Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Olga Simerská

Testování složených hypotéz v regresních modelech s malým počtem pozorování

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Michal Kulich, PhD.

Studijní program: Matematika, Matematická statistika

Ráda bych poděkovala panu Mgr. Michalu Kulichovi, PhD., za jeho velmi milý a vstřícný přístup a za cenné rady a připomínky po celou dobu vedení mé diplomové práce.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 12. 12. 2006 Olga Simerská

Obsah

Název práce: *Testování složených hypotéz v regresních modelech s malým počtem pozorování*
Autor: *Olga Simerská*
Katedra: *Katedra pravděpodobnosti a matematické statistiky*
Vedoucí diplomové práce: *Mgr. Michal Kulich, PhD.*
e-mail vedoucího: *kulich@karlin.mff.cuni.cz*

Abstrakt: Práce zkoumá pomocí simulačních studií chování asymptotických testů při testování hypotézy o nulovosti některých koeficientů v logistických modelech. Vyšetřují se testy založené na věrohodnostím poměru, Waldův a Raův test. Je zformulována teorie potřebná k odvození tvaru statistik asymptotických testů pro testování složených hypotéz v logistické regresi. Na základě numerických výpočtů na simulovaných datech jsou zkoumány vlastnosti hladin významnosti testů při použití kritických hodnot chí-kvadrát rozdělení. Také jsou porovnány síly testů vypočtené na základě empirických kritických hodnot tak, aby všechny zamítaly nulovou hypotézu na 5% hladině. Hlavní pozornost je věnována závislosti těchto hodnot na rozsahu výběru a na volbě některých parametrů modelů.

Klíčová slova: věrohodností poměr, Raova statistika, Waldova statistika, logistická regrese

Title: *Testing composite hypotheses in small sample regression models*
Author: *Olga Simerská*
Department: *Department of Probability and Mathematical Statistics*
Supervisor: *Mgr. Michal Kulich, PhD.*
Supervisor's e-mail address: *kulich@karlin.mff.cuni.cz*

Abstract: The thesis, through a simulation study, examines the behaviour of asymptotic tests for testing hypotheses that several coefficients in logistic models are zero. Likelihood ratio, Wald's, and Rao's tests are considered. The necessary theory is formulated to derive the form of the statistics of asymptotic tests for testing composite hypotheses in logistic regression. Based on the numerical treatment of simulated data, the levels of significance of these tests are investigated, with critical values of the chi-squared distribution. The powers of the tests are then compared, modified empirically so that all tests reject the null hypotesis at the 5% level. The main focus is on the dependence of these values on the sample size and parameter settings.

Keywords: likelihood ratio, Rao's statistic, Wald's statistic, logistic regression

Úvod

Předpokládejme, že máme experiment, ve kterém pro každou jednotku (například osobu) může odezva nabývat jen jedné ze dvou možných hodnot. V praxi se s takovou situací často setkáváme při lékařských studiích, kdy zkoumáme, zda pacient dostal či nedostal alergii, zda zemřel či nezemřel nebo zda se vyléčil či nevyléčil. Zajímá nás, jaký vliv na úspěch či neúspěch měly jisté (vysvětlující) proměnné.

Chceme-li modelovat takový problém, používáme k tomu teorii zobecněných lineárních modelů, speciálně modelů logistické regrese. Významnost vlivu vysvětlujících proměnných na odezvu v těchto modelech lze zjišťovat jedním ze tří asymptotických testů. Jednak je to pomocí Raova (skórového) testu, dále Waldova testu a také pomocí testu založeného na věrohodnostním poměru. Víme, že za nulové hypotézy mají statistiky, na kterých jsou testy založeny, stejné asymptotické rozdělení. Při použití na konečné výběry menšího rozsahu není jejich rozdělení ještě podrobně prozkoumáno. V těchto případech mohou existovat značné rozdíly v chování testů. Jediný způsob, jak lze tyto rozdíly zjistit, je pomocí simulačních metod.

Existují různé experimentální studie zabývající se zkoumáním rozdílů mezi testy ve vybraných modelech (viz například [5]). My se v této diplomové práci budeme snažit pomocí simulační studie prozkoumat chování Waldova, Raova testu a testu poměrem věrohodností v některých modelech logistické regrese. Toto chování budeme chtít popsat v závislosti na struktuře regresního modelu a rozsahu výběru.

Zaměříme se především na následující model. Nechť $Y \sim \mathrm{Alt}\,(\pi)$ značí odezvu, pak

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \left(X_1 X_2\right) \ ,$$

kde $\beta = (\beta_0, \ldots, \beta_3)^T$ je vektor koeficientů modelu, X_1 má alternativní a

X_2 některé spojité rozdělení. V tomto modelu budeme testovat hypotézu, že parametry β_2 a β_3 jsou nulové. Chceme vědět, jak přesně nám test Raův, Waldův a poměrem věrohodností odhadne hladinu významnosti pro některé menší rozsahy výběru, pokud použijeme 95% kritických hodnot chí-kvadrát rozdělení. Také nás zajímá, jestli jsou mezi testy nějaké rozdíly v síle. Pro každý test proto vypočítáme empirické kritické hodnoty, pomocí kterých odhadneme síly testů na stejné hladině a budeme testovat nulovou hypotézu při platnosti některých alternativních hypotéz.

Práce je tematicky rozdělena do čtyř kapitol. V první kapitole uvedeme základní definice a tvrzení týkající se teorie maximální věrohodnosti a zavedeme značení, které budeme nadále v práci používat. Dále se zaměříme na teorii asymptotických testů pro testování složených hypotéz.

Druhá kapitola nás uvede do problematiky logistické regrese. Definujeme si zde zobecněné lineární modely a některé jejich vlastnosti. Vybrané vztahy z první kapitoly odvodíme pro logistickou regresi a s jejich pomocí získáme statistiky asymptotických testů pro tuto regresi.

Ve třetí kapitole popíšeme, jaké jsme pro naši simulační studii volili modely a jejich parametry. Dále se zaměříme na postup při výpočtu odhadů ze simulovaných dat a uvedeme i způsoby ošetření některých praktických problémů, které při simulování dat nastávaly.

Čtvrtá kapitola nás seznámí s výsledky simulačních studií. Některé z nich budou prezentovány pomocí grafů nebo tabulek. Závěrem se pokusíme shrnout, co jsme o chování asymptotických testů zjistili.

K práci je přiloženo CD se zdrojovými kódy k funkcím, pomocí kterých jsme prováděli simulaci dat a výpočet testových statistik a relativních četností zamítnutí.

Kapitola 1

Základní teorie

Tato kapitola popisuje teoretické základy, ze kterých vycházíme při odvozování vztahů v kapitole 2. Dále zde zavedeme způsoby značení v naší práci. Vycházíme z knihy [2] a článku [3].

1.1 Základní definice a tvrzení

V této části popíšeme některé principy metody maximální věrohodnosti.

Definice: *Mějme náhodný vektor* $X = (X_1, \ldots, X_n)^T$ *s hustotou* $p(x, \theta)$, *kde* $\theta \in \Theta$ *(* Θ *je parametrický prostor). Hustota* $p(x, \theta)$ *se nazývá* věrohodnostní funkcí, *pokud je funkcí* θ *při pevné hodnotě* x.

Nechť $X_1, \ldots, X_n$ jsou nezávislá a stejně rozdělená pozorování, kde X_i jsou měřitelná zobrazení $(\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$. $\mathcal{A}$ a $\mathcal{B}$ jsou σ-algebry a $\mathcal{X}$ je výběrový prostor. Každé z X_i nechť má rozdělení $P_0 = P_{\theta_0}$ pro jisté $\theta_0 \in \Theta$ a budiž $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ množina pravděpodobnostních měr na $(\Omega, \mathcal{A})$. Učiňme následující předpoklady :

- *P1)* Nechť parametrický prostor $\Theta \subset \mathbb{R}^m$ obsahuje takový neprázdný otevřený interval γ, že θ_0 patří do γ.
- *P2)* Nechť P_θ má hustotu $f(x; \theta)$ vzhledem k nějaké σ-konečné míře μ.
- *P3)* Nechť nosič $A = \{x : f(x; \theta) > 0\}$ nezávisí na θ.

