}_i] &= (1-p_i)\lambda_i,\\ \operatorname{var}(Y_i|\mathbf{X}_i) &= (1-p_i)\lambda_i\Big[1+\lambda_i(\alpha+p_i)\Big].\end{aligned}$$

Poznamenejme, že oba předpisy splňují obecné tvary odvozené na straně 33 a není je obtížné spočítat.

5.1.2 Konstrukce odhadů regresních koeficientů v ZINB2, α známé

Jak jsme již uvedli, modely ZINB2 se známým parametrem α splňují předpoklady, z nichž jsme vyšli při konstrukci maximálně věrohodných odhadů v případě modelů ZMM. Lze tedy postupovat analogicky jako v podkapitole 4.3.

Z předpisu (5.1) sice ihned dostaneme logaritmickou věrohodnost pozorovaných dat, nicméně její předpis je značně komplikovaný a neumožňuje provést maximalizaci zvlášť pro jednotlivé vektorové parametry. Namísto toho proto opět vyjdeme z logaritmické věrohodnosti úplných dat, jejíž předpis můžeme zkráceně zapsat jako

$$\ell_c(\boldsymbol{\gamma}\,,\boldsymbol{\beta};\alpha) = \ell_c(\boldsymbol{\beta};\alpha) + \ell_c(\boldsymbol{\gamma}),$$

přičemž jednotlivé sčítance mají následující předpisy

$$\begin{aligned}\ell_c(\boldsymbol{\beta};\alpha) &= \sum_{i=1}^{n}(1-Z_i)\left[Y_i\log\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right) - \frac{1}{\alpha}\log(1+\alpha\lambda_i) + c_\alpha(Y_i,1)\right],\\ \ell_c(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\Big[Z_i\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma} - \log\Big(1+\exp\Big(\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}\Big)\Big)\Big].\end{aligned} \tag{5.3}$$

Maximálně věrohodné odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ získáme pomocí EM algoritmu. Předpokládejme, že se algoritmus nachází v j-té iteraci, $j \in \mathcal{N}$, a že z dosavadního průběhu máme k dispozici $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$ a $\mathbf{Z}^{(j-1)}$. Při hledání dalšího přiblížení postupujeme takto:

E krok Pravděpodobnost perfektního stavu v j-tém kroku můžeme vyjádřit jako

$$\begin{aligned}Z_i^{(j)} &= \mathsf{P}\Big(Z_i = 1|Y_i,\mathbf{X}_i,\boldsymbol{\beta}^{(j)}\,,\boldsymbol{\gamma}^{(j)}\Big)\\ &= \begin{cases}\dfrac{1}{1+\exp\Big(-\frac{1}{\alpha}\log(1+\alpha\lambda_i^{(j)})\Big)\exp\Big(-\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}^{(j)}\Big)}, & \text{pokud } Y_i = 0,\\ 0, & \text{jinak.}\end{cases}\end{aligned}$$

Dosadíme-li $Z_i^{(j)}$ do (5.3) za veličiny Z_i, získáme střední hodnotu logaritmické věrohodnosti úplných dat v j-tém kroku, a sice

$$\mathcal{Q}\Big(\boldsymbol{\theta}\,,\boldsymbol{\theta}^{(j)}\Big) = \ell_c\Big(\boldsymbol{\beta};\alpha|\mathbf{Y},\mathbf{V},\mathbf{Z}^{(j)}\Big) + \ell_c\Big(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbf{W},\mathbf{Z}^{(j)}\Big),$$

kde jsme opět označili $\boldsymbol{\theta}^{\mathsf{T}} = \Big(\boldsymbol{\beta}^{\mathsf{T}}\,,\boldsymbol{\gamma}^{\mathsf{T}}\Big)$.

M krok Další iterace odhadů regresních koeficientů $\boldsymbol{\beta}^{(j+1)}$ a $\boldsymbol{\gamma}^{(j+1)}$ zvolíme tak, aby byla střední hodnota logaritmické věrohodnosti $\mathcal{Q}(.)$ maximalizována. Tuto maximalizaci můžeme opět provést zvlášť pro každý vektorový parametr. Hledáme tedy

$$\boldsymbol{\beta}^{(j+1)} = \arg\max_{\boldsymbol{\beta}} \ell_c\Big(\boldsymbol{\beta};\alpha|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big),$$
$$\boldsymbol{\gamma}^{(j+1)} = \arg\max_{\boldsymbol{\gamma}} \ell_c\Big(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\Big).$$

a) Derivováním logaritmické věrohodnosti $\ell_c\Big(\boldsymbol{\beta};\alpha|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big)$ podle $\boldsymbol{\beta}$ dostaneme skórovou funkci $\mathbf{U}_c(\boldsymbol{\beta})$ a tedy i odhadovací rovnici

$$\mathbf{U}_c(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^{n}\Big(1 - \ Z_i^{(j)}\Big)\frac{(Y_i-\lambda_i)}{1+\alpha\lambda_i}\mathbf{V}_i = 0.$$

Předpis $\mathbf{U}_c(\boldsymbol{\beta})$ získáme snadno z (4.7) ze strany 37, dosadíme-li za $V(.)$ a $h_1(.)$ výrazy, které jsme uvedli v úvodní kapitole na straně 10. Řešením odhadovací rovnice je další přiblížení $\boldsymbol{\beta}^{(j+1)}$.

b) Obdobně zderivováním logaritmické věrohodnost $\ell_c(\boldsymbol{\gamma})$ získáme skórovou statistiku $\mathbf{U}_c(\boldsymbol{\gamma})$ a následně odhadovací rovnici. Jelikož jsme zvolili tutéž linkovou funkci $h_2(.)$ jako v případě modelů ZMM, je předpis odhadovací rovnice shodný s předpisem (4.8) uvedeným na straně 37, a sice

$$\sum_{i=1}^{n}\Big[Z_i^{(j)} - p_i\Big]\mathbf{W}_i = 0.$$

Jejím řešením je další přiblížení $\boldsymbol{\gamma}^{(j+1)}$ maximálně věrohodného odhadu parametru $\boldsymbol{\gamma}$.

Oba kroky opakujeme do konvergence. Výsledná přiblížení jsou hledanými maximálně věrohodnými odhady regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$.

Zbývá zvolit výchozí přiblížení $\boldsymbol{\beta}^{(0)}$ a $\boldsymbol{\gamma}^{(0)}$. Tak jako v kapitole 4.3 zvolíme za $\boldsymbol{\beta}^{(0)}$ maximálně věrohodný odhad vektoru regresních koeficientů v příslušném modelu zbaveném nulové složky odezvy (viz též kapitola 2.4 věnovaná ZTNB). Za $\boldsymbol{\gamma}^{(0)}$ pak zvolíme vektor nulový až na první složku, která nechť představuje odhad pravděpodobnosti, že odezva Y_i pochází z perfektního stavu.

5.2 Modely ZINB1

Tak jako v podkapitole věnované modelům ZINB2 zavedeme i nyní modely bez ohledu na to, zda je parametr $\alpha > 0$ známý či nikoli. Posléze se budeme zabývat odhady regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$. Konstrukce maximálně věrohodných odhadů je ovšem v tomto případě značně náročná. Namísto toho proto v podkapitole 5.2.2 navrhneme postup, který sice vychází z konstrukce maximálně věrohodných odhadů v ZIM uvedeného v kapitole 4.3, avšak při hledání dalšího přiblížení $\boldsymbol{\beta}^{(j)}, j \in \mathcal{N}$, používá metodu kvazivěrohodnosti (viz též kapitola 1.3 na straně 12).

V případě, že je parametr α neznámý, jsme nuceni tento postup dále modifikovat. Namísto známého parametru budeme uvažovat jeho přiblížení $\alpha^{(j)}, j \in \mathcal{N}$, které budeme budeme konstruovat v rámci M-kroku EM algoritmu. Tímto případem se budeme zabývat v podkapitole 5.2.3.

5.2.1 Definice modelu ZINB1

Mějme k dispozici datovou strukturu (A) zavedenou na straně 16, přičemž podkladovým rozdělením nechť je nyní *negativně binomické rozdělení typu 1* (NB1). Připomeňme, že dle (1.8) ze strany 12 má náhodná veličina Y řídící se tímto rozdělením hustotu (vůči čítací míře)

$$g(j;\lambda,\alpha) = \frac{\Gamma\left(j+\frac{\lambda}{\alpha}\right)}{\Gamma(j+1)\Gamma\left(\frac{\lambda}{\alpha}\right)}\left(\frac{1}{1+\alpha}\right)^{\frac{\lambda}{\alpha}}\left(\frac{\alpha}{1+\alpha}\right)^{j}, \qquad \begin{array}{l} j \in \mathcal{N}_0, \\ \lambda > 0, \\ \alpha > 0. \end{array} \tag{5.4}$$

Hustota $g(.)$ zřejmě není exponenciálního typu a to ani v případě, kdy je parametr α známý. Jelikož budeme v dalším používat metodu kvazi-věrohodnosti, připomeňme, že v regresním modelu, v něm se odezva řídí rozdělením NB1, platí následující dva vztahy

$$V(\mu_i) = (1+\alpha)\mu_i \quad \text{a} \quad \mu_i = \lambda_i,$$

kde je $\alpha > 0$ známá anebo neznámá konstanta. Více jsme se aplikací metody kvazi-věrohodnosti v případě NB1 zabývali na straně 14.

Z předpisu (5.4) snadno dostaneme rozdělení odezvy v ZINB1, a sice

$$\begin{aligned} \mathsf{P}(Y_i = 0|\mathbf{X}_i) &= p_i + (1-p_i)(1+\alpha)^{-\frac{\lambda_i}{\alpha}}, \\ \mathsf{P}(Y_i = j|\mathbf{X}_i) &= (1-p_i)\frac{\Gamma\left(j+\frac{\lambda_i}{\alpha}\right)}{\Gamma(j+1)\Gamma\left(\frac{\lambda_i}{\alpha}\right)}\left(\frac{1}{1+\alpha}\right)^{\frac{\lambda_i}{\alpha}}\left(\frac{\alpha}{1+\alpha}\right)^{j}, \qquad j \in \mathcal{N}. \end{aligned} \tag{5.5}$$

Linkové funkce jsou tytéž jako v případě ZINB2 a jsou tedy uvedeny v (5.2) na straně 41.

Střední hodnotu a rozptyl odezvy v modelu ZINB1 získáme obdobným postupem jako v případě modelů ZIM uvedeným na straně 33, neboť jsme při něm nepoužili předpoklad, že je podkladové rozdělení exponenciálního typu. Celkem tak dostaneme

$$\begin{aligned} \mathsf{E}[Y_i|\mathbf{X}_i] &= (1-p_i)\lambda_i, \\ \mathrm{var}(Y_i|\mathbf{X}_i) &= (1-p_i)\lambda_i\Big[1+\lambda_i(\alpha+p_i)\Big]. \end{aligned}$$

5.2.2 Konstrukce odhadů regresních koeficientů v ZINB1 se známým parametrem α

Při konstrukci těchto odhadů vyjdeme z postupu, jenž jsme uvedli v kapitole 4.3 pro modely ZIM. Jelikož podkladové rozdělení není exponenciálního typu, mírně zmíněný přístup modifikujeme a budeme hledat maximálně kvazi-věrohodné odhady, které jsme v krátkosti představili v úvodní kapitole na straně 14.

Předpis logaritmické věrohodnosti bychom sice ze vztahu (5.5) dostali snadno, nicméně další postup konstrukce maximálně věrohodných odhadů vede na značně obtížné výrazy. Namísto toho proto opět uvažujme logaritmickou věrohodnost úplných dat $\ell_c(.)$. Obdobnými kroky jako v kapitole 4.3 získáme její předpis, a sice

$$\ell_c(\boldsymbol{\beta}\,,\boldsymbol{\gamma};\alpha) = \ell_c(\boldsymbol{\beta};\alpha) + \ell_c(\boldsymbol{\gamma}),$$

kde jsme označili

$$\begin{aligned}\ell_c(\boldsymbol{\beta};\alpha) = \sum_{i=1}^{n}(1-Z_i)\Bigg[\log\left(\frac{\Gamma\left(Y_i+\frac{\lambda_i}{\alpha}\right)}{\Gamma(Y_i+1)\Gamma\left(\frac{\lambda_i}{\alpha}\right)}\right) - \frac{\lambda_i}{\alpha}\log(1+\alpha) \\ + Y_i\log\left(\frac{\alpha}{1+\alpha}\right)\Bigg], \qquad (5.6)\\ \ell_c(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\Big[Z_i\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma} - \log\left(1+\mathrm{e}^{\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}}\right)\Big].\end{aligned}$$

Rozklad logaritmické věrohodnosti $\ell_c(.)$ nám opět umožní provést maximalizaci zvlášť pro parametr $\boldsymbol{\beta}$ a zvlášť pro parametr $\boldsymbol{\gamma}$. Postupovat budeme podle upraveného EM algoritmu. Předpokládejme tedy, že se nacházíme v jeho j-té iteraci a že tak máme k dispozici $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$ a $Z_i^{(j)}$. Dále pokračujme následujícími kroky:

E krok Hledáme $\mathcal{Q}(.)$, střední hodnotu logaritmické věrohodnosti úplných dat $\ell_c(.)$ za aktuálního přiblížení regresních koeficientů. Nejprve je tedy opět třeba spočítat pravděpodobnost *perfektního stavu*. Připomeňme, že v případě modelů ZIM jsme několika úpravami došli k předpisu (4.6) uvedenému na straně 36. Jelikož jsme přitom nikde nevyužili faktu, že je podkladové rozdělení exponenciálního typu, můžeme i nyní využít obecný tvar (4.6). Dosazením

$$g\left(0;\lambda_i^{(j)},\alpha\right) = (1+\alpha)^{-\frac{\lambda_i^{(j)}}{\alpha}}$$

z předpisu podkladové hustoty (5.4) a snadnou úpravou dostaneme

$$Z_i^{(j)} = \begin{cases}\dfrac{1}{1+(1+\alpha)^{-\frac{\lambda_i^{(j)}}{\alpha}}\exp\left(-\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}^{(j)}\right)}, & \text{pokud } Y_i = 0,\\ 0, & \text{jinak.}\end{cases}$$

Snadno dostaneme i předpis $\mathcal{Q}(.)$, nahradíme-li Z_i v předpise logaritmické věrohodnosti úplných dat (5.6) výrazem $Z_i^{(j)}$, a sice

$$\mathcal{Q}\left(\boldsymbol{\theta}\,,\boldsymbol{\theta}^{(j)}\right) = \ell_c\left(\boldsymbol{\beta};\alpha|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\right) + \ell_c\left(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\right),$$

kde jsme označili $\boldsymbol{\theta}^{\mathsf{T}} = \left(\boldsymbol{\beta}^{\mathsf{T}}\,,\boldsymbol{\gamma}^{\mathsf{T}}\right)$.

M krok Další přiblížení $\boldsymbol{\beta}^{(j+1)}$ a $\boldsymbol{\gamma}^{(j+1)}$ zvolíme tak, aby byl výraz $\mathcal{Q}(.)$ maximalizován. Hledáme tedy opět

$$\boldsymbol{\beta}^{(j+1)} = \arg\max_{\boldsymbol{\beta}} \ell_c\Big(\boldsymbol{\beta};\alpha|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big),$$
$$\boldsymbol{\gamma}^{(j+1)} = \arg\max_{\boldsymbol{\gamma}} \ell_c\Big(\boldsymbol{\gamma}|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\Big).$$

Vzhledem k tomu, že je předpis $\ell_c\Big(\boldsymbol{\gamma}|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\Big)$ stejný jako v případě ZIM či ZINB2, získáme předpis $\boldsymbol{\gamma}^{(j+1)}$ řešením odhadovací rovnice (4.8) uvedené na straně 37.

Při hledání přiblížení $\boldsymbol{\beta}^{(j+1)}$ opět zderivujme výraz $\ell_c(\boldsymbol{\beta};\alpha)$, abychom získali skórovou statistiku $\mathbf{U}_c(\boldsymbol{\beta};\alpha)$ a posléze i odhadovací rovnici. Použijeme přitom *řetízkového pravidla*, dle kterého platí

$$\mathbf{U}_c(\boldsymbol{\beta};\alpha) = \frac{\partial \ell_c(\boldsymbol{\beta};\alpha)}{\partial\,\boldsymbol{\beta}} = \frac{\partial \ell_c(\boldsymbol{\beta};\alpha)}{\partial \lambda_i}\frac{\partial \lambda_i}{\partial\,\boldsymbol{\beta}}. \tag{5.7}$$

Vzhledem k volbě linkové funkce (5.2), tedy $\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}$, snadno dostaneme

$$\frac{\partial \lambda_i}{\partial\,\boldsymbol{\beta}} = \exp(\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta})\mathbf{V}_i = \lambda_i\mathbf{V}_i. \tag{5.8}$$

Všimněme si, že parametr λ_i se v předpise $\ell_c(\boldsymbol{\beta};\alpha)$ (výraz (5.6) na straně 44) vyskytuje mj. v argumentu funkce $\Gamma(.)$. Výpočet derivace $\ell_c(\boldsymbol{\beta};\alpha)$ je tak značně komplikovaný. Využijeme však tzv. *digamma funkci*, která je definována jako

$$\psi(z) = \log'(\Gamma(z)) = \frac{\Gamma'(z)}{\Gamma(z)},$$

čímž se nám získaný předpis značně zjednoduší. Vlastnostmi funkce $\psi(z)$ se zabývá např. kniha Abramowitz a Stegun (1972, s. 258-259).