- *P4)* Nechť $\theta_1, \theta_2 \in \Theta$. Pak platí že $\theta_1 \neq \theta_2$ právě tehdy, když $P_{\theta_1} \neq P_{\theta_2}$.

Sdružená hustota $X = (X_1, \ldots, X_n)^T$ je rovna $f(x_1;\theta) \times \ldots \times f(x_n;\theta)$ vzhledem k míře $\nu = \mu \times \ldots \times \mu$. Zavedeme následující značení

$$p(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \,,$$
$$\ell_n(\theta) = \sum_{i=1}^{n} \log\ f(x_i;\theta) \,,$$

kde $\theta \in \Theta$. Logaritmus sdružené hustoty X značený $\ell_n(\theta)$ jakožto funkce proměnné θ budeme nazývat *logaritmickou věrohodnostní funkcí.*

Věta 1 *Nechť jsou splněny předpoklady* $P1)$ *až* $P4)$, *pak pro každé pevné* $\theta \in \Theta$, $\theta \neq \theta_0$ *platí*

$$P_{\theta_0}\{\, p(x;\theta_0) > p(x;\theta) \,\} \longrightarrow 1 \quad \textit{pro } n \to \infty \,.$$

Definice: *Hodnota* $\hat{\theta}_n$ *parametru* θ, *která maximalizuje věrohodnostní funkci* $p(x,\theta)$ *pro dané* $X = x$ *se nazývá* maximálně věrohodný odhad *parametru* θ.

Dále budeme předpokládat, že systém hustot $\{\, f(x;\theta), \theta \in \Theta \,\}$ je regulární (definice viz [2], kapitola 7.3.5) a má Fisherovu informační matici. Zavedeme značení podle [3]

$$U(\theta|X_i) = \frac{\partial \log f(X_i;\theta)}{\partial\theta} \qquad \text{skórová funkce,}$$

$$U_n(\theta) = \sum_{i=1}^{n} U(\theta|X_i) \qquad \text{skórová statistika,}$$

$$I(\theta|X_i) = -\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta\ \partial\theta^T} \,,$$

$$I_n(\theta) = \sum_{i=1}^{n} I(\theta|X_i) \qquad \text{výběrová Fisherova matice,}$$

$$\mathbb{I}(\theta) = \mathrm{E}_\theta I(\theta|X_i) \qquad \text{Fisherova informační matice .}$$

Nyní můžeme zavést pro praktické použití vhodnější definici maximálně věrohodného odhadu, se kterou budeme v dalším textu pracovat.

Definice: *Hodnota* $\hat{\theta}_n$ *parametru* $\theta \in \Theta$ *se nazývá* maximálně věrohodný odhad *parametru* θ *v modelu* $\mathcal{P}$ *právě tehdy, když řeší věrohodnostní rovnici* $U_n(\hat{\theta}_n) = 0$.

Zde se zaměříme na asymptotické výsledky teorie maximální věrohodnosti. Předtím než tak učiníme, budeme definovat některé další předpoklady regularity

- *P5)* Derivace $f'''_{ijk} = \dfrac{\partial^3 f(x;\theta)}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k}$ existuje pro skoro všechna x , pro všechna $\theta \in \gamma$ a pro všechna $i, j, k = 1, \ldots, m$.

- *P6)* Pro všechna $\theta \in \gamma$ platí

$$\int_A \frac{\partial^2 f(x;\theta)}{\partial\theta_i\,\partial\theta_j}\,\mathrm{d}\mu(x) = 0\ , \quad i, j = 1, \ldots, m\ .$$

- *P7)* Pro všechna $i, j, k = 1, \ldots, m$ existují funkce $M_{ijk} \geq 0$ tak, že $\mathrm{E}_{\theta_0} M_{ijk}(X) < \infty$ a

$$|f'''_{ijk}| \leq M_{ijk} \text{ pro všechna } \theta \in \gamma \text{ a skoro všechna } x \in A\ .$$

Poznamenejme ještě, že pokud jsou splněny podmínky regularity, platí $\mathrm{E}_{\theta_0} U(\theta_0|X_i) = 0$ a $\mathrm{var}_{\theta_0} U(\theta_0|X_i) = \mathbb{I}(\theta_0) > 0$.

Věta 2 *Nechť jsou splněny předpoklady P1) až P7) , pak platí následující tvrzení*

(i) Jestliže $n \to \infty$ *, pak s pravděpodobností blížící se jedné existuje posloupnost řešení* $\hat{\theta}_n$ *věrohodnostní rovnice taková, že* $\hat{\theta}_n \xrightarrow{P} \theta_0$.

(ii) $\dfrac{1}{\sqrt{n}}\, U_n(\theta_0) \xrightarrow{d} N_m(0, \mathbb{I}(\theta_0))$,

(iii) $\sqrt{n}(\hat{\theta}_n - \theta_0) = \dfrac{1}{\sqrt{n}}\, \mathbb{I}^{-1}(\theta_0)\, U_n(\theta_0) + o_P(1) \xrightarrow{d} N_m(0, \mathbb{I}^{-1}(\theta_0))$.

S využitím výsledků z Věty 2 a rozvinutím logaritmu věrohodnosti v Taylorovu řadu lze dokázat následující větu o asymptotických testech založených na věrohodnostní funkci.

Věta 3 *Mějme nezávislé stejně rozdělené veličiny* $X_1, \ldots, X_n$ *s rozdělením* $P_\theta \in \mathcal{P} = \{P_\theta; \theta \in \mathbb{R}^m\}$. *Nechť jsou splněny předpoklady* $P1)$ *až* $P7)$. *Označme*

$$\textit{(i)} \quad \lambda_n = \frac{p(\hat{\theta}_n|X)}{p(\theta_0|X)} \qquad \textit{(věrohodnostní poměr)},$$

$$\textit{(ii)} \quad W_n = n\,(\hat{\theta}_n - \theta_0)^T\, \mathbb{I}(\hat{\theta}_n)\,(\hat{\theta}_n - \theta_0) \quad \textit{(Waldova statistika)},$$

$$\textit{(iii)} \quad R_n = \frac{1}{n}\, U_n(\theta_0|X)^T\, \mathbb{I}^{-1}(\theta_0)\, U_n(\theta_0|X) \quad \textit{(Raova (skórová) statistika)}.$$

Potom za platnosti $H_0 : \theta = \theta_0$ *platí* $\left.\begin{matrix} 2\,\log\,\lambda_n \\ R_n \end{matrix}\right\} \xrightarrow{d} \chi^2_m$ *a je-li funkce* $\mathbb{I}(\theta)$ *spojitá v bodě* θ_0, *pak také* W_n *má asymptoticky* χ^2_m *rozdělení.*

Poznámka: Je-li $\mathbb{I}(\theta)$ spojitá v bodě θ_0, pak můžeme místo matice $\mathbb{I}(\theta_0)$ použít nějaký její konzistentní odhad $\hat{\mathbb{I}}_n(\hat{\theta}_n) = \frac{1}{n}\,\mathbb{I}_n(\hat{\theta}_n)$.

Tyto statistiky se používají na testování hypotézy $\mathrm{H}_0 : \theta = \theta_0$ proti $\mathrm{H}_1 : \theta \neq \theta_0$. Nulovou hypotézu zamítáme, pokud

$$\left.\begin{matrix} 2\,\log\,\lambda_n \\ W_n \\ R_n \end{matrix}\right\} > \chi^2_m(1-\alpha)\,.$$

1.2 Testování složených hypotéz

Předpokládejme, že $X_1, \ldots, X_n$ mají stejné vlastnosti jako v kapitole 1.1. Rozdělení $X = (X_1, \ldots, X_n)$ je tedy závislé na m - rozměrném parametru $\theta \in \Theta \subset \mathbb{R}^m$, $m \geq 2$. Dejme tomu, že nás zajímá ta informace o rozdělení, která je obsažena v jisté k-rozměrné části θ, $1 \leq k < m$. Označme ji τ,

potom $\theta^T = (\tau^T, \psi^T)$ a

$$\tau = \begin{pmatrix} \theta_1 \\ \dots \\ \theta_k \end{pmatrix}, \quad \psi = \begin{pmatrix} \theta_k + 1 \\ \dots \\ \theta_m \end{pmatrix}.$$

Ze zřejmých důvodů se parametr τ nazývá *cílový* a parametr ψ *rušivý*. Je-li skutečná hodnota parametru θ rovna $\theta_0^T = (\tau_0^T, \psi_0^T)$, pak vlastně chceme testovat hypotézu $\mathrm{H}_0^* : \tau = \tau_0$ s tím, že parametr ψ může nabývat libovolné hodnoty z $\mathbb{R}^{m-k}$. Přesněji zapsáno, testujeme tedy složenou hypotézu

$$\begin{array}{lll} H_0^* : \theta & \in & \Theta_0 = \left\{\theta \in \mathbb{R}^m : \tau = \tau_0,\ \psi \in \mathbb{R}^{m-k}\right\} \quad \text{proti složené alternativě} \\ H_1^* : \theta & \notin & \Theta_0 \,. \end{array}$$

Zavedeme nyní značení pro skórovou statistiku a informační matici, rozdělené podobně jako parametr θ

$$U_n(\theta) = \begin{pmatrix} U_{1n}(\theta) \\ U_{2n}(\theta) \end{pmatrix}, \quad \mathbb{I}(\theta) = \begin{pmatrix} \mathbb{I}_{11}(\theta) & \mathbb{I}_{12}(\theta) \\ \mathbb{I}_{21}(\theta) & \mathbb{I}_{22}(\theta) \end{pmatrix}.$$

Pro výpočet testových statistik je třeba zjistit hodnotu maximálně věrohodného odhadu θ (označíme ho $\hat{\theta}_n$) a hodnotu maximálně věrohodného odhadu θ v submodelu za hypotézy $\mathrm{H}_0^* : \tau = \tau_0$. Ten označíme $\tilde{\theta}_n$,

$$\hat{\theta}_n = \begin{pmatrix} \hat{\tau}_n \\ \hat{\psi}_n \end{pmatrix}, \qquad \tilde{\theta}_n = \begin{pmatrix} \tau_0 \\ \tilde{\psi}_n \end{pmatrix}.$$

Parametr $\tilde{\psi}_n$ je řešením rovnice $U_{2n}(\theta) = 0$. Za splnění předpokladů Věty 2 platí

$$\frac{1}{\sqrt{n}}\, U_{2n}(\theta_0) \xrightarrow{d} N_{m-k}(0, \mathbb{I}_{22}(\theta_0)),$$

$$\sqrt{n}\,(\tilde{\psi}_n - \psi_0) \xrightarrow{d} N_{m-k}(0, \mathbb{I}_{22}^{-1}(\theta_0)).$$

Také zavedeme značení jednotlivých bloků inverzní Fisherovy matice

$$\mathbb{I}(\theta)^{-1} = \begin{pmatrix} \mathbb{I}^{11}(\theta) & \mathbb{I}^{12}(\theta) \\ \mathbb{I}^{21}(\theta) & \mathbb{I}^{22}(\theta) \end{pmatrix}, \text{ kde například } \mathbb{I}^{11}(\theta) = (\mathbb{I}_{11} - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}\mathbb{I}_{21})^{-1}(\theta).$$

Nyní již můžeme definovat testové statistiky pro testování H_0^* :

(i) $\lambda_n^* = \dfrac{p(\hat{\theta}_n|X)}{p(\tilde{\theta}_n|X)}$ (věrohodnostní poměr),

(ii) $W_n^* = n\,(\hat{\tau}_n - \tau_0)^T\,(\mathbb{I}^{11}(\hat{\theta}_n))^{-1}\,(\hat{\tau}_n - \tau_0)$ (Waldova statistika),

(iii) $R_n^* = \dfrac{1}{n}\,U_{1n}(\tilde{\theta}_n|X)^T\,\mathbb{I}^{11}(\tilde{\theta}_n)\,U_{1n}(\tilde{\theta}_n|X)$ (Raova (skórová) statistika).

Věta 4 *Nechť $X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené veličiny s rozdělením $P_\theta \in \mathcal{P}$ splňujícím předpoklady P1) až P7) . Nechť platí hypotéza $H_0^* : \theta \in \Theta_0 = \{\theta \in \mathbb{R}^m : \tau = \tau_0\}$. Potom*

$$\left.\begin{array}{c} 2\,\log\,\lambda_n^* \\ W_n^* \\ R_n^* \end{array}\right\} \xrightarrow{d} \chi_k^2\,.$$

Poznámka: Stejně jako ve Větě 3 lze matici $\mathbb{I}^{11}(\tilde{\theta}_n)$ (respektive $\mathbb{I}^{11}(\hat{\theta}_n)$) nahradit nějakým konzistentním odhadem $\hat{\mathbb{I}}_n^{11}(\tilde{\theta}_n)$ (resp. $\hat{\mathbb{I}}_n^{11}(\hat{\theta}_n)$) matice $\mathbb{I}^{11}(\theta_0)$.

Kapitola 2

Teorie logistické regrese

Zde se zaměříme na teorii zobecněných lineárních modelů, především na jeden jejich speciální případ, logistickou regresi. Pro tento model odvodíme tvary L_n^* , W_n^* a R_n^* statistik z Věty 4. Budeme vycházet z knihy [4].

2.1 Zobecněné lineární modely

Definujme si nejprve exponenciální třídu hustot.

Definice: *Mějme míru* P_θ *, která je absolutně spojitá vzhledem k* σ*-konečné míře* μ *a nechť hustota* $f(x) = \dfrac{dP_\theta}{d\mu}$ *má tvar*

$$f(x;\theta,\varphi) = \exp\left\{\frac{x\,\theta - b(\theta)}{a(\varphi)} + c(x,\varphi)\right\} ,$$

kde $\theta \in \mathbb{R}$, $\varphi > 0$ a $a(\varphi) > 0$. *Potom se rozdělení s hustotou* $f(x;\theta,\varphi)$ *nazývá* rozdělení exponenciálního typu *s hustotou v kanonickém tvaru. Parametr* θ *se nazývá* kanonický, φ *se nazývá* disperzní *parametr.*

Tvrzení: Nechť náhodná veličina X má rozdělení exponenciálního typu s hustotou v kanonickém tvaru, nechť $b(\theta) \in C^2(\mathbb{R}^n)$. Potom existuje momentová vytvořující funkce $\varphi_X(t)$, je všude konečná a dvakrát diferencovatelná v bodě nula.