Po sérii úprav dostaneme

$$\begin{aligned}\log'\left(\frac{\Gamma\Big(Y_i+\frac{\lambda_i}{\alpha}\Big)}{\Gamma(Y_i+1)\Gamma\Big(\frac{\lambda_i}{\alpha}\Big)}\right) &= \frac{1}{\alpha}\cdot\left[\frac{\Gamma'\Big(Y_i+\frac{\lambda_i}{\alpha}\Big)}{\Gamma\Big(Y_i+\frac{\lambda_i}{\alpha}\Big)} - \frac{\Gamma'\Big(\frac{\lambda_i}{\alpha}\Big)}{\Gamma\Big(\frac{\lambda_i}{\alpha}\Big)}\right]\\ &= \frac{1}{\alpha}\cdot\left[\psi\left(Y_i+\frac{\lambda_i}{\alpha}\right) - \psi\left(\frac{\lambda_i}{\alpha}\right)\right],\end{aligned}$$

kde derivací je samozřejmě myšlena derivace vzhledem k λ_i. Nyní již snadno získáme předpis $\frac{\partial \ell_c(\boldsymbol{\beta};\alpha)}{\partial \lambda_i}$, a sice

$$\frac{\partial \ell_c(\boldsymbol{\beta};\alpha)}{\partial \lambda_i} = \sum_{i=1}^{n}\frac{(1-Z_i^{(j)})}{\alpha}\cdot\left[\psi\left(Y_i+\frac{\lambda_i}{\alpha}\right) - \psi\left(\frac{\lambda_i}{\alpha}\right) - \lambda_i\log(1+\alpha)\right].$$

Tento výraz spolu s (5.8) dosadíme do předpisu skórové statistiky (5.7) a položíme $\mathbf{U}_c(\boldsymbol{\beta};\alpha) = 0$. Dostaneme tak odhadovací rovnici

$$\sum_{i=1}^{n}\frac{(1-Z_i^{(j)})\lambda_i}{\alpha}\cdot\left[\psi\left(Y_i+\frac{\lambda_i}{\alpha}\right) - \psi\left(\frac{\lambda_i}{\alpha}\right) - \lambda_i\log(1+\alpha)\right]\mathbf{V}_i = 0.$$

Řešení této rovnice označíme $\boldsymbol{\beta}^{(j+1)}$, je totiž dalším přiblížením odhadu regresních koeficientů $\boldsymbol{\beta}$.

Opakování obou kroků algoritmu do konvergence získáme hledané odhady regresních koeficientů.

Zbývá navrhnout počáteční přiblížení regresních koeficientů $\boldsymbol{\beta}^{(0)}$ a $\boldsymbol{\gamma}^{(0)}$. Tak jako dříve, zvolíme za $\boldsymbol{\beta}^{(0)}$ odhad regresních koeficientů $\boldsymbol{\beta}$ v modelu ZTNB1. Tím jsme se v této práci ale nezabývali, čtenář jej však může získat obdobnou modifikací postupu uvedeného v kapitole 2.4. Další možností by mohl být odhad regresních koeficientů $\boldsymbol{\beta}$ v modelu ZIP. Za $\boldsymbol{\gamma}^{(0)}$ pak vezměme opět vektor nulový až na první složku, která představuje odhad pravděpodobnosti perfektního stavu v případě, že $\boldsymbol{\beta}^{(0)}$ je skutečným vektorovým parametrem.

5.2.3 Konstrukce odhadů regresních koeficientů v ZINB1 s neznámým parametrem α

Tak jako v předchozí podkapitole vyjdeme z předpisu logaritmické věrohodnosti úplných dat

$$\ell_c(\boldsymbol{\beta}\,,\boldsymbol{\gamma}\,,\alpha) = \ell_c(\boldsymbol{\beta}\,,\alpha) + \ell_c(\boldsymbol{\gamma}),$$

přičemž předpisy $\ell_c(\boldsymbol{\beta}\,,\alpha)$ a $\ell_c(\boldsymbol{\gamma})$ jsme uvedli v (5.6). Nyní je však parametr α neznámou konstantou, kterou bude třeba nahradit vhodným odhadem. Můžeme přitom postupovat následujícím způsobem, jenž je další modifikací EM algoritmu.

Předpokládejme, že se nacházíme v j-té iteraci, $j \in \mathcal{N}$, popisovaného algoritmu a že tak máme k dispozici $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$, $\alpha^{(j)}$ a $Z_i^{(j-1)}$. Postupujeme následovně:

E krok Tento krok je shodný s E krokem uvedeným v předchozí podkapitole, pouze namísto neznámého parametru α dosazujeme jeho aktuální přiblížení $\alpha^{(j)}$. Získáme tak následující předpis

$$\mathcal{Q}\Big(\boldsymbol{\theta}\,,\boldsymbol{\theta}^{(j)}\Big) = \ell_c\Big(\boldsymbol{\beta};\alpha^{(j)}|\mathbf{Y},\mathbb{V},\mathbf{Z}^{(j)}\Big) + \ell_c\Big(\boldsymbol{\gamma}\,|\mathbf{Y},\mathbb{W},\mathbf{Z}^{(j)}\Big),$$

kde jsme opět označili $\boldsymbol{\theta} = \Big(\boldsymbol{\beta}^{\mathsf{T}}\,,\boldsymbol{\gamma}^{\mathsf{T}}\Big)^{\mathsf{T}}$.

M krok Nyní nejprve zkonstruujeme další přiblížení odhadu parametru α. Vyjdeme přitom z odhadu dispersního parametru φ. Připomeňme, že v případě NB1 platí

$$\varphi = \alpha + 1, \tag{5.9}$$

jak jsme uvedli v úvodní kapitole na straně 14. Za odhad parametru φ v aktuální iteraci EM algoritmu zvolme

$$\varphi^{(j+1)} = \frac{1}{n^\star - v}\sum_{i=1}^{n}(1 - Z_i^{(j)})\frac{\Big(Y_i - \mu_i^{(j)}\Big)^2}{\mu_i}, \quad \text{kde} \quad n^\star = \sum_{i=1}^{n}(1 - Z_i^{(j)}).$$

Odhad vychází z Pearsonova χ^2, jímž odhadujeme dispersní parametr v zobecněných lineárních modelech. Protože musí být současně splněn požadavek nezápornosti parametru α, zvolíme vzhledem k (5.9) další přiblížení tohoto rušivého parametru následovně

$$\alpha^{(j+1)} = \max\Big(0, \varphi^{(j+1)} - 1\Big).$$

Konstrukce dalších přiblížení $\boldsymbol{\beta}^{(j+1)}$ a $\boldsymbol{\gamma}^{(j+1)}$ se shoduje s postupem uvedeným v předchozí podkapitole a získáme je tak řešením následujících rovnic

$$\boldsymbol{\beta}^{(j+1)} = \arg\max_{\boldsymbol{\beta}} \ell_c\Big(\boldsymbol{\beta}; \alpha^{(j+1)} | \mathbf{Y}, \mathbb{V}, \mathbf{Z}^{(j)}\Big),$$
$$\boldsymbol{\gamma}^{(j+1)} = \arg\max_{\boldsymbol{\gamma}} \ell_c\Big(\boldsymbol{\gamma} | \mathbf{Y}, \mathbb{W}, \mathbf{Z}^{(j)}\Big).$$

Výrazy $\ell_c\Big(\boldsymbol{\beta}; \alpha^{(j+1)}\Big)$ a $\ell_c(\boldsymbol{\gamma})$ dostaneme z předpisu (5.6) uvedeného na straně 44 nahrazením parametru α jeho aktuálním přiblížením $\alpha^{(j+1)}$.

Všimněme si, že se parametr α vyskytuje mj. ve jmenovatelích uvedených předpisů a že by tak výše připouštěná možnost $\alpha^{(j+1)} = 0$ mohla být problematická. Využijeme však toho, že pro negativně binomické rozdělení platí, že se pro $\alpha \to 0$ limitně blíží rozdělení Poissonovu. Nastane-li tedy tento případ, získáme další přiblížení $\boldsymbol{\beta}^{(j+1)}$ a $\boldsymbol{\gamma}^{(j+1)}$ a také odhad pravděpodobnosti perfektního stavu $Z_i^{(j+1)}$ v dalším kroku algoritmu dle postupu uvedeného v kapitole 3.3 na straně 28 v případě modelů ZIP.

Oba kroky opakujeme do konvergence, výsledná přiblížení $\hat{\boldsymbol{\beta}}$ a $\hat{\boldsymbol{\gamma}}$ jsou hledanými odhady regresních koeficientů. Zbývá navrhnout vhodná počáteční přiblížení k nastartování algoritmu. V případě parametru α zvolme

$$\alpha^{(0)} = 0.$$

Ve shodě s výše uvedenou poznámkou zvolíme počáteční přiblížení regresorů $\boldsymbol{\beta}^{(0)}$ a $\boldsymbol{\gamma}^{(0)}$ tak, jak jsme v závěru kapitoly 3.3 navrhli pro model ZIP. Vektor $\boldsymbol{\beta}^{(0)}$ buď tedy maximálně věrohodným odhadem regresorů $\boldsymbol{\beta}$ v modelu ZTP, jímž modelujeme vyšetřovaná data zbavená nulové složky odezvy. Vektor $\boldsymbol{\gamma}^{(0)}$ buď tedy nulový až na první složku, která nechť představuje pravděpodobnost *perfektního stavu* v případě, že $\boldsymbol{\beta}^{(0)}$ je skutečnou hodnotou $\boldsymbol{\beta}$.

6. Hradbové modely

Vedle *modelů s nadbytečnými nulami* (ZIM), jimž byly věnovány předešlé kapitoly, existují i další metody snažící se vypořádat s daty zatíženými nadměrným množstvím nul. Již v kapitole 2 jsme poukázali na pokusy oddělit přetíženou nulovou složku a modelovat zbylá data *modely bez nulové odezvy* (ZTM). V této kapitole si představíme další přístup, tzv. *hradbové modely* (*hurdle models*, HM), jenž tuto myšlenku dále rozvíjí. Poprvé byly definovány patrně v článku Mullahy (1986) a zabývala se jimi již řada autorů, zmiňme například práce Gurmu (1998), Min a Agresti (2005) či Saffari et al. (2012).

V případě modelů ZIM jsme předpokládali, že data pochází ze dvou rozdělení. První z nich, označované též jako *perfektní stav*, bylo rozdělení degenerované v nule, druhé, označované též jako *podkladové rozdělení*, pak bylo prvkem třídy rozdělení exponenciálního typu. I *hradbové modely* jsou směsí dvou rozdělení. V jejich případě však s rozdělením degenerovaným v nule mísíme *rozdělení bez nulové odezvy* (viz též kapitola 2). Díky tomu můžeme, jak je poznamenáno např. v práci Min a Agresti (2005), použít modely HM nejen jako alternativu k modelům ZIM, ale i jako *modely s nedostatečným počtem nul* (*zero deflated models*, ZDM), o nichž jsme se zmínili v úvodní kapitole. To samozřejmě v případě modelů ZIM není možné.

Na rozdíl od modelů ZIM pochází nulová pozorování v případě *hradbových modelů* pouze z *perfektního stavu*. Jedná se tak vlastně o speciální případ *modelů s nadbytečnými nulami*. Definice modelů HM je tak přímou aplikací přístupu, jenž jsme použili v případě modelů ZIM. Jelikož *rozdělení bez nulové odezvy* již obecně nejsou děleními exponenciálního typu, odhady regresních koeficientů v HM nezískáme prostou aplikací postupu odvozeného v případě modelů ZIM. Nicméně v dalším uvidíme, že lze s úspěchem použít výsledky kapitol 2.2 a 4.3.

V následujících podkapitolách nejprve na dvou příkladech ilustrujeme rozdílnost přístupů ZIM a HM k datům zatíženým nadbytečnými nulami. Posléze modely HM řádně zavedeme a budeme se věnovat konstrukci odhadů regresních koeficientů v této třídě modelů. Tak jako v případě modelů ZTM a ZIM, i nyní se přitom zaměříme na modely HM založené obecně na diskrétním rozdělení exponenciálního typu a posléze na dva speciální případy – v podkapitole 6.4 na *poissonovský hradbový model* a v podkapitole 6.5 na *negativně binomický hradbový model*. Pro každý vyšetřovaný případ uvedeme obdobně jako v předchozích kapitolách nejprve předpis rozdělení odezvy, její střední hodnoty a rozptylu. Posléze se budeme zabývat konstrukcí odhadů regresních koeficientů, k čemuž i nyní použijeme EM algoritmus.

6.1 Příklad použití modelů HM

Demonstrujme rozdílnost přístupů ZIM a HM na následujících dvou příkladech. Inspirováni aplikací modelů ZIM zmíněnou v práci Shankar et al. (1997) uvažujme následující studii – u náhodně vybraných respondentů se zajímáme o počet dopravních nehod, které způsobili (v pozici řidiče automobilu) v uplynulých deseti letech. Pokud daný člověk není řidičem, bude tento počet zřejmě nulový. Nebýt řidičem tedy odpovídá *perfektnímu stavu*. Avšak i pokud daný re-

spondent řidičem je a aktivně řídí, mohl s jistou pravděpodobností jezdit v uplynulých deseti letech bez nehod. Je tedy zřejmé, že vyšetřovaným datům odpovídá model ZIM.

Naopak datová struktura uvažovaná v článku Hu et al. (2011) odpovídá spíše *hradbovým modelům*. Je sledován počet cigaret vykouřených respondentem za uplynulý měsíc. Je-li zkoumaná osoba nekuřák, je odpověď zřejmě nula (pasivní kouření nepočítáme, stejně jako jsme výše neuvažovali nehody způsobené jedincem nikoli však v pozici řidiče automobilu). To opět odpovídá tzv. *perfektnímu stavu*. Oproti tomu nelze předpokládat, že by kuřák nevykouřil za poslední měsíc ani jednu cigaretu – v takovém případě bychom jej patrně neprávem označili za kuřáka. Nulová odezva tedy oprávněně pochází pouze z *perfektního stavu*, jak předpokládají modely HM.

6.2 Definice

Mějme k dispozici datovou strukturu obdobnou struktuře (A), která byla popsána na straně 16 pro modely ZIM. O hustotě podkladového rozdělení však nyní předpokládejme, že je *hustotou bez nulové odezvy*. Připomeňme, že konstrukci těchto hustot jsme v případě podkladových rozdělení exponencálního typu, na něž se omezíme i v této kapitole, popsali na straně 19. V případě dalších rozdělení postupujeme analogicky. Připomeňme dále, že ve shodě se zmíněným postupem ze strany 19 je nutné předpokládat, že v rozdělení, z něhož při konstrukci *hustoty bez nulové odezvy* vycházíme, je pravděpodobnost nulového pozorování kladná.

Rozdělení odezvy nechť splňuje následující předpis

$$\begin{aligned} &\mathsf{P}(Y_i = 0|\mathbf{X}_i) = p_i, \\ &\mathsf{P}(Y_i = j|\mathbf{X}_i) = (1 - p_i) g_T(Y_i|\, \psi_{\mathrm{i}}\,, \varphi), \end{aligned} \tag{6.1}$$

kde vztah mezi regresory a odezvou určují funkce

$$p_i = \mathrm{logit}^{-1}\left(\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}\right), \qquad \psi_{\mathrm{i}} = h_2^{-1}\left(\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}\right). \tag{6.2}$$

Tak jako v případě modelů ZIM budeme funkce logit(.) a $h_2(.)$ nazývat linkovými funkcemi.

Namísto funkce logit(.) bychom samozřejmě i nyní mohli použít jinou vhodnou funkci s nosičem [0,1], opět se ale omezíme na tuto přirozenou volbu. V pracích věnovaných *hradbovým modelům* jsme se nicméně setkali i s *komplementárním log − log linkem* (například v článku McDowell (2003)), a sice

$$p_i = 1 - \mathrm{e}^{-\,\mathrm{e}^{\mathbf{w}_i^{\mathsf{T}}\boldsymbol{\gamma}}}.$$

6.3 Obecný hradbový model

Nejprve se tedy zabývejme případem, kdy je podkladové rozdělení *hradbového modelu* obecně diskrétní rozdělení exponenciálního typu zbavené nulové složky. V podkapitole 2.2 jsme na straně 19 v (2.3) odvodili předpis hustoty takového rozdělení. Připomeňme její předpis

$$g_T(j;\psi_{\mathrm{i}}\,,\varphi) = \frac{\exp\left(\dfrac{j\,\psi_{\mathrm{i}}}{\varphi} + c(j,\varphi) - c(0,\varphi)\right)}{\exp\left(\dfrac{b(\psi_{\mathrm{i}})}{\varphi} - c(0,\varphi)\right) - 1}, \qquad j \in \mathcal{N}.$$

Dosazením do (6.1) tak ihned máme předpis rozdělení odezvy v *hradbovém modelu*, a sice

$$\begin{aligned} &\mathsf{P}(Y_i = 0|\mathbf{X}_i) = p_i, \\ &\mathsf{P}(Y_i = j|\mathbf{X}_i) = (1-p_i)\frac{\exp\left(\dfrac{j\,\psi_{\mathrm{i}}}{\varphi} + c(j,\varphi) - c(0,\varphi)\right)}{\exp\left(\dfrac{b(\psi_{\mathrm{i}})}{\varphi} - c(0,\varphi)\right) - 1}, \qquad j \in \mathcal{N}, \end{aligned} \tag{6.3}$$

kde vztah mezi regresory a odezvou určují linkové funkce (6.2).

Tak jako v kapitolách 2.2 a 4.1 uvedeme i nyní podmíněnou střední hodnotu a podmíněný rozptyl odezvy. Obdobným postupem jako ve zmíněných kapitolách na straně 20, resp. na straně 33, dostaneme

$$\mathsf{E}[Y_i|\mathbf{X}_i] = \frac{(1-p_i)\mu_i}{1-g(0;\psi_{\mathrm{i}}\,,\varphi)}, \qquad \mathsf{E}\left[Y_i^2|\mathbf{X}_i\right] = \frac{(1-p_i)\mu_i^{[2]}}{1-g(0;\psi_{\mathrm{i}}\,,\varphi)},$$

a z nich dalšími úpravami i

$$\mathrm{var}[Y_i|\mathbf{X}_i] = \frac{1-p_i}{1-g(0;\psi_{\mathrm{i}}\,,\varphi)}\left(\sigma_i^2 + \frac{p_i - g(0;\psi_{\mathrm{i}}\,,\varphi)}{1-g(0;\psi_{\mathrm{i}}\,,\varphi)}\mu_i^{[2]}\right).$$

Přitom μ_i je střední hodnota podkladového rozdělení a, stejně jako ve výše zmiňovaných kapitolách, jsme označili druhý, resp. druhý centrální, moment tohoto rozdělení $\mu_i^{[2]}$, resp. σ_i^2.

Věrohodnost i logaritmická věrohodnost jsou obdobné jako v případě ZIM, které jsme uvedli v (4.3) na straně 35. V našem případě je pouze funkce $g(.)$ nahrazena $g_T(.)$, pro níž zřejmě platí $g_T(0;.) = 0$. Aniž bychom do jejího předpisu dosazovali za hustotu podkladového rozdělení, uveďme předpis logaritmické věrohodnosti

$$\ell(\boldsymbol{\beta}\,,\boldsymbol{\gamma}) = \sum_{i:Y_i=0}\log(p_i) + \sum_{i:Y_i>0}\log(1-p_i) + \sum_{i:Y_i>0}\log(g_T(Y_i;\psi_{\mathrm{i}}\,,\varphi)).$$

Protože na regresorech γ závisí pouze p_i a na regresorech $\boldsymbol{\beta}$ závisí jen ψ_{i}, vidíme, že už nyní lze předpis logaritmické věrohodnosti rozdělit na dvě části a hledat maximálně věrohodné odhady těchto parametrů zvlášť. Přestože bychom mohli postupovat obdobně jako v předchozích kapitolách (připomeňme, že *hradbové modely* lze chápat jako speciální případ *modelů s nadbytečnými nulami*; navíc jsme v

kapitole věnované *negativně binomickým modelům s nadbytečnými nulami* založeným na rozdělení NB1 ukázali, jak lze konstruovat odhady regresních koeficientů v případě, kdy podkladové rozdělení není exponenciálního typu), nabízí se mnohem jednodušší cesta. Ze struktury dat je totiž ihned patrné, která pozorování pocházejí z *perfektního stavu* a která z *podkladového rozdělení*. Definujme proto náhodnou veličinu Z_i předpisem

$$Z_i = \mathbb{1}_{[Y_i=0]}, \quad i = 1, \ldots, n.$$

Logaritmickou věrohodnost pozorovaných dat $(Y_i, \mathbf{X}_i, Z_i), i = 1, \ldots, n$ můžeme vyjádřit opět jako (4.5) na straně 35, a sice

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \ell(\boldsymbol{\beta}) + \ell(\boldsymbol{\gamma}), \tag{6.4}$$

přičemž modifikován byl pouze výraz $\ell(\boldsymbol{\beta})$, jehož předpisem je

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (1 - Z_i) \log(g_T(Y_i; \psi_i, \varphi)) \\ &= \sum_{i=1}^{n} (1 - Z_i) \left[\left(\frac{Y_i \psi_i}{\varphi} + c(Y_i, \varphi) - c(0, \varphi) \right) - \log\left(\mathrm{e}^{\frac{b(\psi_i)}{\varphi} - c(0,\varphi)} - 1 \right) \right].\end{aligned}$$

Připomeňme i předpis $\ell(\boldsymbol{\gamma})$, tedy

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left[Z_i \mathbf{W}_i^{\mathsf{T}} \boldsymbol{\gamma} - \log\left(1 + \mathrm{e}^{\mathbf{W}_i^{\mathsf{T}} \boldsymbol{\gamma}}\right) \right].$$

Vzhledem k separovanosti výrazů závisejících na $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ v (6.4) můžeme hledat jejich odhady odděleně. Derivováním $\ell(\boldsymbol{\beta})$ a následnými úpravami dostaneme skórovou statistiku

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(1 - Z_i^{(j)}\right) \frac{g_T'(Y_i; \psi_i, \varphi)}{g_T(Y_i; \psi_i, \varphi)} \mathbf{V}_i \\ &= \sum_{i=1}^{n} \left(1 - Z_i^{(j)}\right) \left[Y_i - \frac{b'\left(\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}} \boldsymbol{\beta}}\right)}{1 - \exp\left(- \frac{b\left(\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}} \boldsymbol{\beta}}\right)}{\varphi} + c(0, \varphi) \right)} \right] \frac{\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}} \boldsymbol{\beta}}}{\varphi} \mathbf{V}_i.\end{aligned}$$

Řešením odhadovací rovnice $\mathbf{U}\boldsymbol{\beta} = 0$ je kýžený maximálně věrohodný odhad $\hat{\boldsymbol{\beta}}$.

Vzhledem k tomu, že se, jak jsme uvedli výše, předpis $\ell(\boldsymbol{\gamma})$ nezměnil, dostaneme maximálně věrohodný odhad $\hat{\boldsymbol{\gamma}}$ řešením odhadovací rovnice shodné s rovnicí uvedenou v (4.8) na straně 37. Protože nyní nejsou náhodné veličiny Z_i latentní, nahradíme jimi odhady $Z_i^{(j)}$ použité v kapitole věnované *modelům s nadbytečnými nulami*. Získáme tak skórovou funkci logistické regrese, při níž je modelována odezva $\mathbf{Z}$ pomocí regresorů $\mathbb{W}$. Řešením odhadovací této odhadovací rovnice je maximálně věrohodný odhad $\hat{\boldsymbol{\gamma}}$.

6.4 Poissonovský hradbový model

Jak už vyplývá z jejich označení, jsou *poissonovské hradbové modely* (*Poisson hurdle models*, PHM) zřejmě *hradbovými modely*, v nichž je podkladovým rozdělením Poissonovo rozdělení zbavené nulové složky. Spolu s *negativně binomickými hradbovými modely*, jimž je věnována následující podkapitola, patří modely PHM k nejčastěji používaným *hradbovým modelům*. Přesto nejsou v literatuře tolik rozšířeny jako *poissonovské modely s nadbytečnými nulami*, jimž jsou PHM přímou alternativou. Zmiňme zde například práce Miller (2008) či Dalrymple et al. (2003), které se modely PHM zabývají. I samotná myšlenka *hradbových modelů* byla v článku Mullahy (1986) představena právě na modelech PHM.

Konstrukce maximálně věrohodných odhadů v modelech PHM je speciálním případem postupu představeného v předchozí podkapitole. Zde proto uvedeme jen některé důležité kroky.

Rozdělení odezvy v modelech PHM splňuje

$$\begin{aligned}
&\mathsf{P}(Y_i = 0|\mathbf{X}_i) = p_i,\\
&\mathsf{P}(Y_i = j|\mathbf{X}_i) = (1-p_i)\frac{\lambda_i^j}{(\mathrm{e}^{\lambda_i}-1)j!}, \qquad j \in \mathcal{N},
\end{aligned}$$

kde vztah mezi parametry p_i a λ_i a regresory udávají funkce

$$\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \qquad \mathrm{logit}(p_i) = \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}.$$

Vycházeli jsme přitom z předpisu rozdělení odezvy v *poissonovském modelu bez nulové odezvy*, jak bylo uvedeno v (2.3) na straně 21, a z rovnice (6.3) ze strany 50 popisující rozdělení odezvy modelů HM obecně.

Pomineme předpis logaritmické věrohodnosti pozorovaných dat, jenž lze snadno získat z právě uvedeného. Namísto toho uvažujme rovnou logaritmickou věrohodnost pozorovaných dat $(Y_i,\mathbf{X}_i,Z_i), i = 1,\ldots,n$, jejímž předpisem je

$$\begin{aligned}
\ell(\boldsymbol{\beta},\boldsymbol{\gamma}) =& \ell(\boldsymbol{\beta}) + \ell(\boldsymbol{\gamma})\\
=& \sum_{i=1}^{n}\Big[Y_i\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta} - \log\Big(\exp\Big(\mathrm{e}^{\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}}\Big) - 1\Big) - \log(Y_i!)\Big]\\
&+ \sum_{i=1}^{n}\Big[Z_i\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma} - \log\Big(1 + \mathrm{e}^{\mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}}\Big)\Big].
\end{aligned}$$

Veličiny Z_i jsme přitom definovali stejně jako v obecném případě předpisem

$$Z_i = \mathbb{1}_{[Y_i=0]}, \quad i = 1,\ldots,n.$$

Odhadovací rovnice, jejichž řešením jsou hledané maximálně věrohodné odhady regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$, mají tvar.

$$\begin{aligned}
\mathbf{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(1-Z_i)\left[Y_i - \frac{\mathrm{e}^{\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}}}{1-\exp\Big(\mathrm{e}^{\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}}\Big)}\right]\mathbf{V}_i = 0,\\
\mathbf{U}(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}[Z_i - p_i]\mathbf{W}_i = 0.
\end{aligned}$$

6.5 Negativně binomický hradbový model

V dalším speciálním případě *hradbových modelů*, v tzv. *negativně binomickém hradbovému modelu* (*negative binomial hurdle model*, NBHM), je zřejmě podkladovým rozdělením rozdělení negativně binomické zbavené nulové složky. Tímto modelem se poměrně podrobně zabývá práce Saffari et al. (2012), mj. navrhuje odhady regresních koeficientů.Další autoři se o modelech NBHM zmiňují většinou jen krátce jako o možném zobecnění modelů PHM– například práce Min a Agresti (2005) či původní článek představující *hradbové modely*, Mullahy (1986).

Obdobně jako v kapitole 2 věnované *modelům bez nulové odezvy* se i zde omezíme pouze na modely odvozené z *negativně binomického rozdělení typu 2* se známým parametrem α. Toto rozdělení splňuje výše uvedené požadavky a kýžené výsledky tak snadno získáme aplikací výše uvedeného postupu. Lze se domnívat, že bychom v případě modelů NBHM založených na *negativně binomickém rozdělení typu 1* mohli postupovat obdobně jako v kapitole 5.2.2 v případě modelů ZINB založených na tomto rozdělení. Avšak pro omezený rozsah této práce se zde tímto tématem zabývat nebudeme. V dalším proto zkratkou NBHM rozumějme *negativně binomický hradbový model* založený na rozdělení NB2 se známým parametrem α.

Poznamenejme ještě, že je ve zmiňované práci Saffari et al. (2012) navržen postup konstrukce odhadů regresních koeficientů v případě, že je parametr α neznámý (rozdělení, z něhož se při definici modelů NBHM vychází, je rovněž NB2). Postupnými úpravami se autoři dostávají na soustavu dvou rovnic, které již nelze dále zjednodušit, a výsledné odhady posléze získávají aplikací Newtonova-Raphsonova algoritmu.

Nyní však již ke konstrukci maximálně věrohodných odhadů regresních koeficientů v případě modelů NBHM se známým parametrem α. Rozdělení odezvy Y_i v modelech NBHM můžeme vyjádřit jako

$$\begin{aligned}
&\mathsf{P}(Y_i = 0|\mathbf{X}_i) = p_i,\\
&\mathsf{P}(Y_i = j|\mathbf{X}_i) = (1-p_i)\frac{\Gamma\left(j+\frac{1}{\alpha}\right)}{\Gamma(j+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^j \frac{1}{(1+\alpha\lambda_i)^{\frac{1}{\alpha}}-1}, \qquad j \in \mathcal{N},
\end{aligned}$$

kde vztah regresorů na veličin λ_i a p_i vyjadřují stejně jako v případě modelů PHM funkce

$$\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \qquad \operatorname{logit}(p_i) = \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}.$$

Využili jsme přitom předpisu *negativně binomického rozdělení bez nulové odezvy* uvedeného v (2.5) na straně 22 a rovnice (6.3) ze strany 50 popisující rozdělení odezvy HM obecně.

Zavedeme-li náhodné veličiny Z_i opět předpisem

$$Z_i = \mathbb{1}_{[Y_i=0]}, \quad i = 1,\ldots,n,$$

dostaneme po sérii úprav logaritmickou věrohodnost $(Y_i,\mathbf{X}_i,Z_i)$, $i = 1,\ldots,n$, a sice

$$\begin{aligned}\ell(\boldsymbol{\beta}\,,\boldsymbol{\gamma}) &= \ell(\boldsymbol{\beta}) + \ell(\boldsymbol{\gamma}) \\ &= \sum_{i=1}^{n}(1-Z_i)\left[\log\left(\frac{\Gamma\left(Y_i+\frac{1}{\alpha}\right)\alpha^{Y_i}}{\Gamma(Y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\right) + Y_i\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta} - Y_i\log\left(1+\alpha\,\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)\right. \\ &\qquad\qquad \left. - \log\left(\left(1+\alpha\exp\left(\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}\right)\right)^{\frac{1}{\alpha}}-1\right)\right] \\ &\quad + \sum_{i=1}^{n}\left[Z_i\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma} - \log\left(1+\mathrm{e}^{\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}}\right)\right].\end{aligned}$$

Z ní získáme maximálně věrohodné odhady vektorů regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ řešením odhadovacích rovnic

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(1-Z_i)\frac{\left(Y_i-\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)\left(1+\alpha\,\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)^{\frac{1}{\alpha}}-Y_i}{\left(1+\alpha\,\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)\left(\left(1+\alpha\,\mathrm{e}^{\mathbf{V}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)^{\frac{1}{\alpha}}-1\right)}\mathbf{V}_i = 0,$$

$$\mathbf{U}(\boldsymbol{\gamma}) = \sum_{i=1}^{n}[Z_i - p_i]\mathbf{W}_i = 0.$$

7. Simulace

V závěru práce demonstrujeme použití představených modelů ve dvou simulačních studiích. V podkapitole 7.1 se budeme zabývat daty odpovídajícími *poissonovskému modelu s nadbytečnými nulami* (ZIP) s relativně jednoduchou strukturou regresorů. Jednotlivými volbami regresních koeficientů získáme data s různou očekávanou proporcí nadbytečných nul a různou střední hodnotou podkladového rozdělení. Na data posléze aplikujeme modely *poissonovské* (Poiss) a *negativně binomické regrese* (NB), dále pak modely ZIP, ZINB, PHM a NBHM, a srovnáme jejich schopnost odhadovat zvolené regresní koeficienty.

V podkapitole 7.2 přejdeme k poněkud komplikovanější struktuře regresorů, čímž se pokusíme nastínit situaci poněkud více odpovídající reálnému použití zkoumaných modelů. Vzhledem k časové náročnosti jednotlivých výpočtů již nebudeme uvažovat různé volby regresních koeficientů.

Veškeré simulace provedeme v R, za nímž stojí R Core Team (2015). Zdrojové kódy obou simulačních studií jsou k dispozici jako elektronické přílohy v systému SIS. Při zpracování dat zatížených nadbytečnými nulami vyjdeme především z práce Zeileis et al. (2008), kde autoři popisují dostupné knihovny a funkce k modelování dat výše zmíněnými modely. Tabulka 7.1 zachycuje námi použité funkce ke konstrukci bodových odhadů regresních koeficientů v uvedených modelech. Intervalové odhady jsme ve všech vyšetřovaných případech získali z funkce

Tabulka 7.1: Knihovny a funkce použité v R ke konstrukci odhadů regresních koeficientů v uvedených modelech.

Modely	Funkce	Knihovna
Poiss, NB	`glm()`	`stats`
ZIP, ZINB	`zeroinfl()`	`pscl`
PHM, NBHM	`hurdle()`	`pscl`

`confint.default()` (knihovna `stats`). Z nápovědy k této funkci je patrné, že se jedná o Waldovy intervaly spolehlivosti založené na asymptotických vlastnostech maximálně věrohodných odhadů regresních koeficientů.

Poznamenejme ještě, že všechny uvažované balíčky rozumí pod označením *negativně binomické rozdělení* rozdělení, které jsme v předchozích kapitolách nazývali NB2, a rozdělení NB1 v nich není implementováno. V dalším se proto při aplikování *modelů s nadbytečnými nulami* založených na *negativně binomickém rozdělení* omezíme na modely NB2, ZINB2 a NBHM2, přičemž číslovku z jejich názvech již nebudeme uvádět.

Abychom mohli přístupy použitých modelů navzájem srovnat, budeme u *hradbových modelů* namísto odhadů regresních koeficientů $\boldsymbol{\gamma}$ uvádět hodnoty $-\boldsymbol{\gamma}$. Vyplývá to přímo z implementace funkce `hurdle()`, v níž je regresory $\mathbf{W}_i$ modelována pravděpodobnost kladné odezvy, tedy *neperfektního stavu*. Tu můžeme ve shodě s označením uvedeném v kapitole 6 vyjádřit jako $q_i = 1 - p_i$. Vzhledem k volbě linkové funkce $h_1(.)$ (viz (6.2) na straně 49) platí

$$\operatorname{logit}(q_i) = \log\left(\frac{q_i}{1-q_i}\right) = \log\left(\frac{1-p_i}{p_i}\right) = -\operatorname{logit}(p_i) = -\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma}.$$

7.1 Simulační studie 1

V rámci této studie uvažujme datovou strukturu odpovídající *poissonovskému modelu s nadbytečnými nulami*, jenž jsme představili v kapitole 3. Připomeňme, že dle (3.2) uvedené na straně 25 má rozdělení odezvy následující předpis

$$\begin{aligned}\mathsf{P}(Y_i = 0|\mathbf{X}_i) &= p_i + (1-p_i)\,\mathrm{e}^{-\lambda_i}\,,\\ \mathsf{P}(Y_i = j|\mathbf{X}_i) &= (1-p_i)\frac{\mathrm{e}^{-\lambda_i}\,\lambda_i^j}{j!}, \qquad j \in \mathcal{N}.\end{aligned}$$

kde

$$\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \qquad \mathrm{logit}(p_i) = \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}.$$

Naším cílem bude ukázat odlišnosti v odhadech regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$, modelujeme-li taková data některým z následujících modelů – Poiss, NB, ZIP, ZINB, PHM anebo NBHM. Zajímat nás bude rovněž schopnost pokrývat skutečnou hodnotu regresních koeficientů 95% intervaly spolehlivosti (Waldovými) a vůbec to, zda pokus aplikovat daný model na připravená data selže či nikoli.