Označme nosič $A = \{x : f(x;\theta,\varphi) > 0\}$. Platí

$$\begin{aligned}
\mathrm{E}\ \mathrm{e}^{tX} &= \int_A \exp\left\{\frac{x(ta(\varphi)+\theta) - b(\theta)}{a(\varphi)} + c(x,\varphi)\right\} \mathrm{d}x \\
&= \exp\left\{\frac{b\,(t\,a(\varphi)+\theta) - b(\theta)}{a(\varphi)}\right\} \cdot \\
&\cdot \int_A \exp\left\{\frac{x\,(ta(\varphi)+\theta) - b\,(ta(\varphi)+\theta)}{a(\varphi)} + c\right\} \mathrm{d}x = \\
&= \exp\left\{\frac{b(ta(\varphi)+\theta) - b(\theta)}{a(\varphi)}\right\} = \varphi_X(t)\,,
\end{aligned}$$

$$\begin{aligned}
\varphi'_X(t) &= \varphi_X(t)\ b'(ta(\varphi)+\theta)\,, \\
\varphi''_X(t) &= \varphi'_X(t)\ b'(ta(\varphi)+\theta) + \varphi_X(t)\ b''(ta(\varphi)+\theta)\,.
\end{aligned}$$

Na základě vlastností momentové vytvořující funkce pak můžeme psát

$$\begin{aligned}
\mathrm{E}X &= \varphi'_X(0) = b'(\theta)\ , \\
\mathrm{E}X^2 &= \varphi''_X(0) = \left[b'(\theta)\right]^2 + b''(\theta)\ a(\varphi)\ , \\
\mathrm{var}X &= \mathrm{E}X^2 - (\mathrm{E}X)^2 = b''(\theta)\ a(\varphi)\ .
\end{aligned}$$

Nadále budeme předpokládat, že $a(\varphi) \equiv \varphi > 0$. Označíme

$$\mu = b'(\theta)\ , \tag{2.1a}$$

$$\mathrm{var}X = \varphi\ b''(\theta) = \varphi\ V(\mu)\ . \tag{2.1b}$$

Definice: *Mějme nyní nezávislé náhodné veličiny* $Y_1,\ldots,Y_n$, $Y = (Y_1,\ldots,Y_n)$, *a vektory regresorů* $x_i = (x_{i1},\ldots,x_{ip})^T$, $i = 1,\ldots,n$, $p << n$, *které tvoří regresní matici* $X_{n\times p} = (x_1^T,\ldots,x_n^T)$. *Rozdělení každého* Y_i *závisí na* x_i *skrze parametr* $\beta = (\beta_1,\ldots,\beta_p)^T$. *Potom definujeme* zobecněný lineární model *následujícími předpoklady:*

1. Závislé proměnná Y_i má rozdělení exponenciálního typu s hustotu ve tvaru

$$f(y_i;\theta_i,\varphi) = \exp\left\{\frac{y_i\,\theta_i - b(\theta_i)}{a(\varphi)}\,a_i + c(y_i,\varphi)\right\}\,,$$

kde $b \in C^2(\mathbb{R}^n)$, θ_i je parametr závisející na x_i a a_i je známá konstanta nazývaná *apriorní váha*.

2. Parametr θ_i závisí na x_i skrze lineární prediktor

$$\eta_i = x_i^T\,\beta\,.$$

3. Existuje známá ostře monotonní funkce $g \in C^2(\mathbb{R}^n)$ taková, že

$$\eta_i = g(\mu_i)\,,\ \text{kde}\quad \mu_i = \mathrm{E}Y_i\,.$$

Tato funkce se nazývá *linková*.

Poznamenejme ještě, že z rovnosti (2.1a) platí $\mu_i = b'(\theta_i)$ a tudíž pro zobecněné lineární modely máme čtyři ekvivalentní parametrizace

$$\begin{aligned}\eta &= (\eta_1,\ldots,\eta_n)\,,\\ \mu &= (\mu_1,\ldots,\mu_n)\,,\\ \theta &= (\theta_1,\ldots,\theta_n)\,,\\ \beta &= (\beta_1,\ldots,\beta_p)\,.\end{aligned}$$

Parametrizace η, μ a θ mají n složek, které jsou funkcemi p složek hlavní parametrizace β.

Definice: *Linková funkce g se nazývá kanonická právě tehdy, když $\eta_i = g(\mu_i) = \theta_i$.*

2.2 Logistická regrese

Zaměříme se zde na jeden konkrétní případ zobecněných lineárních modelů. Nechť Y_i nabývá pouze dvou hodnot. Často jimi značíme úspěch $Y_i = 1$ a

neúspěch $Y_i = 0$ (například úspěšnost léčby), $Y_i \sim Alt(\pi_i)$, $\pi_i \in (0,1)$ a tudíž $\mathrm{E}\, Y_i = P(Y_i = 1) = \pi_i$.

Hustotu alternativního rozdělení můžeme upravit následujícím způsobem

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left\{ y_i \ \log\ \frac{\pi_i}{1-\pi_i} - \left(\log\ \frac{1}{1-\pi_i}\right)\right\}$$

na hustotu rozdělení exponenciálního typu, kde $\varphi = 1$, $a_i = 1$, $c(y_i, \varphi) = 0$,

$$\theta_i = \log\ \frac{\pi_i}{1-\pi_i}\ ,$$
$$b(\theta_i) = \log\ \frac{1}{1-\pi_i} = \log\left(e^{\theta_i}+1\right),$$

pro $i = 1, \ldots, n$.
Z rovností (2.1) lze tedy odvodit známé vlastnosti alternativního rozdělení

$$\mathrm{E}Y_i = b'(\theta_i) = \frac{e^{\theta_i}}{1+e^{\theta_i}} = \pi_i\ ,$$
$$\mathrm{var}\, Y_i = b''(\theta_i) = \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = \pi_i\,(1-\pi_i) = V(\pi_i) > 0\ .$$

Existuje několik typů linkových funkcí vhodných pro binární odezvu, které zobrazují jednotkový interval na celou reálnou osu $g : (0,1) \to (-\infty, \infty)$. V této práci se zaměříme na logistický link, který je nejčastěji používaný a také nejsnadněji interpretovatelný,

$$g(\pi_i) = \log\ \frac{\pi_i}{1-\pi_i} = x_i^T\ \beta = \eta_i = \theta_i\ .$$

Z poslední rovnosti plyne, že logistický link je linkem kanonickým. Ekvivalentně můžeme model zapsat pomocí šance odezvy na úspěch nebo její pravděpodobnosti

$$\frac{\pi_i}{1-\pi_i} = \mathrm{e}^{\,\beta_0+\beta_1\,x_1+\ldots+\beta_p\,x_p}\ ,$$
$$\pi_i = g^{-1}(x_i^T\ \beta) = \frac{e^{x_i^T\,\beta}}{1+e^{x_i^T\,\beta}}\ .$$

2.3 Asymptotické testy pro logistickou regresi

Nyní si odvodíme některé vztahy z kapitoly 1.1 pro modely logistické regrese. Značení zachováme stejné jako v celé této kapitole.

Logaritmická věrohodnostní funkce má tvar

$$\ell_n(\theta(\beta); Y) = \log \prod_{i=1}^{n} f(y_i; \theta_i) = \sum_{i=1}^{n} \left(Y_i\, \theta_i - b(\theta_i)\right) \ . \tag{2.2}$$

Abychom mohli použít teorii maximální věrohodnosti, musí být splněn předpoklad nezávislosti a stejného rozdělení $Y_1, \ldots, Y_n$. V našem případě nejsou však náhodné veličiny $Y_1, \ldots, Y_n$ stejně rozdělené. Budeme ale předpokládat, že $(Y_1, x_1), \ldots, (Y_n, x_n)$ je náhodný výběr z $(p+1)$-rozměrného rozdělení s hustotou $g(y, x) = f(y \mid x)\, h(x)$, kde

$$f(y, x) = \exp\left\{y\, \theta(x) - b(\theta(x))\right\} \ .$$

je podmíněná hustota Y_i, je-li dáno $x_i = x$, a $h(x)$ je marginální hustota x_i.

Na takovou hustotu $g(y, x)$ můžeme teorii maximální věrohodnosti použít. Vyjádříme-li ji ve tvaru logaritmické věrohodnostní funkce, dostaneme

$$\sum_{i=1}^{n} \log f(y_i \mid \beta_i, x_i) + \sum_{i=1}^{n} \log\ h(x_i) \ .$$

Jestliže tuto hustotu derivujeme podle β, člen $\sum_{i=1}^{n} \log\ h(x_i)$ vypadne, protože nezávisí na β. Nadále tedy můžeme pracovat s tvarem věrohodnostní funkce uvedeném v (2.2).