Pro jednoduchost jsme se rozhodli uvažovat šest nezávislých regresorů, přičemž jak λ_i, střední hodnotu podkladového rozdělení, tak p_i, očekávanou proporci nadbytečných nul v datech, ovlivňují tři z nich. V obou případech se jedná o absolutní člen, regresor generovaný ze spojitého rozdělení a regresor generovaný z diskrétního rozdělení. Na základě výsledků předchozích pokusů jsme zvolili za spojité rozdělení gamma rozdělení, a sice konkrétně

$$V_{i1} \sim \Gamma(8{,}4), \qquad W_{i1} \sim \Gamma(8{,}4)$$

s hustotou

$$f(x) = \frac{4^8}{\Gamma(8)}x^{8-1}\,\mathrm{e}^{-4x} = \frac{4096}{315}x^7\,\mathrm{e}^{-4x}, \qquad x > 0,$$

za diskrétní rozdělení rozdělení alternativní

$$V_{i2} \sim Alt(0.3), \qquad W_{i2} \sim Alt(0.3),$$

přičemž zřejmě platí $\mathsf{P}(V_{i2} = 1) = \mathsf{P}(W_{i2} = 1) = 0.3$. To spolu s následující volbou regresních koeficientů

$$\boldsymbol{\beta} = (\beta_0, 0.15, 0.25)^\mathsf{T}, \qquad \boldsymbol{\gamma} = (\gamma_0, 0.2, 0.3)^\mathsf{T}$$

zajistí, že s pravděpodobností alespoň 95 % platí pro lineární prediktory

$$\mathbf{V}_i^\mathsf{T}\boldsymbol{\beta} - \beta_0 \in \Big(0.14, 0.71\Big), \qquad \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma} - \gamma_0 \in \Big(0.18, 0.90\Big).$$

Nyní již volbou regresních koeficientů β_0 a γ_0 uvedenou v tabulce 7.2 rozlišíme případy, jimž se budeme v dalším věnovat. Každou volbu přitom vyšetříme pro dva rozsahy výběru, $n = 50$ a $n = 500$, abychom mohli sledovat jejich efekt na získané odhady. Přehled scénářů, jimž se v této simulační studii budeme věnovat, obsahuje tabulka 7.3. Kromě volby regresních koeficientů β_0 a γ_0 specifickou pro daný scénář a intervalů spolehlivosti pro λ_i a p_i uvedených již v tabulce 7.2 v ní nalezneme odkazy na tabulky shrnující výsledky simulací dle daného scénáře.

Tabulka 7.2: Alespoň 95% intervaly spolehlivosti pro λ_i a p_i při různých volbách regresních koeficientů $\boldsymbol{\beta}_0$ a $\boldsymbol{\gamma}_0$ v simulační studii 1.

Hodnota koef.	Oček. hodnota	Hodnota koef.	Oček. hodnota
$\beta_0 = 0.20$	$\lambda_i \in$ (1.4, 2.5)	$\gamma_0 = -2.30$	$p_i \in$ (0.1, 0.2)
$\beta_0 = 1.50$	$\lambda_i \in$ (5.0, 8.8)	$\gamma_0 = -0.60$	$p_i \in$ (0.4, 0.6)
$\beta_0 = 2.35$	$\lambda_i \in$ (12.0, 21.1)	$\gamma_0 = 1.30$	$p_i \in$ (0.8, 0.9)

Tabulka 7.3: Přehled nastavení parametrů simulační studie 1 a intervaly, jimž s alespoň 95% pravděpodobností náleží parametry podmíněného rozdělení odezvy.

	Volby regr.		Intervaly spol.		Výsledky
Označení	β_0	γ_0	λ_i	p_i	Tab. (strana)
Scénář A	0.20	−2.30	(1.4, 2.5)	(0.1, 0.2)	A.1 (str. 73)
Scénář B	0.20	−0.60	(1.4, 2.5)	(0.4, 0.6)	A.2 (str. 74)
Scénář C	0.20	1.30	(1.4, 2.5)	(0.8, 0.9)	A.3 (str. 75)
Scénář D	1.50	−2.30	(5.0, 8.8)	(0.1, 0.2)	7.4 (str. 60)
Scénář E	1.50	−0.60	(5.0, 8.8)	(0.4, 0.6)	A.4 (str. 76)
Scénář F	1.50	1.30	(5.0, 8.8)	(0.8, 0.9)	A.5 (str. 77)
Scénář G	2.35	−2.30	(12.0, 21.1)	(0.1, 0.2)	A.6 (str. 78)
Scénář H	2.35	−0.60	(12.0, 21.1)	(0.4, 0.6)	A.7 (str. 79)
Scénář I	2.35	1.30	(12.0, 21.1)	(0.8, 0.9)	A.8 (str. 80)

Zde v této podkapitole se budeme zabývat pouze scénářem D, zbylé výsledky jsou ponechány do přílohy A1 na straně 71 a následujících.

Než se však budeme zabývat výsledky, věnujme se krátce postupu, jímž jsme se při simulacích řídili. Nejprve jsme vygenerovali regresory V_{ij} a W_{ij} z výše definovaných rozdělení. Z nich a ze zvolených regresorů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ jsme dle příslušného scénáře spočítali parametry modelu λ_i a p_i. Pak jsme vygenerovali latentní veličiny $Z_i \sim Alt(p_i)$ určující, zda pochází Y_i z *perfektního* stavu. V takovém případě jsme zvolili $Y_i = 0$, v opačném jsme vygenerovali hodnotu odezvy z Poissonova rozdělení se střední hodnotou λ_i. Na takto vytvořenou datovou strukturu jsme se postupně pokusili aplikovat modely Poiss, NB, ZIP, ZINB, PHM a NBHM, přičemž jsme použili funkce uvedené v tabulce 7.1. Pokud daný pokus zkonvergoval, uložili jsme pro další zpracování jak odhady regresních koeficientů v tomto modelu, tak i příslušné směrodatné odchylky a 95% intervaly spolehlivosti.

Celý proces jsme pro každý ze scénářů zopakovali tisíckrát, načež jsme pro každý model spočítali, kolikrát se nám jej podařilo aplikovat. (Poznamenejme, že za nezkonvergované jsme prohlásili i ty úspěšné případy, kdy byla směrodatná odchylka některého z odhadů, resp. absolutní hodnota odhadu samotného, větší než 100) Ze zkonvergovaných případů jsme dále určili průměrné odhady regresních koeficientů, průměrné směrodatné odchylky, směrodatné odchylky odhadů jednotlivých regresních koeficientů napříč simulacemi a úspěšnost intervalů spolehlivosti v pokrytí skutečných hodnot regresních koeficientů.

Tabulka 7.4 zachycuje souhrnné výsledky v případě **scénáře D**. Přestože je proporce *perfektního stavu* nízká (viz tabulka 7.3), modely ZIP, ZINB, PHM a NBHM zhruba v 8 % případů při rozsah výběru velikosti $n = 50$ nezkonvergovaly. Při rozsahu výběru velikosti $n = 500$ již zkonvergovaly všechny aplikované

modely. Ze zmíněné tabulky 7.4 je ihned patrné, že klasický *poissonovský model* a *negativně binomický model* dávají horší odhady regresních koeficientů, přičemž selhávají především v odhadování absolutního členu. Odhady zbylých dvou regresních koeficientů β_1 a β_2 jsou již v pořádku, nicméně průměrná směrodatná odchylka SE a směrodatná odchylka SE^{sim} jsou ve srovnání s dalšími modely vesměs vyšší.

Zbylé modely (ZIP, ZINB, PHM a NBHM) jsou na tom lépe – zaměříme-li se na regresní koeficienty ovlivňující podkladové rozdělení, jsou odhady všech čtyř modelů relativně přesné už pro rozsah výběru $n = 50$; s rostoucím rozsahem výběru se samozřejmě snížila velikost směrodatných odchylek. Pokrytí skutečných hodnot intervaly spolehlivosti zůstává ve všech případech, až na odhady koeficientu β_2 na hladině 95 %.

U regresních koeficientů $\boldsymbol{\gamma}$ je situace mírně odlišná, v případě rozsahu výběru $n = 50$ nedává žádný z použitých modelů přesvědčivé výsledky (všimněme si především velikosti směrodatných odchylek). Musíme však mít na zřeteli, že očekávaná proporce nadbytečných nul je v tomto scénáři nízká (dle tabulky 7.3 očekáváme, že tvoří pouze $10 - 20$ % pozorování). Nepatrně lepších výsledků přitom dosahují *hradbové modely*. Pro větší rozsah výběru již dostáváme rozumné odhady regresních koeficientů a rozdíly mezi uvažovanými modely se stírají.

Právě zmíněné poznatky můžeme vysledovat i v simulacích dle ostatních scénářů. Poněkud podrobněji se jim budeme věnovat až v příloze na straně 71. Zde v bodech uvedeme některé závěry, které lze v získaných datech vysledovat.

- V případě nižších hodnot λ_i pochází značná část nul v odezvě z *podkladového rozdělení*, což odporuje předpokladům *hradbových modelů*. Přesto tyto modely dávaly při nízkých rozsazích výběru srovnatelné nebo dokonce lepší odhady $\boldsymbol{\gamma}$ než *modely s nadbytečnými nulami*. Výjimkou přitom byl absolutní člen γ_0, který, dle očekávání, zpravidla nedokázaly *hradbové modely* dobře odhadnout. Pro vyšší rozsahy výběru již dávaly *modely s nadbytečnými nulami* lepší odhady těchto regresních koeficientů.

- Se vzrůstající hodnotou λ_i pocházelo z *podkladového rozdělení* méně nul, *hradbové modely* tak odhadovaly regresní koeficienty $\boldsymbol{\gamma}$ vesměs stejně dobře jako *modely s nadbytečnými nulami*. K obdobným odhadům jsme dospěli už ve scénářích, v nichž byly očekávané hodnoty λ_i v intervalu (5.0,8.8).

- *Modely s nadbytečnými nulami* i *hradbové modely* dávaly při větších rozsazích výběru zpravidla obdobně dobré odhady regresních koeficientů $\boldsymbol{\beta}$. Při nižším rozsahu výběru mely *hradbové modely* problém především v případě, kdy byla očekávaná hodnota λ_i nízká a většina pozorování pocházela z *perfektního stavu* (viz scénář C, tabulka A.3). S touto nepříznivou situací, kdy odhadované regresní koeficienty ve skutečnosti ovlivňují pouze nevelký počet pozorování, si o něco lépe poradily oba *modely s nadbytečnými nulami*. V ostatních situacích pak i pro nižší rozsahy výběru byly odhady získané ze všech čtyř zmiňovaných modelů víceméně srovnatelné.

  Je vcelku zajímavé, že s výjimkou absolutního členu β_0 dávaly standardní modely Poiss a NB často obdobně dobré odhady regresních koeficientů $\boldsymbol{\beta}$. Zvlášť pro nižší rozsahy výběru však odhady v modelu NB trpěly většími směrodatnými odchylkami, intervalové odhady v modelu Poiss pak nejméně

často pokrývaly skutečné hodnoty regresních koeficientů. Dle očekávání byly odhady přesnější v případech, kdy byla proporce nadbytečných nul nízká (srov. např. výsledky simulací dle scénáře G (tabulka A.6) s výsledky dle scénáře I (tabulka A.8)).

- Problémy s konvergencí použitých modelů jsme se setkali především v situacích, kdy bylo očekávané množství nadbytečných nul příliš velké či naopak příliš malé. Modely Poiss a NB zpravidla zkonvergovaly ve více případech než zbylé modely. Obecně patrně nelze říci, že jsou v tomto ohledu lepší *hradbové modely* nebo *modely s nadbytečnými nulami* – např. při simulacích dle scénáře A (tabulka A.1) s menším rozsahem výběru (viz tabulka A.1) zkonvergovaly oba používané *modely s nadbytečnými nulami* pouze v cca $60 - 70$ % případů, zatímco aplikování *hradbových modelů* proběhlo téměř bez problémů. Při simulacích dle scénáře C (tabulka A.3) byla situace obrácená, byť rozdíly nebyly zdaleka tak markantní. Při větších rozsazích výběru již všechny používané modely zpravidla zkonvergovaly.

7.2 Simulační studie 2

Nyní se zaměříme na případ s poněkud komplikovanější strukturou regresorů. Nejen že zvýšíme jejich počet, regresory již nadále nebudou nezávislé a navíc některé z nich budou ovlivňovat jak pravděpodobnost *perfektního* stavu tak i střední hodnotu podkladového rozdělení. Tím bude opět rozdělení Poissonovo a model popisující skutečnou strukturu dat bude opět *poissonovský model s nadbytečnými nulami* (ZIP), jak jsme jej představili v kapitole 3. Na vybudovanou strukturu dat, kterou níže představíme, aplikujeme jako v předchozí simulační studii modely Poiss, NB, ZIP, ZINB, PHM a NBHM a srovnáme, jak se výsledky jednotlivých přístupů liší.

Při konstrukci dat se inspirujme reálným příkladem, který je zmíněn v článku Cheung (2002), když na něm demonstruje použití právě modelu ZIP. Úkolem je pokusit se na základě údajů o porodní váze, věku a pohlaví dítěte a několika charakteristik týkajících se jeho matky popsat motorický vývoj dítěte ve věku přibližně 22 měsíců. Ve shodě s běžnou praxí hodnotí rozvinutost motoriky dítěte výsledkem testu, během něhož mělo postavit věž z kostek. Nyní poněkud nepatřičné označení *perfektní* stav je přitom spjato se situací, kdy dítě nezvládne napodobit zadavatele testu ani jednou kostkou.

V práci Cheung (2002) vychází autor z dat *The 1970 British Birth Cohort Study*, kterou jsme však neměli k dispozici. Jelikož je naším cílem především ukázat, jak se chovají výše zmíněné modely na datech s komplikovanější strukturou regresorů, než byla uvažována v simulační studii 1, jeho aplikací ZIP se zde pouze inspirujeme a vygenerujeme obdobnou datovou strukturu.

Budeme uvažovat celkem 7 regresorů, přičemž tři z nich budou diskrétní a zbylé spojité. Charakteristiky rozdělení, z nichž budeme regresory generovat, stejně jako míru jejich korelovanosti přitom odvodíme z datového souboru `Kojeni`, jenž je používán v práci Zvára (2008) a který byl sesbírán pro diplomovou práci Hajná (1995). Datový soubor se sestává z informacích o dětech v prvních měsících jejich života a o jejich rodičích a zvolili jsme je právě pro jejich tématickou blízkost k datům vyšetřovaným v článku Cheung (2002).

Tabulka 7.4: Souhrnné výsledky 1000 simulací dle **scénáře D** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (1.35, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.25, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	1.18	0.22	0.29	0.85	0.05	0.05	0.07	0.88	0.09	0.15	0.21	0.87
NB	1000	1.19	0.31	0.29	0.94	0.05	0.07	0.07	0.95	0.09	0.22	0.21	0.95
ZIP	934	1.34	0.23	0.24	0.95	0.05	0.05	0.05	0.96	0.09	0.16	0.16	0.95
ZINB	929	1.34	0.24	0.24	0.96	0.05	0.05	0.05	0.96	0.09	0.16	0.16	0.95
PHM	937	1.34	0.23	0.23	0.95	0.05	0.05	0.05	0.96	0.09	0.16	0.15	0.96
NBHM	937	1.34	0.23	0.23	0.96	0.05	0.05	0.05	0.96	0.09	0.16	0.15	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	934	-2.52	2.09	2.30	0.96	0.09	0.37	0.47	0.97	0.36	1.67	1.28	0.98
ZINB	929	-2.51	1.81	2.29	0.96	0.09	0.37	0.47	0.97	0.35	1.23	1.09	0.98
PHM	937	-2.28	1.43	1.63	0.95	0.07	0.33	0.41	0.96	0.31	0.95	0.92	0.97
NBHM	937	-2.28	1.43	1.63	0.95	0.07	0.33	0.41	0.96	0.31	0.95	0.92	0.97

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (1.35, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.25, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	1.20	0.07	0.09	0.42	0.05	0.02	0.02	0.87	0.10	0.05	0.06	0.84
NB	1000	1.20	0.10	0.09	0.67	0.05	0.02	0.02	0.97	0.10	0.07	0.06	0.95
ZIP	1000	1.35	0.07	0.07	0.95	0.05	0.02	0.02	0.96	0.10	0.05	0.05	0.94
ZINB	1000	1.35	0.07	0.07	0.95	0.05	0.02	0.02	0.96	0.10	0.05	0.05	0.94
PHM	1000	1.35	0.07	0.07	0.95	0.05	0.02	0.02	0.96	0.10	0.05	0.05	0.94
NBHM	1000	1.35	0.07	0.07	0.95	0.05	0.02	0.02	0.96	0.10	0.05	0.05	0.94

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-2.26	0.41	0.41	0.95	0.10	0.09	0.09	0.95	0.18	0.28	0.28	0.96
ZINB	1000	-2.27	0.41	0.42	0.95	0.10	0.09	0.09	0.95	0.18	0.28	0.28	0.96
PHM	1000	-2.18	0.39	0.39	0.94	0.10	0.09	0.09	0.95	0.17	0.27	0.27	0.96
NBHM	1000	-2.18	0.39	0.39	0.94	0.10	0.09	0.09	0.95	0.17	0.27	0.27	0.96

Tabulka 7.5: Volba regresorů pro druhou simulační studii. Kromě označení regresorů a informace o rozdělení, z něhož bude v dalším generován obsahuje tabulka i název veličiny datového souboru `Kojeni`, z něhož byl daný regresor odvozen. Regresory V_{i0} a W_{i0} představují absolutní člen a nejsou v tabulce zmíněny.