Skórovou statistiku pro logistickou regresi získáme derivací (2.2) podle β pomocí řetízkového pravidla

$$U_n(\beta \mid Y) = \frac{\partial \ell_n(\beta \mid Y)}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{\partial \ell_n}{\partial \theta_i}\, \frac{\partial \theta(\pi_i)}{\partial \pi_i}\, \frac{\partial \pi(\eta_i)}{\partial \eta_i}\, \frac{\partial \eta_i}{\partial \beta}\right] \ ,$$

$$\frac{\partial \ell_n}{\partial \theta_i} = Y_i - b'(\theta_i) = Y_i - \pi_i \ ,$$

$$\frac{\partial\theta(\pi_i)}{\partial\pi_i} = \frac{1}{\pi_i\,(1-\pi_i)} = \frac{1}{V(\pi_i)}\,,$$

$$\frac{\partial\pi(\eta_i)}{\partial\eta_i} = \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} = V(\pi_i)\,,$$

$$\frac{\partial\eta_i}{\partial\beta} = x_i\,.$$

Tedy

$$U_n(\beta \mid Y) = \sum_{i=1}^{n}(Y_i - \pi_i)\,x_i\,.$$

Maximálně věrohodný odhad β je podle definice takový, který splňuje rovnici

$$U_n(\hat\beta \mid Y) = \sum_{i=1}^{n}(Y_i - \hat\pi_i)\,x_i = 0\,, \quad \text{kde } \hat\pi_i = \frac{e^{x_i^T\,\hat\beta}}{1+e^{x_i^T\,\hat\beta}}\,.$$

Ukážeme, že pro každou hodnotu n existuje právě jeden maximálně věrohodný odhad $\hat\beta$. Výběrovou Fisherovu matici můžeme vyjádřit ve tvaru

$$I_n(\beta \mid Y) = -\frac{\partial U_n(\beta \mid Y)}{\partial\beta^T} = -\sum_{i=1}^{n}\left[\frac{\partial(Y_i-\pi_i)\,x_i}{\partial\pi_i}\,\frac{\partial\pi(\eta_i)}{\partial\eta_i}\,\frac{\partial\eta_i}{\partial\beta^T}\right]\,,$$

$$I_n(\beta) = \sum_{i=1}^{n} V(\pi_i)\,x_i\,x_i^T\,.$$

Za předpokladu plné hodnosti matice X je $I_n(\beta) > 0 \quad \forall\beta$ a logaritmická věrohodnostní funkce je ostře konkávní funkcí β.

Odhad $\hat\beta$ v praxi získáváme pomocí metody iterativních nejmenších vážených čtverců (Iterative Weighted Least Squares), která je založena na následující větě.

Věta 5 *Maximálně věrohodný odhad $\hat\beta$ v zobecněném lineárním modelu splňuje rovnici*

$$\hat\beta = (X^T\,\hat W\,X)^{-1}\,(X^T\,\hat W\,\hat Z)\,,$$

kde

$$\hat{W} = diag\{w(\hat{\pi}_i)\} \quad \textit{je váhová matice}\,,$$

$$w(\hat{\pi}_i) = \frac{1}{V(\hat{\pi}_i)\ [g'(\hat{\pi}_i)]^2}\,,$$

$$\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_n)\,,$$

$$\hat{Z}_i = \hat{\eta}_i + (Y_i - \hat{\pi}_i)\, g'(\hat{\pi}_i)\,,$$

$$\hat{\eta}_i = x_i^T\, \hat{\beta}\,.$$

Pokud je linková funkce g kanonická, jako v případě logistické regrese, platí tyto vztahy

$$g^{-1}(\theta_i) = b(\theta_i)\,,$$

$$g\left(b'(\theta_i)\right) = \theta_i\,.$$

Derivujeme-li vztah podle θ_i, pak $g'(\pi_i)\, b''(\theta_i) = 1$. Z toho vyplývá, že

$$g'(\hat{\pi}_i) = \frac{1}{V(\hat{\pi}_i)},$$

$$w(\hat{\pi}_i) = V(\hat{\pi}_i)\,.$$

Dokážeme si nyní Větu 5 pro případ logistické regrese.

Důkaz: Stačí dokázat platnost rovnosti

$$(X^T\, \hat{W}\, X)\, \hat{\beta} = X^T\, \hat{W}\, \hat{Z}\,.$$

Z definice maximálně věrohodného odhadu víme, že $U_n(\hat{\beta}) = 0$.
Proto

$$\left(\sum_{i=1}^{n} V(\hat{\pi}_i)\, x_i\, x_i^T\right) \hat{\beta} = \left(\sum_{i=1}^{n} V(\hat{\pi}_i)\, x_i\, x_i^T\right) \hat{\beta} + U_n(\hat{\beta}) =$$

$$= \left(\sum_{i=1}^{n} V(\hat{\pi}_i)\, x_i\, x_i^T\right) \hat{\beta} + \sum_{i=1}^{n} (Y_i - \hat{\pi}_i)\, x_i =$$

$$= \sum_{i=1}^{n} V(\hat{\pi}_i)\, x_i \left[x_i^T\, \hat{\beta} + (Y_i - \hat{\pi}_i)\, \frac{1}{V(\hat{\pi}_i)}\right] = X^T\, \hat{W}\, \hat{Z}\,. \qquad \square$$

Podrobnější popis metody lze najít v [4], kapitola 2.5.

K odvození testových statistik asymptotických testů pro logistickou regresi již stačí aplikovat Věty 2 a 4. Z Věty 2 a z centrální limitní věty vyplývají následující vztahy

$$\frac{1}{\sqrt{n}}\, U_n(\beta/Y) \xrightarrow{d} N_m(0, \mathbb{I}(\beta))\,,$$

$$\sqrt{n}\,(\hat{\beta} - \beta) \xrightarrow{d} N_m(0, \mathbb{I}^{-1}(\beta))\,.$$

Konzistentní odhad Fisherovy informační matice označíme $\hat{\mathbb{I}}_n$

$$\hat{\mathbb{I}}_n(\hat{\beta}) = \frac{1}{n}\, X^T\, \hat{W}\, X \xrightarrow{P} \mathbb{I}(\beta)\,,$$

$$\hat{\text{cov}}(\hat{\beta}) = (X^T\, \hat{W}\, X)^{-1}\,.$$

Rozdělíme nyní parametr β na dvě části, stejně jako v kapitole 1.2, $1 \leq k\ < p$

$$\tau = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_k \end{pmatrix}\,, \quad \psi = \begin{pmatrix} \beta_k + 1 \\ \dots \\ \beta_p \end{pmatrix}\,.$$

a testujeme složenou hypotézu $H_0 : \beta \in B_0 = \left\{\beta \in \mathbb{R}^p : \tau = \tau_0,\ \psi \in \mathbb{R}^{p-k}\right\}$. Při splnění předpokladů $P1)$ - $P7)$ můžeme použít testové statistiky z Věty 4 upravené pro parametr β

$$2\ \log \lambda_n^* = 2\ \log \frac{p(\hat{\beta}\ \mid\ Y)}{p(\tilde{\beta}\ \mid\ Y)} = 2\ \sum_{i=1}^{n} Y_i(\hat{\theta}_i - \tilde{\theta}_i) - \left[b(\hat{\theta}_i) - b(\tilde{\theta}_i)\right]\,,$$

$$W_n^* = n\,(\hat{\tau} - \tau_0)^T (\hat{\mathbb{I}}^{11}(\hat{\beta}))^{-1} (\hat{\tau} - \tau_0)\,,$$

$$R_n^* = \frac{1}{n}\, U_{1n}(\tilde{\beta}\,|\,Y)^T\, \hat{\mathbb{I}}^{11}(\tilde{\beta})\, U_{1n}(\tilde{\beta}\,|\,Y)\,,$$

kde $\hat{\theta}_i = x_i^T \hat{\beta}$ a $\tilde{\theta}_i = x_i^T \tilde{\beta}$. Používáme skórovou statistiku a informační matici odvozené pro logistickou regresi. Každá ze statistik má asymptoticky χ^2_k rozdělení.

Pokud τ_0 je vektor samých nul, neboli pokud testujeme nulovost některých koeficientů modelu, můžeme místo statistiky $2 \log \lambda^*_n$ použít statistiku založenou na devianci širšího modelu a jeho submodelu.

Definice: *Pokud ℓ^*_n je logaritmická věrohodnost v saturovaném modelu, statistika*

$$D(Y, \hat{\beta}) = 2\ [\ell^*_n(Y) - \ell_n(Y, \beta)]$$

se nazývá deviance.

Saturovaný model je takový, v kterém $p = n$, tedy délka vektoru odezvy Y je stejná jako délka vektoru koeficientů modelu β a každé pozorování má svůj vlastní parametr. V tomto modelu jsou vyhlazené hodnoty Y_i rovny pozorovaným, $\hat{\pi}_i = Y_i$, a $\ell^*_n(Y)$ je maximální dosažitelná věrohodnost pro pozorování $Y_1, \ldots, Y_n$.

V logistické regresi, kde $Y_i \sim \mathrm{Alt}(\pi_i)$, platí pro saturovaný model $Y_i = \hat{\pi}_i$ a

$$\ell^*_n(Y) = \sum_{i=1}^{n} Y_i\ \log Y_i + (1 - Y_i)\ \log(1 - Y_i) = Y_i\ \log \frac{Y_i}{1 - Y_i} + \log\,(1 - Y_i)\,.$$

Deviance se v tomto případě vypočítá jako

$$D(Y, \hat{\beta}) = 2 \sum_{i=1}^{n} Y_i\ \left[\log \frac{Y_i}{1 - Y_i} - \log \frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right] + \log\,(1 - Y_i) - \log\,(1 - \hat{\pi}_i) =$$

$$= 2 \sum_{i=1}^{n} Y_i\ \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i)\ \log \frac{1 - Y_i}{1 - \hat{\pi}_i}\,.$$

Je-li $\hat{\beta}$ odhad β v širším modelu a $\tilde{\beta}$ odhad β v submodelu za platnosti H_0, pak platí

$$D(Y, \tilde{\beta}) - D(Y, \hat{\beta}) \xrightarrow{d} \chi^2_k \, .$$

K tomuto vztahu dojdeme jednoduchou úpravou na statistiku poměrem věrohodností

$$\begin{aligned} D(Y, \tilde{\beta}) - D(Y, \hat{\beta}) &= 2 \, \left(\ell^*_n(Y) - \ell_n(\tilde{\beta}; Y) \right) - 2 \, \left(\ell^*_n(Y) - \ell_n(\hat{\beta}; Y) \right) = \\ &= 2 \, \left(\ell_n(\hat{\beta}; Y) - \ell_n(\tilde{\beta}; Y) \right) \, . \end{aligned}$$

Kapitola 3

Metodika simulačních studií

Cílem této práce je zkoumat vlastnosti testu založeného na věrohodnostním poměru, Waldova testu a Raova testu (pro test poměrem věrohodností používáme dále značení LR test) pomocí simulovaných dat. Bude nás zajímat, jak se chovají ve zvolených modelech logistické regrese. Na základě simulovaných dat vypočteme odhady hladiny významnosti testů za předpokladu použití kritických hodnot chí-kvadrát rozdělení $\chi^2_{k\,(1-\alpha)}$, kde k jsou stupně volnosti a hladinu významnosti α volíme rovnou 5 procentům. Dále vypočteme síly testů použitím empiricky získaných kritických hodnot takových, že všechny tři testy při jejich použití zamítají nulovou hypotézu na stejné (5%) hladině. Tyto výpočty budeme provádět pro různé hodnoty parametrů našich modelů a pro měnící se počet pozorování.

Naším úkolem je porovnat průběh odhadnutých hladin významnosti sledovaných testů při zvyšujícím se počtu pozorování. Současně přitom chceme sledovat rozdíly ve velikostech „empiricky upravených“ sil (viz kapitola 3.2.2) těchto testů a zjistit, jaký vliv na velikost těchto dvou hodnot mají určité kombinace parametrů modelu. Vhodný typ studie pro takový problém je faktoriálový design (viz [4], kapitola 1.2.3).

Faktoriálový design je typ statistické studie, ve kterém je každý prediktor kategorická veličina s více jak jednou kategorií. Každá kategorie každého prediktoru se ve studii vyskytuje v kombinaci s každou kategorií každého jiného prediktoru, což nám umožňuje zkoumat vliv všech kategorií a jejich interakcí na závisle proměnnou.

V našem případě jsou prediktory zvolené parametry modelu logistické regrese. Každému z nich přiřadíme dvě hodnoty (dvě kategorie) a naší závisle

proměnnou budou postupně relativní četnosti zamítnutí nulové hypotézy u LR, Waldova a Raova testu při její platnosti (odhad hladiny významnosti testu) a neplatnosti (odhad síly testu).

Všechny zde uvedené výpočty byly prováděny pomocí statistického programu R 2.3.0 (viz `http://www.r-project.org`). Zdrojové kódy vybraných funkcí jsou uvedeny na přiloženém CD.

3.1 Popis modelů

Nechť Y značí binární odezvu v modelu logistické regrese, $Y \sim \mathrm{Alt}\,(\pi)$. Například se může jednat o diagnózu jisté nemoci s kategoriemi *přítomna, nepřítomna*. Označme

$$\mathrm{E}Y = P(Y = 1) = \pi(X)\,,$$

kde X je regresní matice. Závislost π na hodnotách X je v našem modelu dána vztahem

$$g(\pi) = \log\left(\frac{\pi(X)}{1-\pi(X)}\right) = X\beta\,,$$

přičemž β je vektor neznámých parametrů. Podle teorie z kapitoly 3 je g linková funkce, která je v logistických modelech kanonická a nazývá se *logit*.

Popíšeme nyní dva logistické modely, se kterými budeme dále pracovat. V prvním modelu, který budeme nazývat *model s interakcemi*, je regresní matice $X = (X_1, X_2, X_1\,X_2)$ tvořena regresorem X_1, který má alternativní rozdělení, regresorem X_2 s jistým spojitým rozdělením, které později podrobněji popíšeme, a regresorem interakce X_1 a X_2. Závislost π na X je pro tento model následující

$$g(\,\pi(X)\,) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3\,(X_1\,X_2)\,, \tag{3.1}$$

kde $-\infty < \beta_j < +\infty$, $j = 0, \ldots, 3$, jsou složky vektoru β.

Druhý model obsahuje stejné regresory X_1 a X_2 jako model s interakcemi, ale nepředpokládá mezi nimi interakci. Nazývejme ho *jednoduchý model*. Platí

pro něj

$$g(\ \pi(X)\ ) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2\ , \tag{3.2}$$

kde $\gamma^T = (\gamma_0, \gamma_1, \gamma_2)$ má reálné složky.

Volba těchto regresorů byla motivována některými reálnými příklady. Nezávisle proměnná X_1 v našich modelech může například značit pohlaví pacienta (muž/žena) nebo jeho rasu (běloch/černoch), proměnná X_2 pak jeho věk.

Pro demonstraci některých vztahů v modelech budeme nadále v této práci používat příklad závislosti výskytu jisté nemoci na pohlaví a věku pacienta. Regresor značící pohlaví bude v našem případě nabývat hodnoty 1, pokud je pacient muž, a hodnoty 0, pokud je žena.

Zaměříme se zde detailněji na rozdělení regresorů a na výpočet jejich středních hodnot, které budeme později potřebovat.