Velič.	Rozdělení	Regresory	Veličina v `Kojeni` [jednotka]
X_{i1}	$N(3.50, 0.50)$	V_{i1}, W_{i1}	Porodní hmotnost [kg]
X_{i2}	$N(0.50, 0.15)$	V_{i2}, W_{i2}	Porodní délka [m]
X_{i3}	$N(1.70, 0.05)$	V_{i3}	Výška matky [m]
X_{i4}	$N(2.60, 0.40)$	W_{i3}	Věk matky [10 let]
X_{i5}	$Alt(0.70)$	V_{i4}, W_{i4}	Používání dudlíku
X_{i6}	$Alt(0.60)$	V_{i5}	Jde o plánované dítě
X_{i7}	$Alt(0.30)$	W_{i5}	Narozeno mimo porodnici

Seznam předpokládaných regresorů včetně rozdělení, z něhož je budeme generovat, obsahuje tabulka 7.5. Předpokládanou míru korelovanosti jednotlivých vysvětlujících veličin zachycuje matice (7.1). Pro první čtyři veličiny je odvozena z datového souboru `kojeni`, zbylé tři veličiny jsou pro jednoduchost zvoleny nezávislé. Celkem je zvolena následovně

$$\operatorname{corr}(\mathbb{X}) = \begin{pmatrix} 1.00 & 0.80 & 0.30 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.80 & 1.00 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.30 & 0.20 & 1.00 & -0.10 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & -0.10 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}. \tag{7.1}$$

To nám v dalším umožní získat veličiny $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ a $\mathbf{X}_4$ transformací

$$(\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4) = \left(\sqrt{0.50}\,\tilde{\mathbf{X}}_1,\sqrt{0.15}\,\tilde{\mathbf{X}}_2,\sqrt{0.05}\,\tilde{\mathbf{X}}_3,\sqrt{0.40}\,\tilde{\mathbf{X}}_4\right),$$

přičemž $\left(\tilde{X}_{i1}, \tilde{X}_{i2}, \tilde{X}_{i3}, \tilde{X}_{i4}\right)^\mathsf{T}$ generujeme z mnohorozměrného normálního rozdělení s parametry

$$\mu = \left(\frac{3.50}{\sqrt{0.50}}, \frac{0.50}{\sqrt{0.15}}, \frac{1.70}{\sqrt{0.05}}, \frac{2.60}{\sqrt{0.40}}\right)^\mathsf{T}, \qquad \Sigma = \begin{pmatrix} 1.00 & 0.80 & 0.30 & 0.00 \\ 0.80 & 1.00 & 0.20 & 0.00 \\ 0.30 & 0.20 & 1.00 & -0.10 \\ 0.00 & 0.00 & -0.10 & 1.00 \end{pmatrix}.$$

Veličiny $\mathbf{X}_5$, $\mathbf{X}_6$ a $\mathbf{X}_7$ získáme jako nezávislé výběry z alternativního rozdělení s výše uvedenými parametry.

Jak jsme již uvedli výše, vybudujeme datovou strukturu tak, aby odpovídala *poissonovskému modelu s nadbytečnými nulami*. Tak jako v kapitole 3 budeme i nyní předpokládat, že vztah mezi regresory a podmíněnou střední hodnotou podkladového rozdělení a pravděpodobností *perfektního* stavu určují následující vztahy

$$\log(\lambda_i) = \mathbf{V}_i^\mathsf{T}\boldsymbol{\beta}, \qquad \operatorname{logit}(p_i) = \mathbf{W}_i^\mathsf{T}\boldsymbol{\gamma}.$$

Zbývá tedy zvolit skutečnou hodnotu regresních koeficientů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$. S přihlédnutím k simulační studii 1 učiníme tuto volbu tak, aby s alespoň 95% pravděpodobností platilo

$$\lambda_i \in (2.00, 6.00), \qquad p_i \in (0.10, 0.20),$$

což odpovídá situaci zachycené ve scénáři D. Zmíněný požadavek přitom splňuje například následující volba regresních koeficientů

$$\boldsymbol{\beta} = (0.4, 0.2, 0.1, 0.3, -0.2, -0.3)^\mathsf{T},$$
$$\boldsymbol{\gamma} = (-2.5, 0.1, 0.2, 0.1, 0.1, -0.1)^\mathsf{T},$$

kterou budeme v dalším uvažovat.

Tak jako v simulační studii 1 rozlišíme několik případů. Ty se však nyní budou lišit pouze rozsahem výběru – postupně vyšetříme situaci, kdy máme k dispozici 50, 250, 500 a 1000 pozorování. Na vygenerovaná data se pokusíme aplikovat opět modely Poiss, NB, ZIP, ZINB, PHM a NBHM a zaznamenáme jak odhady regresních koeficientů a příslušné směrodatné odchylky, tak i schopnost získaných 95% (Waldových) intervalů spolehlivosti pokrývat skutečné hodnoty koeficientů. Celý postup pro každý rozsah výběru 1000 zopakujeme a souhrnné výsledky uložíme do tabulky – jejich přehled dává tabulka 7.6. Zde se budeme v krátkosti zabývat opět pouze jedním případem, a sice tím, při němž je rozsah výběru zvolen $n = 250$. Ostatní souhrnné tabulky se nacházejí v příloze na straně 81 a následující, zde v závěru kapitoly uvedeme pouze krátký komentář.

Tabulka 7.6: Přehled vyšetřovaných případů v simulační studii 2.

Rozsah výběru	Číslo tabulky	Strana
$n = 50$	A.9	81
$n = 250$	7.7	64
$n = 500$	A.10	82
$n = 1000$	A.11	83

Zabývejme se tedy případem, kdy jsme měli k dispozici $n = 250$ pozorování a krátce okomentujme souhrnné výsledky, které dává tabulka 7.7. Všimněme si, že všechny použité modely dávají vesměs dobré odhady regresních koeficientů $\boldsymbol{\beta}$. Poněkud překvapující mohou být odhady $\boldsymbol{\beta}_0$ modely Poiss a NB, které nejsou příliš odlišné od odhadů tohoto regresního koeficientu ostatními modely. Připomeňme však, že očekávaná proporce nadbytečných nul je nízká a většina pozorování tak pochází z *podkladového rozdělení*.

Podstatně zajímavější je situace v případě odhadů regresních koeficientů $\boldsymbol{\gamma}$. Oba *modely s nadbytečnými nulami* dávají v průměru vcelku dobré odhady, byť s velkými směrodatnými odchylkami. Naopak *hradbové modely* mají problém nejen s odhadem absolutního členu γ_0 (toho jsme byli svědky i v několika případech v simulační studii 1), nýbrž i odhady dalších regresních koeficientů jsou nevalné. Jistým vysvětlením může být relativně nízká očekávaná hodnota λ_i – pak samozřejmě nemalé procento nulových pozorování pochází z *podkladového rozdělení*, což odporuje předpokladům PHM i NBHM. V první simulační studii jsme viděli, že se vzrůstající očekávanou hodnotou tohoto parametru se zlepšovala schopnost *hradbových modelů* odhadovat regresní koeficienty $\boldsymbol{\gamma}$.

Z tabulky A.9 vidíme, že pro nižší rozsahy výběrů mají *hradbové modely* a především pak *modely s nadbytečnými nulami* problémy s konvergencí. Všechny použité modely však už pro takto nízký rozsah výběru dávají obdobně dobré odhady regresních koeficientů $\boldsymbol{\beta}$. Očekáváme, že ve vyšetřovaných datech bylo jen několik málo nadbytečných nul, není proto divu, že jsou odhady regresních koeficientů $\boldsymbol{\gamma}$ vesměs tristní.

S roztoucím rozsahem výběru se odhady $\hat{\boldsymbol{\beta}}$ samozřejmě zpřesňují a totéž platí i o odhadech $\hat{\boldsymbol{\gamma}}$ v případě modelů ZIP a ZINB. Jak si nicméně můžeme všimnout v tabulkách A.10 a A.11, ani pro větší rozsahy výběru nedávají *hradbové modely* dobré odhady rogresních koeficientů $\boldsymbol{\gamma}$. Připomeňme, že v první simulační studii jsme se s tímto problémem setkávali především u absolutního členu γ_0; s rostoucím rozsahem výběru navíc daný problém zpravidla vymizel.

Tabulka 7.7: Souhrnné výsledky 1000 simulací sim. studie 2 pro rozsah výběru $n = 250$. Jsou uvedeny počty úsp. pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 250$; $\boldsymbol{\beta} = (0.40, 0.20, 0.10, 0.30, -0.20, -0.30)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.50, 0.10, 0.20, 0.10, 0.10, -0.10)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	0.31	0.32	0.40	0.88	0.19	0.08	0.11	0.86	0.07	0.15	0.20	0.84
NB	1000	0.31	0.44	0.40	0.96	0.19	0.12	0.11	0.96	0.07	0.21	0.20	0.95
ZIP	1000	0.39	0.34	0.33	0.96	0.20	0.09	0.09	0.95	0.10	0.16	0.16	0.94
ZINB	996	0.38	0.35	0.33	0.96	0.20	0.09	0.09	0.96	0.10	0.16	0.16	0.94
PHM	1000	0.39	0.34	0.33	0.96	0.20	0.09	0.09	0.95	0.10	0.16	0.16	0.95
NBHM	1000	0.38	0.35	0.33	0.96	0.20	0.09	0.09	0.95	0.10	0.16	0.16	0.95

Model	Zkonv.	$\hat{\beta}_3$	$\mathrm{SE}_{\hat{\beta}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_3}$	$\mathrm{q}_{\hat{\beta}_3}$	$\hat{\beta}_4$	$\mathrm{SE}_{\hat{\beta}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_4}$	$\mathrm{q}_{\hat{\beta}_4}$	$\hat{\beta}_5$	$\mathrm{SE}_{\hat{\beta}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_5}$	$\mathrm{q}_{\hat{\beta}_5}$
Poiss	1000	0.30	0.16	0.21	0.87	-0.21	0.07	0.10	0.85	-0.30	0.07	0.09	0.88
NB	1000	0.30	0.22	0.21	0.97	-0.21	0.10	0.10	0.95	-0.30	0.10	0.09	0.97
ZIP	1000	0.30	0.17	0.17	0.96	-0.20	0.08	0.08	0.94	-0.30	0.07	0.07	0.94
ZINB	996	0.30	0.17	0.17	0.96	-0.20	0.08	0.08	0.95	-0.30	0.07	0.07	0.95
PHM	1000	0.30	0.17	0.17	0.96	-0.20	0.08	0.08	0.94	-0.30	0.07	0.07	0.95
NBHM	1000	0.30	0.18	0.17	0.96	-0.20	0.08	0.08	0.95	-0.30	0.07	0.07	0.95

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-2.72	1.72	1.82	0.94	0.12	0.49	0.50	0.95	0.17	0.89	0.91	0.95
ZINB	996	-2.76	1.74	1.87	0.94	0.13	0.50	0.51	0.95	0.17	0.90	0.92	0.95
PHM	1000	1.89	1.38	1.45	0.12	0.04	0.40	0.41	0.95	-0.09	0.73	0.76	0.93
NBHM	1000	1.89	1.38	1.45	0.12	0.04	0.40	0.41	0.95	-0.09	0.73	0.76	0.93

Model	Zkonv.	$\hat{\gamma}_3$	$\mathrm{SE}_{\hat{\gamma}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_3}$	$\mathrm{q}_{\hat{\gamma}_3}$	$\hat{\gamma}_4$	$\mathrm{SE}_{\hat{\gamma}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_4}$	$\mathrm{q}_{\hat{\gamma}_4}$	$\hat{\gamma}_5$	$\mathrm{SE}_{\hat{\gamma}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_5}$	$\mathrm{q}_{\hat{\gamma}_5}$
ZIP	1000	0.11	0.32	0.33	0.95	0.15	0.46	0.48	0.96	-0.14	0.47	0.47	0.96
ZINB	996	0.11	0.33	0.34	0.95	0.15	0.47	0.48	0.96	-0.14	0.47	0.48	0.96
PHM	1000	-0.10	0.27	0.28	0.90	-0.24	0.39	0.40	0.90	0.11	0.38	0.39	0.93
NBHM	1000	-0.10	0.27	0.28	0.90	-0.24	0.39	0.40	0.90	0.11	0.38	0.39	0.93

Závěr

S daty zatíženými nadbytečnými nulami se setkáváme v mnoha oblastech, v nichž je předmětem zájmu nějaká celočíselná veličina – od sledování počtů živočichů ohrožených druhů až po ukazatele kvality výrobních procesů. V uplynulých více než 50 letech bylo navrženo několik způsobů, jak se s modelováním takových dat vypořádat. Byla tak definována celá řada modelů, které souhrnně řadíme do třídy *modelů s upraveným počtem nul* (*zero modified models*, ZMM). V této práci jsme se zabývali jejími třemi podtřídami, a sice *modely bez nulové odezvy* (*zero truncated models*, ZTM), *modely s nadbytečnými nulami* (*zero inflated models*, ZIM) a *hradbovými modely* (*hurdle models*, HM). Modely každé z uvedených tříd jsme řádně definovali a posléze jsme se zaměřili na konstrukci odhadů regresních koeficientů.

První pokusy modelovat data s nadbytečnými nulami vedly na modely ZTM. Stačilo totiž odstranit nulovou složku a modelovat pouze zbylá data. Konstrukcí modelů této podtřídy ZMM jsme se zabývali v kapitole 2. Uvedených poznatků pak využívají i *hradbové modely*, jimž je věnována kapitola 6.

Obě podtřídy modelů ZMM uvažující nulovou složku odezvy (tj. modely ZIM a HM) předpokládají, že data zatížená nadbytečnými nulami pocházejí ze směsi dvou rozdělení – z rozdělení degenerovaného v nule, tzv. *perfektního stavu*, a z nějakého vhodného, tzv. *podkladového*, rozdělení. Modely HM lze přitom chápat jako speciální případ ZIM, kdy je za podkladové rozdělení voleno rozdělení bez nulové složky. Konstrukce odhadů regresních koeficientů, jak jsme ukázali v kapitole 6.3, se však v případě HM značně zjednoduší.

Stojíme-li v praxi před úkolem modelovat počty událostí či jevů, používáme běžně poissonovský regresní model. Obdobně tomu je i v případě modelů ZMM – patrně nejčastější volbou *podkladového rozdělení*, s níž jsme se v literatuře setkali, je rozdělení Poissonovo. Rovněž článek Lambert (1992), v němž byly poprvé představeny modely ZIM, popisuje použité metody právě na modelech založených na Poissonově rozdělení, na tzv. *poissonovských modelech s nadbytečnými nulami* (*zero inflated Poisson models*, ZIP). V kapitole 3 jsme představili mj. postup, jímž jsou v práci Lambert (1992) konstruovány maximálně věrohodné odhady regresních koeficientů v modelech ZIP. Tento postup jsme v následujících kapitolách zobecnili a použili ke konstrukci odhadů i v dalších modelech ZIM.

Poissonovské regresní modely patří do obecnější třídy, tzv. *zobecněných lineárních modelů*, v nichž se odezva řídí rozdělením *exponenciálního typu*. Tím jsme se v této práci inspirovali a v každé z výše uvedených podtříd modelů ZMM (tj. v modelech ZTM, ZIM a HM) jsme zavedli model vycházející obecně z rozdělení exponenciálního typu. Vzhledem k praktickým aplikacím modelů ZMM, jimiž se zpravidla modelují počty událostí či jevů, jsme se přitom omezili na diskrétní rozdělení náležící do této třídy. Vedle řádné definice takto zvolených modelů jsme uvedli i postup, jímž lze v daných modelech konstruovat maximálně věrohodné odhady regresních koeficientů. Jelikož jsme v literatuře vesměs nenarazili na práce zobecňující modely ZMM tímto směrem, modifikovali jsme postup běžně používaný v případě modelů dané podtřídy ZMM založených na Poissonově či negativně binomickém rozdělení.

Přirozenou a hojně používanou alternativou poissonovských regresních modelů v případě, kdy jsou data zatížena nadměrnou disperzí, jsou modely založené na negativně binomickém rozdělení. V literatuře se však setkáváme s jistou nekonzistencí, když je tímto pojmem označováno hned několik rozdělení. V úvodní kapitole jsme popsali dvě patrně nejrozšířenější z nich, nazývané též *negativně binomické rozdělení typu 1* (NB1), resp. *typu 2* (NB2). Rozdělení NB2, z něhož vychází většina prací věnovaných modelům ZMM založených na negativně binomickém rozdělení, je přitom při známém parametru α rozdělením exponenciálního typu. Výsledky jsme tak snadno získali aplikací obecného přístupu k modelům dané podtřídy ZMM založených na rozdělení exponenciálního typu.

Rozdělení NB1 naopak rozdělením exponenciálního typu není a to ani v případě, kdy je parametr α znán. Předpis rozdělení odezvy v modelech založených na NB1 je navíc komplikovanější než v případě NB2 a v literatuře se s výsledky o modelech ZMM založených na NB1 setkáváme jen zřídkakdy. V kapitole 5.2 jsme však ukázali, že modifikací původního algoritmu pro modely ZIM založené na rozdělení exponenciálního typu dokážeme konstruovat odhady regresních koeficientů i v případě modelů ZINB1 (ať už je parametr α znán či nikoliv). Upravený postup přitom používá v M-kroku EM algoritmu ke konstrukci přiblížení odhadu $\hat{\boldsymbol{\beta}}$ metodu kvazi-věrohodnosti. Lze se navíc domnívat, že obdobným přístupem získáme odhady regresních koeficientů i v dalších případech – ať už v modelech ZTM či HM založených na rozdělení NB1, resp. obecně v modelech ZMM založených na dalších rozděleních splňujících předpoklady použitých metod. Vzhledem k omezenému rozsahu této práce jsme se tímto tématem více nezabývali, domníváme se však, že je hodno další pozornosti.

V závěru práce jsme na dvou simulačních studiích demonstrovali použití představených modelů a srovnali jsme získané výsledky se standardními regresními modely (omezili jsme se přitom na modely založené na Poissonově a negativně binomickém rozdělení typu 2, neboť patří zdaleka k nejrozšířenějším modelům). V obou případech odpovídala struktura vyšetřovaných dat předpokladům *poissonovských modelů s nadbytečnými nulami*. V první studii je na devíti scénářích ukázáno, jak se mění schopnost uvažovaných modelů odhadovat regresní koeficienty při různých proporcích nadbytečných nul a různých středních hodnotách podkladového rozdělení. Výsledky této studie jsou shrnuty na straně 58 a následující. Druhá simulační studie, jejíž výsledky jsou shrnuty na straně 62, je pak věnována modelování dat s komplikovanější strukturou, jenž je inspirována reálným použitím modelů ZMM.

Literatura

ABRAMOWITZ, M. a STEGUN, I. A. *Handbook of mathematical functions*. New York : Dover Publications, 9. vyd., 1972. ISBN 0486612724.

ANDĚL, J. *Základy matematické statistiky*. Praha : Matfyzpress, 3. vyd., 2011. ISBN 978-80-7378-162-0.

CHEUNG, Y. B. Zero-inflated models for regression analysis of count data. *Statistics in Medicine*. 2002, 21, 10, s. 1461–1469.

CHIN, H. C. a QUDDUS, M. A. Modeling Count Data with Excess Zeroes. *Sociological Methods*. 2003, 32, 1, s. 90–116. doi: 10.1177/0049124103253459.

COHEN, A. C. Estimating the Parameter in a Conditional Poisson Distribution. *Biometrics*. 1960, 16, 2.

DALRYMPLE, M. L., HUDSON, I. L. a FORD, R. P. K. Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS. *Computational Statistics*. 2003, 41, 3-4, s. 491–504. ISSN 01679473. doi: 10.1016/S0167-9473(02)00187-1.

DEMPSTER, A. P., LAIRD, N. M. a RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*. 1977, 39, 1, s. 1–38.

DENG, D. a PAUL, S. R. Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica*. 2005, 15, 1, s. 257–276. ISSN 1017-0405.

DIETZ, E. a BÖHNING, D. On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics*. 2000, 34, 4, s. 441–459.

GHOSH, S. et al. The k-ZIG. *Biometrics*. 2012, 68, 3, s. 878–885. ISSN 0006341x. doi: 10.1111/j.1541-0420.2011.01729.x.

GREEN, P. J. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society*. 1984, 46, 2, s. 149–192. ISSN 1369-7412.

GROGGER, J. T. a CARSON, R. T. Models for Truncated Counts. *Journal of applied econometrics*. 1991, 6, 3, s. 225–238. ISSN 0883-7252. doi: 10.1002/jae.3950060302.

GURMU, S. Generalized hurdle count data regression models. *Economics Letters*. 1998, 58, 3, s. 263–268. ISSN 01651765. doi: 10.1016/S0165-1765(97)00295-4.

HAJNÁ, P. Vliv biosociálních faktorů na délku kojení a závislost vybraných antropometrických charakteristik na způsobu výživy dítěte v prvních šesti měsících života. Diplomová práce, Přírodovědecká fakulta Univerzity Karlovy v Praze, Praha, 1995.

HALL, D. B. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*. 2000, 56, 4, s. 1030–1039. ISSN 0006341x.

HEILBRON, D. C. Zero-Altered and other Regression Models for Count Data with Added Zeros. *Biometrical Journal*. 1994, 36, 5, s. 531–547. ISSN 03233847. doi: 10.1002/bimj.4710360505.

HEYDE, C. a MORTON, R. Quasi-Likelihood and Generalizing the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996, 58, 2, s. 317–327.

HU, M.-C., PAVLICOVA, M. a NUNES, E. V. Zero-Inflated and Hurdle Models of Count Data with Extra Zeros. *The American Journal of Drug and Alcohol Abuse*. 2011, 37, 5, s. 367–375. ISSN 0095-2990. doi: 10.3109/00952990.2011.597280.

LAMBERT, D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*. 1992, 34, 1.

LAWLESS, J. F. Negative binomial and mixed Poisson regression. *The Canadian Journal o f Statistics*. 1987, 15, 3, s. 209–225.

LEE, A. H. et al. Truncated negative binomial mixed regression modelling of ischaemic stroke hospitalizations. *Statistics in Medicine*. 2003, 22, 7, s. 1129–1139. ISSN 0277-6715. doi: 10.1002/sim.1419.

LIM, H. K., SONG, J. a JUNG, B. C. Score tests for zero-inflation and overdispersion in two-level count data. *Computational Statistics*. 2013, 61, s. 67–82.

MCCULLAGH, P. a NELDER, J. A. *Generalized linear models*. Boca Raton : Chapman, 2. vyd., 1998. ISBN 04-123-1760-5.

MCDOWELL, A. From the help desk: hurdle models. *The Stata Journal*. 2003, 3, 2, s. 178–184.

MCLACHLAN, G. J. a KRISHNAN, T. *The EM algorithm and extensions*. Hoboken, N. J. : Wiley-Interscience, 2. vyd., 2008. ISBN 04-712-0170-7.

MILLER, J. M. Comparing Poisson, Hurdle, and Zip Model Fit under Varying Degrees of Skew and Zero-Inflation. Disertační práce, University of Florida, University of Florida, 2008.

MIN, A. a CZADO, C. Testing for zero-modification in count regression models. *Statistica Sinica*. 2010, 20, 1, s. 323–341.

MIN, Y. a AGRESTI, A. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*. 2005, 5, 1, s. 1–19. ISSN 1471082x. doi: 10.1191/1471082X05st084oa.

MULLAHY, J. Specification and testing of some modified count data models. *Journal of Econometrics*. 1986, 33, 3, s. 341–365.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

RAO, A. K. a SUMATHI, K. On Estimation and tests for zero inflated regression models. In *InterStat*, [Blacksburg, Va., 2009. Virginia Tech].

RIDER, P. R. Truncated Poisson Distributions. *Journal of the American Statistical Association*. 1953, 48, 264, s. 826–830.

RIDOUT, M., HINDE, J. a DEMÉTRIO, C. G. B. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics*. 2001, 57, 1, s. 219–223.

SAFFARI, S., ADNAN, R. a GREENE, W. Hurdle negative binomial regression model with right censored count data. *SORT*. 2012, 36, 2, s. 181–194. ISSN 1696-2281.

SEN, P. K. a MOTTA SINGER, J. *Large sample methods in statistics*. New York : Chapman, 1993. ISBN 04-120-4221-5.

SHANKAR, V., MILTON, J. a MANNERING, F. Modeling accident frequencies as zero-altered probability processes. *Accident Analysis*. 1997, 29, 6, s. 829–837. ISSN 00014575. doi: 10.1016/S0001-4575(97)00052-3.

SINGH, J. A Characterization of Positive Poisson Distribution and its Statistical Application. *SIAM Journal on Applied Mathematics*. 1978, 34, 3, s. 545–548. ISSN 0036-1399. doi: 10.1137/0134043.

BROEK, J. A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics*. 1995, 51, 2, s. 738–743. doi: 10.2307/2532959.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*. 1974, 61, 3, s. 439–447. ISSN 0006-3444. doi: 10.1093/biomet/61.3.439.

WELSH, A. et al. Modelling the abundance of rare species. *Ecological Modelling*. 1996, 88, 1-3, s. 297–308.

ZEILEIS, A., KLEIBER, C. a JACKMAN, S. Regression Models for Count Data in R. *Journal of statistical software*. 2008, 27, 8.

ZVÁRA, K. *Regrese*. Praha : Matfyzpress, 1. vyd., 2008. ISBN 978-80-7378-041-8.

Seznam tabulek

Přílohy

A1 Výsledky simulační studie 1

Na následujících stranách uvedeme souhrnné výsledky simulačních studií rozdělené do několika tabulek podle jednotlivých nastavení. Nejprve však uvedeme krátký komentář, v němž se pokusíme podtrhnout některé odlišnosti, které jsou ve výsledcích simulací dle jednotlivých scénářů patrné. Budeme se přitom odkazovat na očekávané hodnoty λ_i a p_i, které pro jednotlivé scénáře shrnuje tabulka 7.3 uvedená na straně 57.

Scénář A (tabulka A.1) Díky nízké proporci nadbytečných nul dostáváme poměrně dobré odhady regresních koeficientů $\boldsymbol{\beta}$ již pro nižší rozsahy výběrů (pro model Poiss a NB samozřejmě s výjimkou absolutního členu).

Všimněme si, že modely s nadbytečnými nulami (tj. ZIP a ZINB) mají problémy s konvergencí. Pro nižší rozsah výběru navíc nedávají dobré odhady koeficientů $\boldsymbol{\gamma}$.

Vzhledem k nižší hodnotě λ_i obsahují pozorování nepocházející z *perfektního stavu* nižší hodnoty a tedy i větší množství nul, než tomu bylo v případě scénáře D. To je v rozporu s předpoklady *hradbových modelů*. Oba zkoumané modely, PHM a NBHM, tak mají problém především s odhadem absolutního členu γ_0. Pro větší rozsahy výběru však dávají lepší odhady ostatních regresních koeficientů $\boldsymbol{\gamma}$ než *modely s nadbytečnými nulami*.

Scénář B (tabulka A.2) Na rozdíl od scénáře A jsme zvýšili proporci nadbytečných nul. Značně se tím zmenšil problém modelů zkonvergovat – s výjimkou ZINB již všechny pokusy zkonvergovaly ve více než 95 % případů.

Kromě absolutního členu β_0 jsou odhady regresních koeficientů $\boldsymbol{\beta}$ napříč modely obdobně dobré, a to pro oba zkoumané rozsahy výběru. Co se regresních koeficientů $\boldsymbol{\gamma}$ týče, jsme opět svědky jevu, na nějž jsme již poukázali v případě scénáře A – modely ZIP a ZINB dávají horší odhady pro rozsah výběru $n = 50$ než *hradbové modely*, pro větší rozsah výběru je již situace srovnatelná. *Hradbové modely* navíc i nyní dávají nesprávné odhady absolutního členu γ_0 (pro nižší rozsah výběru to platí i pro oba *modely s nadbytečnými nulami*).

Scénář C (tabulka A.3) Připomeňme, že očekávaná proporce nadbytečných nul byla zvolena mezi 80-90 %, odhady regresních koeficientů $\boldsymbol{\beta}$ tak pro nižší rozsahy výběru vesměs selhávají. Pro rozsah výběru $n = 50$ (vyjma absolutního členu β_0) dává nejlepší výsledky patrně model Poiss. Jinak zůstává vše jako v předchozím případě, pouze se nadále stírá rozdíl mezi odhady γ_0 *hradbovými modely* a *modely s nadbytečnými nulami*.

Při nízkém rozsahu výběru mají všechny zkoumané modely problémy s konvergencí – modely Poiss a NB zkonvergovaly ve zhruba 85 % případů, zbylé modely pak pouze v 50-65 % případů. Nicméně s rostoucím rozsahem výběru vzroste i očekávaný počet pozorování pocházejících z *podkladového rozdělení* a tyto problémy vymizí.

Scénář E (tabulka A.4) Nastavení regresních koeficientů $\boldsymbol{\beta}$ odpovídá scénáři D, zvětšena byla pouze očekávaná proporce nadbytečných nul. Závěry jsou obdobné jako v předchozích případech, všimněme si tedy především toho, že rozdíl mezi *modely s nadbytečnými nulami* a *hradbovými modely* se zde již téměř nevyskytuje a nesetkáme se s ním ani v dalších scénářích. To proto, že očekávaná střední hodnota podkladového rozdělení je již natolik velká, že většina nulových pozorování pochází skutečně z *perfektního* stavu.

Scénář F (tabulka A.5) Tak jako v případě scénáře C jde o případ s největší očekávanou proporcí nadbytečných nul a ke konstrukci odhadů je proto rovněž zapotřebí větší rozsah výběru. Pro $n = 50$ mají všechny uvažované modely problémy s konvergencí, které se se vzrůstajícím n vytrácejí. Všimněme si také, že získané odhady mají vesměs větší (průměrné) směrodatné odchylky oproti předchozímu případu. Modely Poiss a NB mají problém nejen s odhadem absolutního členu β_0, jimi stanovené intervaly spolehlivosti pokryly skutečné hodnoty pouze v 60 %, resp. v 85 % případů a to i při rozsahu výběru velikosti $n = 500$. Ostatní modely dávají srovnatelné odhady.

Scénář G (tabulka A.6) Ve zbylých třech scénářích je separace *perfektního* a *neperfektního* stavu největší. Díky tomu získáváme dobré odhady $\boldsymbol{\beta}$ u všech modelů (v případě *poissonovského* a *negativně binomického modelu* opět s výjimkou absolutního členu) a to i při menším zkoumaném rozsahu výběru. Stejně jako u scénářů A a D způsobuje malá očekávaná proporce nadbytečných nul to, že je k přesnějším odhadům $\boldsymbol{\gamma}$ zapotřebí větších rozsahů výběru a že pro nižší rozsahy výběru použité modely v některých případech nezkonvergují.

Scénář H (tabulka A.7) Závěry jsou obdobné jako v předchozím případě. Odhady $\boldsymbol{\gamma}$ jsou při stejně velkém rozsahu výběru přesnější, než tomu bylo u scénáře G, nicméně toto zlepšení jsme mohli pozorovat i mezi scénáři A a B a také D a E (tehdy došlo ke stejnému zvýšení očekávané proporce nadbytečných nul).

Scénář I (tabulka A.8) Očekávaná proporce nadbytečných nul je opět již natolik velká, že v případě $n = 50$ selhává asi 20 % pokusů o aplikaci modelu na data (v případě modelů Poiss a NB jich selhává pouze asi 10 % případů, nicméně obdržené odhady regresních koeficientů $\boldsymbol{\beta}$ jsou o něco horší než odhady získané ostatními modely). Pro větší rozsahy výběrů se situace poněkud zlepšuje, problémy jsme zaznamenali pouze v případě modelu NBHM.

Tabulka A.1: Souhrnné výsledky 1000 simulací dle **scénáře A** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (0.30, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.25, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$SE_{\hat{\beta}_0}$	$SE^{sim}_{\hat{\beta}_0}$	$q_{\hat{\beta}_0}$	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	$SE^{sim}_{\hat{\beta}_1}$	$q_{\hat{\beta}_1}$	$\hat{\beta}_2$	$SE_{\hat{\beta}_2}$	$SE^{sim}_{\hat{\beta}_2}$	$q_{\hat{\beta}_2}$
Poiss	1000	0.13	0.38	0.41	0.90	0.05	0.09	0.09	0.92	0.09	0.26	0.31	0.91
NB	1000	0.13	0.42	0.41	0.95	0.05	0.10	0.09	0.94	0.09	0.29	0.31	0.93
ZIP	761	0.33	0.42	0.42	0.95	0.04	0.09	0.09	0.95	0.09	0.29	0.31	0.93
ZINB	634	0.30	0.44	0.42	0.95	0.05	0.10	0.09	0.97	0.10	0.30	0.31	0.93
PHM	997	0.33	0.50	0.50	0.94	0.04	0.11	0.11	0.95	0.06	0.34	0.36	0.96
NBHM	997	0.30	0.53	0.51	0.95	0.04	0.12	0.12	0.97	0.06	0.36	0.37	0.97

Model	Zkonv.	$\hat{\gamma}_0$	$SE_{\hat{\gamma}_0}$	$SE^{sim}_{\hat{\gamma}_0}$	$q_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$SE_{\hat{\gamma}_1}$	$SE^{sim}_{\hat{\gamma}_1}$	$q_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$SE_{\hat{\gamma}_2}$	$SE^{sim}_{\hat{\gamma}_2}$	$q_{\hat{\gamma}_2}$
ZIP	761	-3.97	12.44	9.83	0.97	-0.10	1.23	2.19	0.99	1.14	18.27	5.69	1.00
ZINB	634	-4.26	14.46	10.04	0.97	-0.12	1.38	2.26	0.99	1.93	16.22	5.69	1.00
PHM	997	-1.13	1.05	1.11	0.79	0.05	0.24	0.26	0.96	0.16	0.73	0.75	0.97
NBHM	997	-1.13	1.05	1.11	0.79	0.05	0.24	0.26	0.96	0.16	0.73	0.75	0.97

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (0.30, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.25, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$SE_{\hat{\beta}_0}$	$SE^{sim}_{\hat{\beta}_0}$	$q_{\hat{\beta}_0}$	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	$SE^{sim}_{\hat{\beta}_1}$	$q_{\hat{\beta}_1}$	$\hat{\beta}_2$	$SE_{\hat{\beta}_2}$	$SE^{sim}_{\hat{\beta}_2}$	$q_{\hat{\beta}_2}$
Poiss	1000	0.14	0.11	0.13	0.70	0.05	0.03	0.03	0.92	0.10	0.08	0.09	0.92
NB	1000	0.14	0.13	0.13	0.79	0.05	0.03	0.03	0.96	0.10	0.09	0.09	0.95
ZIP	1000	0.30	0.13	0.13	0.95	0.05	0.03	0.03	0.95	0.10	0.09	0.09	0.95
ZINB	995	0.29	0.13	0.13	0.95	0.05	0.03	0.03	0.96	0.10	0.09	0.09	0.95
PHM	1000	0.30	0.15	0.14	0.95	0.05	0.03	0.03	0.96	0.10	0.10	0.10	0.96
NBHM	1000	0.29	0.15	0.15	0.95	0.05	0.03	0.03	0.97	0.10	0.10	0.10	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$SE_{\hat{\gamma}_0}$	$SE^{sim}_{\hat{\gamma}_0}$	$q_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$SE_{\hat{\gamma}_1}$	$SE^{sim}_{\hat{\gamma}_1}$	$q_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$SE_{\hat{\gamma}_2}$	$SE^{sim}_{\hat{\gamma}_2}$	$q_{\hat{\gamma}_2}$
ZIP	1000	-2.36	0.83	0.87	0.94	0.10	0.15	0.20	0.95	0.21	0.63	0.68	0.98
ZINB	995	-2.70	1.82	2.45	0.96	0.12	0.20	0.37	0.95	0.29	1.81	1.39	0.98
PHM	1000	-1.08	0.30	0.33	0.05	0.05	0.07	0.07	0.88	0.10	0.21	0.22	0.92
NBHM	1000	-1.08	0.30	0.33	0.05	0.05	0.07	0.07	0.88	0.10	0.21	0.22	0.92