Pro regresor pohlaví platí

$$X_1 \sim\ \mathrm{Alt}(q)\ ,\ \mathrm{E}X_1 = q\ ,\ q \in (0,1)\ .$$

Regresor věku (X_2) jsme pro modely (3.1) a (3.2) v jednom případě simulovali daty z normálního rozdělení $\mathrm{N}(\mu, \sigma^2)\ ,\ \mu \in \mathbb{R}\ ,\ \sigma^2 > 0$ (používáme klasické značení uvedené například v [2], kapitola 1.3.2) s následujícími parametry a střední hodnotou

$$X_2 \sim \begin{cases} \mathrm{N}(40, 9) & \text{když } X_1 = 1 \\ \mathrm{N}(45, 13) & \text{když } X_1 = 0 \end{cases}\ , \tag{3.3}$$

$$\mathrm{E}X_2 = 40\ q + 45\ (1 - q)$$

V druhém případě jsme předpokládali, že X_2 pochází z gamma rozdělení $\mathrm{Ga}(a, p)\ ,\ a > 0\ ,\ p > 0$, s hustotou a střední hodnotou

$$f(x) = \frac{a^p}{\Gamma(p)}\ \mathrm{e}^{-a\,x} x^{p-1}\ ,\ x > 0\ ,\ \mathrm{E}X = \frac{p}{a}$$

a má parametry a střední hodnotu

$$X_2 \sim \begin{cases} \mathrm{Ga}(2; 0{,}5) & \text{když } X_1 = 1 \\ \mathrm{Ga}(3; 0{,}5) & \text{když } X_1 = 0 \end{cases}\ , \tag{3.4}$$

$$\mathrm{E}X_2 = 4\,q + 6\,(1-q)$$

Ve třetím případě jsme X_2 simulovali jako náhodnou veličinu s logaritmicko normálním rozdělením, $\mathrm{LN}(\mu, \sigma^2)\,,\ \mu \in \mathbb{R}\,,\ \sigma^2 > 0$,

$$f(x) = \frac{1}{\sqrt{2\,\pi}\,\sigma\,x}\ \exp\left\{-\frac{(\log x - \mu)^2}{2\,\sigma^2}\right\}\,,\ x > 0\,,\ \mathrm{E}X = \mathrm{e}^{\mu+\frac{\sigma^2}{2}}\,.$$

Parametry a střední hodnota X_2 jsou v tomto případě

$$X_2 \sim \begin{cases} \mathrm{LN}(3;0{,}3) & \text{když } X_1 = 1 \\ \mathrm{LN}(2{,}5;0{,}4) & \text{když } X_1 = 0 \end{cases}\,, \tag{3.5}$$

$$\mathrm{E}X_2 = \mathrm{e}^{3{,}045}\,q + \mathrm{e}^{2{,}58}\,(1-q)\,.$$

V modelu s interakcemi je navíc regresor X_1X_2, který je s pravděpodobností q roven X_2, s pravděpodobností $(1-q)$ je nulový. Při rozdělení (3.3) (respektive (3.4) a (3.5)) je jeho střední hodnota rovna $40\,q$ (respektive $4\,q$ a $\mathrm{e}^{3{,}045}\,q$).

Pro každý z modelů (3.1) a (3.2) máme tedy tři různé příklady rozdělení veličin.

3.2 Výpočetní postupy

V této části popíšeme postup, který jsme použili pro výpočet hladin významnosti sledovaných testů. Dále ukážeme, jakým způsobem jsme získali empiricky upravené kritické hodnoty asymptotických testů pro měření jejich sil.

Pro modely popsané v kapitole 3.1 jsme testovali tyto hypotézy:
V modelu s interakcemi (3.1) nás zajímalo, zda jsou parametry β_2 a β_3 nenulové. Pokud bychom uvažovali příklad vlivu pohlaví a věku na diagnózu nemoci, chtěli jsme vlastně zjistit, zda existuje závislost výskytu nemoci na věku a zároveň zda obě pohlaví mají průběh výskytu nemoci s věkem rozdílný.

Rozdělme nyní parametr β na dvě části $\beta^T = (\psi^T, \tau^T)$, kde

$$\psi = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}\,,\quad \tau = \begin{pmatrix}\beta_2\\ \beta_3\end{pmatrix}\,.$$

Z kapitoly 2.2 vyplývá, že τ je cílový a ψ rušivý parametr a testujeme tedy složenou hypotézu

$$H_0 : \beta \in B_0 = \left\{\beta \in \mathbb{R}^4 : \tau = (0,0)^T, \psi \in \mathbb{R}^2\right\} \quad \text{proti složené alternativě}$$
$$H_1 : \beta \notin B_0 \,.$$

V jednoduchém modelu (3.2) jsme testovali hypotézu o nulovosti koeficientu γ_2. Cílový parametr je tedy jednorozměrný a nulová hypotéza zní $H_0^+ : \gamma \in B_0^+ = \left\{\gamma_2 = 0, (\gamma_0, \gamma_1)^T \in \mathbb{R}^2\right\}$.

3.2.1 Výpočet hladin významnosti testů

Simulovali jsme nezávislá pozorování se stejným rozdělením vysvětlujících proměnných jako v modelech (3.1) a (3.2) a s rozsahem výběru n. Do vzorců na výpočet testových statistik sledovaných testu odvozených v kapitole 3.3 jsme dosadili nasimulované hodnoty. Statistiky pro naše účely označíme

$$\begin{aligned} T^1(n) &:= 2\ \log \lambda_n^* \,, \\ T^2(n) &:= W_n^* \,, \\ T^3(n) &:= R_n^* \,. \end{aligned} \tag{3.6}$$

Tyto statistiky mají asymptoticky chí-kvadrát rozdělení χ^2_k. Počet stupňů volnosti k je při testování H_0 roven dvěma a při testování H_0^+ je jedna. Hladinu významnosti v této studii volíme $\alpha = 5$ %.

Hypotézu H_0 (respektive H_0^+) jsme zamítali tehdy, když statistiky $T^1(n)$, $T^2(n)$ a $T^3(n)$ s nasimulovanými hodnotami přesáhly hodnotu kvantilu $\chi^2_{k\ (0,95)}$ za platnosti H_0 (respektive H_0^+).

Pro získání odhadu hladin významnosti LR, Waldova a Raova testu při použití kritických hodnot chí-kvadrát rozdělení jsme provedli 5000 opakování výpočtu statistik. Pravděpodobnosti zamítnutí hypotézy jsme odhadli pomocí relativních četností

$$\widehat{P}\left(T^j(n) \geq \chi^2_{k\ (0,95)}\right) = \frac{\sum_{l=1}^{5000} \boldsymbol{I}_{\left[T_v^j \geq \chi^2_{2\ (0,95)}\right]}}{5000} \,,$$

T_v^j je v-tá nasimulovaná hodnota statistiky $T^j(n)$, $j = 1, 2, 3$ a $\boldsymbol{I}$ je indikátor množiny.

3.2.2 Výpočet empiricky upravených sil testů

Nadále budeme používat značení (3.6) z předchozí kapitoly. Předpokládejme, že bychom síly sledovaných testů počítali odhadem pravděpodobností

$$P\left(T^j(n) \geq \chi^2_{k\,(1-\alpha)} \mid H_1\right) ,$$

pro $j = 1, 2, 3$ a k jako v kapitole 3.2.1. Odhady sil testů by v tomto případě byly velmi ovlivněny velikostmi rozdílů mezi α hladinou a hladinou významnosti testů při použití kritických hodnot $\chi^2_{k\,(1-\alpha)}$.

Protože bychom chtěli síly testů rozumně porovnat, potřebujeme použít pro každý test jiné kritické hodnoty. Jedná se o takové kritické hodnoty, které budou za platnosti nulové hypotézy u všech tří testů zamítat tuto hypotézu na stejné α hladině.

Pro testovou statistiku $T^j(n)$ dostaneme kritickou hodnotu $K^j_{k,(1-\alpha)}(n)$, která splňuje vztah

$$P\left(T^j(n) \geq K^j_{k,(1-\alpha)}(n) \mid H_0\right) = \alpha$$

pro $j = 1, 2, 3$. Na základě 5000 simulací odhadneme $K^j_{k,(1-\alpha)}(n)$ při hodnotě $\alpha = 5\%$ a rozsahu výběru n. Pro každý test $j = 1, 2, 3$ bude tedy empirická kritická hodnota rovna 95% kvantilu získaného z 5000 vygenerovaných hodnot testové statistiky $T^j(n)$ za nulové hypotézy.

Takto modifikované síly testů budeme nazývat *empiricky upravené* síly.

3.3 Volba parametrů

Zajímá nás, jak se chování LR, Waldova a Raova testu mění s rostoucím rozsahem výběru. Také chceme znát vliv některých parametrů modelů na toto chování. V této části popíšeme, jaké rozsahy výběrů jsme volili a u kterých

parametrů jsme sledovali vliv.