Tabulka A.2: Souhrnné výsledky 1000 simulací dle **scénáře B** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (0.30, 0.05, 0.10)^\mathsf{T}$, $\boldsymbol{\gamma} = (-0.50, 0.10, 0.20)^\mathsf{T}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	999	-0.37	0.49	0.69	0.66	0.04	0.11	0.16	0.84	0.03	0.35	0.50	0.84
NB	999	-0.36	0.71	0.72	0.85	0.04	0.16	0.17	0.95	0.03	0.50	0.51	0.95
ZIP	962	0.29	0.62	0.67	0.93	0.04	0.14	0.15	0.93	0.04	0.42	0.49	0.93
ZINB	859	0.26	0.64	0.68	0.93	0.04	0.14	0.16	0.93	0.05	0.43	0.48	0.94
PHM	983	0.28	0.69	0.71	0.95	0.04	0.16	0.16	0.95	0.04	0.48	0.53	0.96
NBHM	989	0.23	0.75	0.74	0.96	0.04	0.17	0.17	0.96	-0.04	0.92	1.10	0.97

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	962	-0.98	2.05	3.26	0.98	0.15	0.37	0.64	0.98	0.26	1.62	1.66	0.98
ZINB	859	-1.07	3.13	3.68	0.98	0.10	0.47	1.05	0.98	0.44	2.43	2.14	0.98
PHM	983	-0.09	0.96	1.03	0.94	0.10	0.23	0.25	0.95	0.18	0.68	0.70	0.96
NBHM	989	-0.09	0.96	1.04	0.93	0.10	0.23	0.25	0.95	0.18	0.68	0.70	0.96

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (0.30, 0.05, 0.10)^\mathsf{T}$, $\boldsymbol{\gamma} = (-0.50, 0.10, 0.20)^\mathsf{T}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	-0.38	0.15	0.20	0.03	0.05	0.03	0.05	0.85	0.09	0.10	0.14	0.85
NB	1000	-0.38	0.22	0.20	0.12	0.05	0.05	0.05	0.97	0.09	0.15	0.14	0.97
ZIP	1000	0.29	0.18	0.17	0.96	0.05	0.04	0.04	0.96	0.10	0.12	0.12	0.95
ZINB	1000	0.28	0.18	0.18	0.96	0.05	0.04	0.04	0.96	0.10	0.12	0.12	0.95
PHM	1000	0.29	0.19	0.18	0.96	0.05	0.04	0.04	0.96	0.10	0.13	0.13	0.96
NBHM	1000	0.28	0.20	0.19	0.96	0.05	0.04	0.04	0.97	0.10	0.13	0.13	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-0.53	0.35	0.37	0.94	0.10	0.08	0.08	0.94	0.21	0.24	0.25	0.94
ZINB	1000	-0.57	0.37	0.39	0.94	0.11	0.08	0.08	0.94	0.21	0.24	0.25	0.94
PHM	1000	-0.06	0.28	0.30	0.64	0.09	0.07	0.07	0.93	0.18	0.20	0.21	0.94
NBHM	1000	-0.06	0.28	0.30	0.64	0.09	0.07	0.07	0.93	0.18	0.20	0.21	0.94

Tabulka A.3: Souhrnné výsledky 1000 simulací dle **scénáře C** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (0.30, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (1.30, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	840	-1.76	1.00	1.60	0.45	0.02	0.23	0.37	0.79	0.17	0.71	0.93	0.88
NB	839	-1.80	1.68	2.11	0.73	0.02	0.39	0.50	0.92	0.20	1.14	1.05	0.97
ZIP	660	-0.15	1.69	2.71	0.91	0.05	0.39	0.66	0.90	0.13	0.95	1.25	0.93
ZINB	594	-0.04	1.79	3.22	0.90	0.02	0.42	0.77	0.89	0.13	0.97	1.33	0.92
PHM	524	-0.08	3.82	6.02	0.98	0.07	0.95	1.69	0.98	-0.12	3.07	3.15	0.98
NBHM	561	-0.33	3.53	7.24	0.96	0.07	0.78	1.89	0.96	-0.51	5.55	4.35	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	660	0.81	2.04	3.16	0.99	0.18	0.48	0.70	0.99	-0.07	1.36	1.43	0.99
ZINB	594	0.73	2.17	3.50	0.99	0.20	0.52	0.83	0.99	-0.11	1.55	1.78	0.99
PHM	524	1.38	1.42	1.49	0.97	0.15	0.35	0.39	0.97	0.07	1.00	0.82	0.98
NBHM	561	1.40	1.43	1.50	0.97	0.15	0.35	0.39	0.98	0.06	1.00	0.81	0.98

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (0.30, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (1.30, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	-1.61	0.28	0.44	0.00	0.04	0.06	0.10	0.78	0.08	0.20	0.32	0.78
NB	1000	-1.61	0.50	0.45	0.02	0.04	0.11	0.10	0.97	0.08	0.35	0.32	0.96
ZIP	1000	0.30	0.35	0.37	0.94	0.04	0.08	0.08	0.94	0.08	0.24	0.25	0.94
ZINB	997	0.25	0.37	0.43	0.93	0.04	0.08	0.08	0.95	0.08	0.25	0.26	0.95
PHM	1000	0.31	0.36	0.37	0.94	0.04	0.08	0.08	0.96	0.08	0.25	0.26	0.96
NBHM	1000	0.28	0.39	0.39	0.96	0.05	0.09	0.08	0.96	0.08	0.26	0.27	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	1.27	0.45	0.46	0.94	0.11	0.11	0.11	0.95	0.22	0.33	0.32	0.96
ZINB	997	1.14	0.61	1.31	0.94	0.11	0.12	0.27	0.95	0.23	0.59	0.68	0.96
PHM	1000	1.55	0.43	0.44	0.91	0.10	0.10	0.11	0.95	0.22	0.32	0.31	0.96
NBHM	1000	1.55	0.43	0.44	0.91	0.10	0.10	0.11	0.95	0.22	0.32	0.31	0.96

Tabulka A.4: Souhrnné výsledky 1000 simulací dle **scénáře E** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (1.35, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-0.50, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	0.66	0.29	0.57	0.41	0.04	0.07	0.13	0.66	0.08	0.20	0.39	0.72
NB	1000	0.67	0.70	0.65	0.86	0.04	0.16	0.16	0.97	0.08	0.49	0.41	0.98
ZIP	998	1.34	0.31	0.32	0.95	0.05	0.07	0.07	0.95	0.10	0.21	0.23	0.96
ZINB	998	1.34	0.32	0.32	0.96	0.05	0.07	0.07	0.96	0.10	0.22	0.23	0.96
PHM	998	1.34	0.31	0.31	0.95	0.05	0.07	0.07	0.95	0.10	0.21	0.22	0.96
NBHM	998	1.34	0.32	0.31	0.96	0.05	0.07	0.07	0.96	0.10	0.22	0.22	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	998	-0.51	0.96	0.99	0.96	0.10	0.23	0.24	0.96	0.21	0.68	0.72	0.96
ZINB	998	-0.52	0.97	0.99	0.96	0.10	0.23	0.24	0.96	0.21	0.68	0.72	0.96
PHM	998	-0.48	0.95	0.97	0.96	0.10	0.22	0.23	0.96	0.20	0.67	0.71	0.96
NBHM	998	-0.48	0.95	0.97	0.96	0.10	0.22	0.23	0.96	0.20	0.67	0.71	0.96

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (1.35, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-0.50, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	0.68	0.09	0.16	0.00	0.05	0.02	0.04	0.69	0.09	0.06	0.12	0.71
NB	1000	0.69	0.22	0.16	0.09	0.05	0.05	0.04	0.99	0.09	0.16	0.12	0.99
ZIP	1000	1.35	0.09	0.09	0.96	0.05	0.02	0.02	0.96	0.10	0.06	0.07	0.94
ZINB	1000	1.35	0.09	0.09	0.96	0.05	0.02	0.02	0.96	0.10	0.06	0.07	0.94
PHM	1000	1.35	0.09	0.09	0.96	0.05	0.02	0.02	0.96	0.10	0.06	0.07	0.94
NBHM	1000	1.35	0.09	0.09	0.97	0.05	0.02	0.02	0.96	0.10	0.06	0.07	0.94

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-0.51	0.28	0.28	0.96	0.10	0.06	0.06	0.96	0.21	0.20	0.20	0.95
ZINB	1000	-0.51	0.28	0.28	0.96	0.10	0.07	0.06	0.96	0.21	0.20	0.20	0.95
PHM	1000	-0.49	0.28	0.28	0.95	0.10	0.06	0.06	0.95	0.20	0.20	0.20	0.95
NBHM	1000	-0.49	0.28	0.28	0.95	0.10	0.06	0.06	0.95	0.20	0.20	0.20	0.95

Tabulka A.5: Souhrnné výsledky 1000 simulací dle **scénáře F** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (1.35, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (1.30, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	894	-0.66	0.58	1.39	0.22	0.03	0.13	0.32	0.59	0.10	0.42	0.85	0.67
NB	891	-0.61	1.43	1.95	0.57	0.01	0.33	0.48	0.84	0.10	0.98	1.01	0.93
ZIP	776	1.19	0.87	1.73	0.94	0.06	0.21	0.41	0.94	0.06	0.51	0.75	0.93
ZINB	763	1.23	0.89	1.68	0.95	0.05	0.21	0.41	0.94	0.06	0.52	0.73	0.93
PHM	768	1.25	0.88	1.43	0.97	0.06	0.21	0.34	0.97	0.09	0.50	0.58	0.96
NBHM	770	1.25	0.89	1.43	0.97	0.06	0.21	0.34	0.97	0.04	0.64	0.97	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	776	1.19	1.39	1.45	0.97	0.15	0.34	0.37	0.97	0.09	1.06	0.86	0.98
ZINB	763	1.18	1.39	1.45	0.97	0.15	0.34	0.37	0.97	0.10	0.97	0.80	0.98
PHM	768	1.25	1.36	1.43	0.97	0.15	0.33	0.36	0.97	0.10	0.96	0.79	0.98
NBHM	770	1.25	1.36	1.43	0.97	0.15	0.33	0.36	0.97	0.10	0.96	0.79	0.98

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (1.35, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (1.30, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	-0.58	0.16	0.37	0.00	0.05	0.04	0.09	0.60	0.09	0.12	0.28	0.61
NB	1000	-0.57	0.38	0.38	0.01	0.05	0.09	0.09	0.86	0.09	0.27	0.28	0.85
ZIP	1000	1.34	0.17	0.17	0.94	0.05	0.04	0.04	0.94	0.10	0.12	0.12	0.93
ZINB	1000	1.34	0.18	0.18	0.95	0.05	0.04	0.04	0.95	0.10	0.12	0.12	0.94
PHM	1000	1.34	0.17	0.17	0.94	0.05	0.04	0.04	0.94	0.10	0.12	0.12	0.94
NBHM	1000	1.34	0.18	0.17	0.95	0.05	0.04	0.04	0.95	0.10	0.12	0.12	0.94

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	1.29	0.39	0.41	0.94	0.11	0.09	0.10	0.94	0.21	0.29	0.29	0.95
ZINB	1000	1.29	0.39	0.41	0.94	0.11	0.09	0.10	0.94	0.21	0.29	0.29	0.95
PHM	1000	1.30	0.39	0.41	0.94	0.11	0.09	0.10	0.94	0.21	0.29	0.29	0.95
NBHM	1000	1.30	0.39	0.41	0.94	0.11	0.09	0.10	0.94	0.21	0.29	0.29	0.95

Tabulka A.6: Souhrnné výsledky 1000 simulací dle **scénáře G** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (2.00, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.25, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	1.83	0.16	0.26	0.69	0.05	0.04	0.06	0.77	0.08	0.11	0.18	0.79
NB	1000	1.83	0.30	0.27	0.95	0.05	0.07	0.06	0.96	0.08	0.21	0.18	0.98
ZIP	924	1.98	0.16	0.16	0.94	0.05	0.04	0.04	0.95	0.10	0.11	0.11	0.96
ZINB	924	1.98	0.17	0.16	0.95	0.05	0.04	0.04	0.95	0.10	0.12	0.11	0.96
PHM	924	1.98	0.16	0.16	0.94	0.05	0.04	0.04	0.95	0.10	0.11	0.11	0.96
NBHM	924	1.98	0.17	0.16	0.95	0.05	0.04	0.04	0.95	0.10	0.12	0.11	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	924	-2.36	1.44	1.62	0.96	0.08	0.33	0.35	0.97	0.33	0.96	0.91	0.97
ZINB	924	-2.36	1.44	1.62	0.96	0.08	0.33	0.35	0.97	0.33	0.96	0.91	0.97
PHM	924	-2.36	1.44	1.62	0.96	0.08	0.32	0.35	0.97	0.33	0.95	0.90	0.97
NBHM	924	-2.36	1.44	1.62	0.96	0.08	0.32	0.35	0.97	0.33	0.95	0.90	0.97

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (2.00, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.25, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	1.84	0.05	0.07	0.19	0.05	0.01	0.02	0.79	0.10	0.03	0.05	0.79
NB	1000	1.84	0.09	0.07	0.65	0.05	0.02	0.02	0.98	0.10	0.07	0.05	0.98
ZIP	1000	2.00	0.05	0.05	0.94	0.05	0.01	0.01	0.94	0.10	0.03	0.03	0.94
ZINB	1000	2.00	0.05	0.05	0.94	0.05	0.01	0.01	0.94	0.10	0.03	0.03	0.95
PHM	1000	2.00	0.05	0.05	0.94	0.05	0.01	0.01	0.94	0.10	0.03	0.03	0.94
NBHM	999	2.00	0.05	0.05	0.94	0.05	0.01	0.01	0.94	0.10	0.03	0.03	0.95

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-2.25	0.40	0.39	0.95	0.10	0.09	0.09	0.95	0.18	0.28	0.26	0.97
ZINB	1000	-2.25	0.40	0.39	0.95	0.10	0.09	0.09	0.95	0.18	0.28	0.26	0.97
PHM	1000	-2.25	0.40	0.39	0.95	0.10	0.09	0.09	0.95	0.18	0.28	0.26	0.97
NBHM	999	-2.25	0.40	0.39	0.95	0.10	0.09	0.09	0.95	0.18	0.28	0.26	0.97

Tabulka A.7: Souhrnné výsledky 1000 simulací dle **scénáře H** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (2.00, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-0.50, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	1.34	0.21	0.49	0.30	0.04	0.05	0.11	0.57	0.08	0.15	0.38	0.58
NB	1000	1.33	0.24	0.49	0.32	0.04	0.05	0.12	0.60	0.08	0.16	0.39	0.60
ZIP	1000	2.00	0.22	0.22	0.94	0.05	0.05	0.05	0.95	0.09	0.15	0.16	0.94
ZINB	1000	2.00	0.22	0.22	0.95	0.05	0.05	0.05	0.95	0.09	0.15	0.16	0.95
PHM	1000	2.00	0.22	0.22	0.94	0.05	0.05	0.05	0.95	0.09	0.15	0.16	0.94
NBHM	1000	2.00	0.22	0.22	0.95	0.05	0.05	0.05	0.95	0.09	0.15	0.16	0.95

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-0.59	0.96	1.02	0.96	0.12	0.22	0.24	0.96	0.21	0.67	0.68	0.96
ZINB	1000	-0.59	0.96	1.02	0.96	0.12	0.22	0.24	0.96	0.21	0.67	0.68	0.96
PHM	1000	-0.59	0.96	1.01	0.96	0.12	0.22	0.24	0.96	0.21	0.67	0.67	0.96
NBHM	1000	-0.59	0.96	1.01	0.96	0.12	0.22	0.24	0.96	0.21	0.67	0.67	0.96

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (2.00, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-0.50, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	1.33	0.06	0.16	0.00	0.05	0.01	0.04	0.56	0.09	0.04	0.11	0.58
NB	1000	1.33	0.06	0.16	0.00	0.05	0.01	0.04	0.56	0.09	0.04	0.11	0.58
ZIP	1000	2.00	0.06	0.07	0.95	0.05	0.01	0.01	0.95	0.10	0.04	0.04	0.95
ZINB	1000	2.00	0.06	0.07	0.95	0.05	0.01	0.01	0.95	0.10	0.04	0.04	0.95
PHM	1000	2.00	0.06	0.07	0.95	0.05	0.01	0.01	0.95	0.10	0.04	0.04	0.95
NBHM	1000	2.00	0.06	0.07	0.95	0.05	0.01	0.01	0.95	0.10	0.04	0.04	0.95

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-0.51	0.28	0.27	0.96	0.10	0.06	0.07	0.96	0.19	0.20	0.20	0.95
ZINB	1000	-0.51	0.28	0.27	0.96	0.10	0.06	0.07	0.96	0.19	0.20	0.20	0.95
PHM	1000	-0.51	0.28	0.27	0.96	0.10	0.06	0.07	0.96	0.19	0.20	0.20	0.95
NBHM	1000	-0.51	0.28	0.27	0.96	0.10	0.06	0.07	0.96	0.19	0.20	0.20	0.95

Tabulka A.8: Souhrnné výsledky 1000 simulací dle **scénáře I** sim. studie 1. Jsou uvedeny počty úspěšných pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (2.00, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (1.30, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$SE_{\hat{\beta}_0}$	$SE^{sim}_{\hat{\beta}_0}$	$q_{\hat{\beta}_0}$	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	$SE^{sim}_{\hat{\beta}_1}$	$q_{\hat{\beta}_1}$	$\hat{\beta}_2$	$SE_{\hat{\beta}_2}$	$SE^{sim}_{\hat{\beta}_2}$	$q_{\hat{\beta}_2}$
Poiss	905	0.12	0.40	1.28	0.17	-0.00	0.09	0.33	0.47	0.16	0.29	0.82	0.48
NB	899	0.05	0.79	1.76	0.28	0.02	0.18	0.48	0.62	0.20	0.55	0.90	0.62
ZIP	815	1.97	0.56	0.83	0.96	0.05	0.14	0.22	0.97	0.09	0.33	0.40	0.95
ZINB	797	1.99	0.62	1.17	0.96	0.03	0.16	0.37	0.97	0.14	0.35	0.51	0.96
PHM	815	1.97	0.55	0.68	0.96	0.06	0.14	0.18	0.97	0.09	0.32	0.36	0.95
NBHM	814	1.97	0.57	0.68	0.96	0.06	0.14	0.18	0.97	0.10	0.33	0.37	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$SE_{\hat{\gamma}_0}$	$SE^{sim}_{\hat{\gamma}_0}$	$q_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$SE_{\hat{\gamma}_1}$	$SE^{sim}_{\hat{\gamma}_1}$	$q_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$SE_{\hat{\gamma}_2}$	$SE^{sim}_{\hat{\gamma}_2}$	$q_{\hat{\gamma}_2}$
ZIP	815	1.22	1.37	1.60	0.93	0.16	0.34	0.39	0.96	0.03	0.95	0.86	0.97
ZINB	797	1.16	1.44	1.75	0.93	0.16	0.35	0.42	0.96	0.01	0.97	0.91	0.97
PHM	815	1.22	1.37	1.60	0.93	0.16	0.33	0.38	0.96	0.03	0.95	0.86	0.97
NBHM	814	1.22	1.37	1.60	0.93	0.16	0.33	0.38	0.96	0.03	0.95	0.86	0.97

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (2.00, 0.05, 0.10)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (1.30, 0.10, 0.20)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$SE_{\hat{\beta}_0}$	$SE^{sim}_{\hat{\beta}_0}$	$q_{\hat{\beta}_0}$	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	$SE^{sim}_{\hat{\beta}_1}$	$q_{\hat{\beta}_1}$	$\hat{\beta}_2$	$SE_{\hat{\beta}_2}$	$SE^{sim}_{\hat{\beta}_2}$	$q_{\hat{\beta}_2}$
Poiss	1000	0.09	0.12	0.36	0.00	0.05	0.03	0.08	0.46	0.09	0.08	0.25	0.49
NB	1000	0.09	0.12	0.36	0.00	0.05	0.03	0.08	0.46	0.09	0.08	0.25	0.49
ZIP	1000	2.00	0.12	0.12	0.94	0.05	0.03	0.03	0.95	0.10	0.08	0.08	0.96
ZINB	1000	2.00	0.12	0.12	0.95	0.05	0.03	0.03	0.96	0.10	0.09	0.08	0.96
PHM	1000	2.00	0.12	0.12	0.94	0.05	0.03	0.03	0.95	0.10	0.08	0.08	0.96
NBHM	921	2.00	0.12	0.12	0.95	0.05	0.03	0.03	0.96	0.10	0.08	0.08	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$SE_{\hat{\gamma}_0}$	$SE^{sim}_{\hat{\gamma}_0}$	$q_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$SE_{\hat{\gamma}_1}$	$SE^{sim}_{\hat{\gamma}_1}$	$q_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$SE_{\hat{\gamma}_2}$	$SE^{sim}_{\hat{\gamma}_2}$	$q_{\hat{\gamma}_2}$
ZIP	1000	1.28	0.39	0.39	0.95	0.11	0.09	0.09	0.95	0.22	0.29	0.29	0.96
ZINB	1000	1.28	0.39	0.39	0.95	0.11	0.09	0.09	0.95	0.22	0.29	0.29	0.96
PHM	1000	1.28	0.39	0.39	0.95	0.11	0.09	0.09	0.95	0.22	0.29	0.29	0.96
NBHM	921	1.27	0.39	0.39	0.94	0.10	0.09	0.09	0.95	0.22	0.29	0.28	0.96

A2 Výsledky simulační studie 2

Tabulka A.9: Souhrnné výsledky 1000 simulací sim. studie 2 pro rozsah výběru $n = 50$. Jsou uvedeny počty úsp. pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 50$; $\boldsymbol{\beta} = (0.40, 0.20, 0.10, 0.30, -0.20, -0.30)^\mathsf{T}$, $\boldsymbol{\gamma} = (-2.50, 0.10, 0.20, 0.10, 0.10, -0.10)^\mathsf{T}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	0.31	0.75	0.97	0.87	0.18	0.20	0.25	0.88	0.07	0.35	0.45	0.87
NB	1000	0.31	0.96	0.99	0.94	0.18	0.25	0.25	0.94	0.07	0.45	0.46	0.93
ZIP	648	0.41	0.81	0.84	0.94	0.20	0.21	0.22	0.95	0.10	0.38	0.40	0.93
ZINB	616	0.40	0.82	0.85	0.95	0.20	0.21	0.22	0.95	0.11	0.38	0.40	0.93
PHM	854	0.40	0.82	0.82	0.96	0.20	0.21	0.22	0.96	0.11	0.38	0.39	0.94
NBHM	854	0.40	0.83	0.82	0.96	0.20	0.22	0.22	0.96	0.11	0.38	0.39	0.95

Model	Zkonv.	$\hat{\beta}_3$	$\mathrm{SE}_{\hat{\beta}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_3}$	$\mathrm{q}_{\hat{\beta}_3}$	$\hat{\beta}_4$	$\mathrm{SE}_{\hat{\beta}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_4}$	$\mathrm{q}_{\hat{\beta}_4}$	$\hat{\beta}_5$	$\mathrm{SE}_{\hat{\beta}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_5}$	$\mathrm{q}_{\hat{\beta}_5}$
Poiss	1000	0.28	0.38	0.50	0.87	-0.21	0.18	0.23	0.87	-0.30	0.17	0.22	0.86
NB	1000	0.29	0.49	0.51	0.94	-0.21	0.23	0.24	0.94	-0.30	0.21	0.22	0.94
ZIP	648	0.28	0.41	0.42	0.95	-0.21	0.19	0.19	0.96	-0.30	0.18	0.19	0.94
ZINB	616	0.29	0.41	0.42	0.95	-0.21	0.19	0.19	0.96	-0.30	0.18	0.19	0.94
PHM	854	0.28	0.41	0.42	0.95	-0.20	0.19	0.19	0.95	-0.30	0.18	0.19	0.95
NBHM	854	0.28	0.42	0.42	0.96	-0.20	0.19	0.19	0.96	-0.30	0.18	0.19	0.95

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	648	-4.38	8.15	8.90	0.96	0.29	1.66	2.45	0.98	0.41	3.04	4.67	0.98
ZINB	616	-4.80	8.26	11.29	0.96	0.32	1.92	3.69	0.98	0.51	3.68	7.83	0.98
PHM	854	1.99	3.57	4.16	0.75	0.06	1.05	1.25	0.96	-0.16	1.90	2.20	0.95
NBHM	854	1.99	3.57	4.16	0.75	0.06	1.05	1.25	0.96	-0.16	1.90	2.20	0.95

Model	Zkonv.	$\hat{\gamma}_3$	$\mathrm{SE}_{\hat{\gamma}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_3}$	$\mathrm{q}_{\hat{\gamma}_3}$	$\hat{\gamma}_4$	$\mathrm{SE}_{\hat{\gamma}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_4}$	$\mathrm{q}_{\hat{\gamma}_4}$	$\hat{\gamma}_5$	$\mathrm{SE}_{\hat{\gamma}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_5}$	$\mathrm{q}_{\hat{\gamma}_5}$
ZIP	648	0.10	1.11	1.72	0.98	0.11	3.44	2.40	0.99	-0.04	4.01	2.91	0.99
ZINB	616	0.14	1.30	2.46	0.98	-0.09	3.45	3.07	0.98	-0.04	3.36	3.58	0.99
PHM	854	-0.07	0.70	0.81	0.96	-0.16	1.00	1.00	0.98	0.04	0.99	1.01	0.97
NBHM	854	-0.07	0.70	0.81	0.96	-0.16	1.00	1.00	0.98	0.04	0.99	1.01	0.97

Tabulka A.10: Souhrnné výsledky 1000 simulací sim. studie 2 pro rozsah výběru $n = 500$. Jsou uvedeny počty úsp. pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 500$; $\boldsymbol{\beta} = (0.40, 0.20, 0.10, 0.30, -0.20, -0.30)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.50, 0.10, 0.20, 0.10, 0.10, -0.10)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	0.32	0.23	0.29	0.86	0.18	0.06	0.07	0.87	0.08	0.11	0.14	0.87
NB	1000	0.31	0.31	0.29	0.96	0.18	0.08	0.07	0.96	0.08	0.15	0.13	0.96
ZIP	1000	0.41	0.24	0.24	0.95	0.20	0.06	0.06	0.95	0.11	0.11	0.11	0.95
ZINB	1000	0.41	0.24	0.25	0.95	0.20	0.06	0.06	0.95	0.11	0.11	0.11	0.95
PHM	1000	0.41	0.24	0.25	0.94	0.20	0.06	0.06	0.95	0.11	0.11	0.11	0.94
NBHM	1000	0.41	0.24	0.25	0.95	0.20	0.06	0.06	0.96	0.11	0.11	0.11	0.94

Model	Zkonv.	$\hat{\beta}_3$	$\mathrm{SE}_{\hat{\beta}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_3}$	$\mathrm{q}_{\hat{\beta}_3}$	$\hat{\beta}_4$	$\mathrm{SE}_{\hat{\beta}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_4}$	$\mathrm{q}_{\hat{\beta}_4}$	$\hat{\beta}_5$	$\mathrm{SE}_{\hat{\beta}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_5}$	$\mathrm{q}_{\hat{\beta}_5}$
Poiss	1000	0.30	0.12	0.15	0.87	-0.21	0.05	0.06	0.88	-0.30	0.05	0.06	0.89
NB	1000	0.30	0.16	0.15	0.97	-0.21	0.07	0.06	0.97	-0.30	0.07	0.06	0.98
ZIP	1000	0.30	0.12	0.12	0.95	-0.20	0.05	0.05	0.96	-0.30	0.05	0.05	0.96
ZINB	1000	0.30	0.12	0.12	0.95	-0.20	0.05	0.05	0.96	-0.30	0.05	0.05	0.96
PHM	1000	0.30	0.12	0.12	0.95	-0.20	0.05	0.05	0.96	-0.30	0.05	0.05	0.96
NBHM	1000	0.30	0.12	0.12	0.95	-0.20	0.05	0.05	0.96	-0.30	0.05	0.05	0.96

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-2.52	1.17	1.19	0.95	0.09	0.34	0.34	0.95	0.22	0.61	0.62	0.95
ZINB	1000	-2.54	1.18	1.20	0.95	0.10	0.34	0.34	0.95	0.22	0.61	0.63	0.95
PHM	1000	1.77	0.96	0.99	0.01	0.06	0.28	0.28	0.95	-0.13	0.51	0.52	0.91
NBHM	1000	1.77	0.96	0.99	0.01	0.06	0.28	0.28	0.95	-0.13	0.51	0.52	0.91

Model	Zkonv.	$\hat{\gamma}_3$	$\mathrm{SE}_{\hat{\gamma}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_3}$	$\mathrm{q}_{\hat{\gamma}_3}$	$\hat{\gamma}_4$	$\mathrm{SE}_{\hat{\gamma}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_4}$	$\mathrm{q}_{\hat{\gamma}_4}$	$\hat{\gamma}_5$	$\mathrm{SE}_{\hat{\gamma}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_5}$	$\mathrm{q}_{\hat{\gamma}_5}$
ZIP	1000	0.10	0.22	0.22	0.95	0.11	0.31	0.31	0.96	-0.13	0.32	0.32	0.96
ZINB	1000	0.10	0.22	0.23	0.95	0.11	0.31	0.31	0.95	-0.13	0.32	0.32	0.96
PHM	1000	-0.09	0.19	0.19	0.83	-0.20	0.27	0.27	0.84	0.11	0.27	0.26	0.90
NBHM	1000	-0.09	0.19	0.19	0.83	-0.20	0.27	0.27	0.84	0.11	0.27	0.26	0.90

Tabulka A.11: Souhrnné výsledky 1000 simulací sim. studie 2 pro rozsah výběru $n = 1000$. Jsou uvedeny počty úsp. pokusů o aplikaci modelu (sloupec Zkonv.), pro každý regr. koeficient pak průměrné hodnoty odhadu a příslušné směr. odchylky (SE), dále směr. odchylka odhadů napříč simulacemi (SE^{sim}) a odhad pravděpodobnosti pokrytí skutečné hodnoty 95% konfidenčním intervalem (q).

Rozsah výběru $n = 1000$; $\boldsymbol{\beta} = (0.40, 0.20, 0.10, 0.30, -0.20, -0.30)^{\mathsf{T}}$, $\boldsymbol{\gamma} = (-2.50, 0.10, 0.20, 0.10, 0.10, -0.10)^{\mathsf{T}}$.

Model	Zkonv.	$\hat{\beta}_0$	$\mathrm{SE}_{\hat{\beta}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_0}$	$\mathrm{q}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\mathrm{SE}_{\hat{\beta}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_1}$	$\mathrm{q}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\mathrm{SE}_{\hat{\beta}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_2}$	$\mathrm{q}_{\hat{\beta}_2}$
Poiss	1000	0.32	0.16	0.21	0.85	0.18	0.04	0.05	0.84	0.07	0.07	0.10	0.87
NB	1000	0.32	0.22	0.21	0.95	0.18	0.06	0.05	0.95	0.07	0.10	0.10	0.96
ZIP	1000	0.40	0.17	0.17	0.95	0.20	0.04	0.04	0.95	0.10	0.08	0.08	0.96
ZINB	1000	0.40	0.17	0.17	0.95	0.20	0.04	0.04	0.95	0.10	0.08	0.08	0.96
PHM	1000	0.40	0.17	0.17	0.95	0.20	0.04	0.04	0.95	0.10	0.08	0.08	0.96
NBHM	1000	0.40	0.17	0.17	0.95	0.20	0.04	0.04	0.96	0.10	0.08	0.08	0.97

Model	Zkonv.	$\hat{\beta}_3$	$\mathrm{SE}_{\hat{\beta}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_3}$	$\mathrm{q}_{\hat{\beta}_3}$	$\hat{\beta}_4$	$\mathrm{SE}_{\hat{\beta}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_4}$	$\mathrm{q}_{\hat{\beta}_4}$	$\hat{\beta}_5$	$\mathrm{SE}_{\hat{\beta}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\beta}_5}$	$\mathrm{q}_{\hat{\beta}_5}$
Poiss	1000	0.30	0.08	0.10	0.87	-0.21	0.04	0.05	0.85	-0.30	0.03	0.04	0.87
NB	1000	0.30	0.11	0.10	0.97	-0.21	0.05	0.05	0.96	-0.30	0.05	0.04	0.97
ZIP	1000	0.30	0.08	0.08	0.95	-0.20	0.04	0.04	0.95	-0.30	0.04	0.04	0.94
ZINB	1000	0.30	0.09	0.08	0.95	-0.20	0.04	0.04	0.95	-0.30	0.04	0.04	0.94
PHM	1000	0.30	0.09	0.08	0.96	-0.20	0.04	0.04	0.95	-0.30	0.04	0.04	0.95
NBHM	1000	0.30	0.09	0.09	0.96	-0.20	0.04	0.04	0.95	-0.30	0.04	0.04	0.95

Model	Zkonv.	$\hat{\gamma}_0$	$\mathrm{SE}_{\hat{\gamma}_0}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_0}$	$\mathrm{q}_{\hat{\gamma}_0}$	$\hat{\gamma}_1$	$\mathrm{SE}_{\hat{\gamma}_1}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_1}$	$\mathrm{q}_{\hat{\gamma}_1}$	$\hat{\gamma}_2$	$\mathrm{SE}_{\hat{\gamma}_2}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_2}$	$\mathrm{q}_{\hat{\gamma}_2}$
ZIP	1000	-2.54	0.82	0.82	0.94	0.11	0.23	0.24	0.95	0.21	0.42	0.43	0.96
ZINB	1000	-2.55	0.82	0.82	0.94	0.11	0.24	0.24	0.95	0.21	0.43	0.43	0.96
PHM	1000	1.80	0.68	0.68	0.00	0.04	0.20	0.20	0.94	-0.13	0.36	0.37	0.84
NBHM	1000	1.80	0.68	0.68	0.00	0.04	0.20	0.20	0.94	-0.13	0.36	0.37	0.84

Model	Zkonv.	$\hat{\gamma}_3$	$\mathrm{SE}_{\hat{\gamma}_3}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_3}$	$\mathrm{q}_{\hat{\gamma}_3}$	$\hat{\gamma}_4$	$\mathrm{SE}_{\hat{\gamma}_4}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_4}$	$\mathrm{q}_{\hat{\gamma}_4}$	$\hat{\gamma}_5$	$\mathrm{SE}_{\hat{\gamma}_5}$	$\mathrm{SE}^{\mathrm{sim}}_{\hat{\gamma}_5}$	$\mathrm{q}_{\hat{\gamma}_5}$
ZIP	1000	0.09	0.16	0.16	0.94	0.10	0.22	0.22	0.95	-0.11	0.22	0.23	0.96
ZINB	1000	0.09	0.16	0.16	0.94	0.10	0.22	0.22	0.95	-0.11	0.22	0.23	0.96
PHM	1000	-0.09	0.13	0.13	0.71	-0.19	0.19	0.19	0.64	0.10	0.19	0.19	0.82
NBHM	1000	-0.09	0.13	0.13	0.71	-0.19	0.19	0.19	0.64	0.10	0.19	0.19	0.82