Rozsahy výběru:

Rozsah výběru značíme v této práci písmenem n. Snažili jsme se volit n taková, která dostatečně postihnou postupnou konvergenci odhadnutých hladin významnosti testů (při použití kritických hodnot $\chi^2_{k\,(0,95)}$ kvantilu) k 5% hladině. Pro malé rozsahy výběru měla metoda maximální věrohodnosti často konvergenční problémy nebo se vyskytovaly časté případy *„dokonale rozdělených“* dat (viz kapitola 3.4). Z těchto důvodů jsme v modelu s interakcemi, zadaném vztahem (3.1), volili nejmenší velikost počtu pozorování 80. Další velikosti n v tomto modelu byly $150, 250, 500, 700$ a 850.

V jednoduchém modelu, zadaném rovností (3.2), byly rozsahy výběru n rovny $30, 50, 80, 150, 250$ a 500.

Parametry modelů:

Dále je třeba zvolit hodnoty parametrů logistických modelů (3.1) a (3.2). Jde o velikost pravděpodobnosti regresoru pohlaví q a o koeficienty modelů β a γ. Hodnoty některých z nich jsme pevně volili podle principu faktoriálového designu, hodnoty ostatních jsme přizpůsobili podle požadavků uvedených na konci této kapitoly.

Podívejme se nyní na volbu parametrů, jejichž hodnoty jsme určovali pevně. Volili jsme je stejné pro oba modely. V úvodu kapitoly 3 jsme se zmínili, že každému parametru přiřadíme jen dvě takové hodnoty. Je to z toho důvodu, že při použití faktoriálového designu se vzrůstajícím počtem kategorií u sledovaných parametrů velmi rychle roste počet potřebných výpočtů. K vysvětlení volby parametrů budeme používat uvažovaný příklad diagnózy nemoci.

- Prvním sledovaným parametrem bude absolutní člen β_0 (respektive γ_0).

  Výraz $\beta_0 + v\,\beta_2$ nebo $\gamma_0 + v\,\gamma_2$ nám dává logaritmus šance na výskyt nemoci u pohlaví, jehož pravděpodobnost je $1 - q$ (v našem příkladu ženy) ve věku v. Označme

  $$O_Z = \mathrm{e}^{\beta_0}\,,$$

  pro γ_0 analogicky. Potom šance na výskyt nemoci u ženy ve věku v je rovna $O_Z \cdot \mathrm{e}^{v\beta_2}$.

Abychom dostatečně pokryli rozsah možných hodnot této šance, volíme první hladinu parametru $\beta_0 = -3{,}6$ a druhou 0,05 (stejně pro γ_0). Šance je potom v prvním případě rovna přibližně $0{,}03 \cdot \mathrm{e}^{v\beta_2}$. V druhém případě je rovna asi $1{,}05 \cdot \mathrm{e}^{v\beta_2}$.

- Druhým sledovaným parametrem bude pravděpodobnost pozitivní odezvy, kterou značíme π.

  Zajímá nás, zda se testy chovají jinak, když nasimulovaná pozorování mají tuto pravděpodobnost nízkou, $\pi = 0{,}15$, a přibližně poloviční, $\pi = 0{,}48$.

- Pouze u modelů s interakcemi (3.1) budeme sledovat koeficient β_3.

  Exponenciála koeficientu interakce věku s pohlavím říká, kolikrát je poměr šancí pro muže (při zvýšení věku o jeden rok) větší než poměr šancí pro ženy (při zvýšení věku o jeden rok). Budeme ji značit $\frac{or_M}{or_Z}$. Hladiny β_3 volíme 0,2 a 0,005. Parametr $\frac{or_M}{or_Z}$ je přibližně 1,22 a 1, tedy poměr šancí pro muže je asi o 20 % větší a poměry jsou téměř stejné.

Volba hladin sledovaných parametrů a velikost hodnot q a zbývajících koeficientů modelů β_1 a β_2 (resp. γ_1 a γ_2) byla provedena následovně.

Vypočetli jsme hodnotu q z parametrů π a β (resp. γ). Postup si ukážeme pouze pro model s interakcemi (3.1) s normálním rozdělením regresoru X_2 (viz (3.3)). Při jiném rozdělení regresorů a pro model (3.2) jsme postupovali podobně.

Při pevných hodnotách π platí

$$\log \frac{\pi}{1-\pi} = \mathrm{E}\left(\log \frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \mathrm{E}X_1 + \beta_2 \mathrm{E}X_2 + \beta_3 \mathrm{E}\left(X_1 X_2\right) .$$

Po dosazení středních hodnot odvozených v kapitole 3.1 dostáváme

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1\, q + \beta_2\, (40q + 45 - 45q) + \beta_3\, 40q$$

$$q = \frac{\ln \frac{\pi}{1-\pi} - \beta_0 - \beta_2\ 45}{\beta_1 + \beta_2\ (40 - 45) + \beta_3\ 40}\ .$$

Protože q je pravděpodobnost, musí být splněno $q \in (0\ ,1)$, to částečně omezilo volbu hodnot parametrů.

Další omezující podmínkou byl požadavek, aby síly testů v intervalech vymezených rozsahy výběru pokrývaly co nejširší spektrum hodnot. Z toho důvodu jsme volili parametry tak, aby odhady velikosti sil testů byly při 200 pozorování přibližně v intervalu $\langle 0{,}4\ ;0{,}6\rangle$. Odhady sil pro tyto účely byly čistě orientační a proto nám stačily odhady při použití kritických hodnot $\chi^2_{k\ (0{,}95)}$.

Dodejme, že pro nízký počet pozorování bylo třeba se vyvarovat případů častého výskytu „dokonalého rozdělení“ dat, které popisujeme v následující kapitole 3.4. Výskyt těchto případů také ovlivnil volbu hodnot parametrů modelu.

Poznámka:
Dále budeme v textu používat značení $OR_M = \mathrm{e}^{\beta_1}$ (stejné pro γ_1). Poměr šancí na nemoc u mužů proti ženám ve věku v v našem příkladu je tedy roven $OR_M \cdot \mathrm{e}^{v\beta_3}$.

3.4 Poznámky k postupu při simulacích

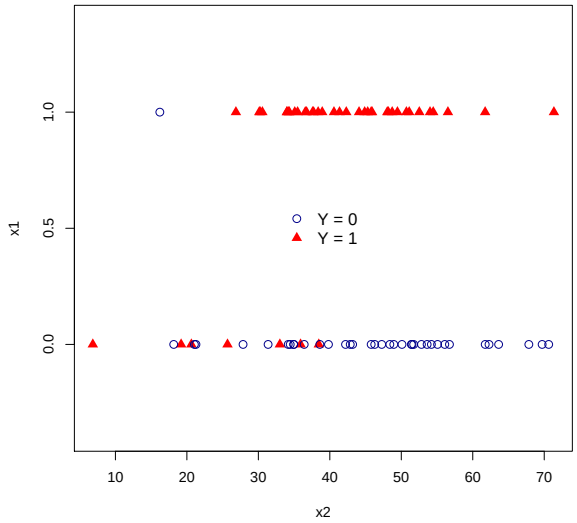
Zdrojové kódy k vybraným funkcím jsou uvedeny na přiloženém CD.

Během simulačních výpočtů docházelo k situacím, kdy nebylo možno správně vypočítat testové statistiky. V této části popíšeme o jaké situace šlo a pomocí jakých úprav ve zdrojových kódech jsme je řešili.

Hodnoty testových statistik nebylo možno zjistit v situacích, kdy rozsah výběru nasimulovaných dat byl malý a data byla tudíž řídká. Algoritmus v těchto případech nebyl schopen správně vypočítat odhady modelu a dával zkreslené výsledky. Důvodem byl výskyt jevu, který je popsaný v [1] (kapitola 5.5.5) jako *'perfect discrimination'*.

Dokonalé rozdělení (*'perfect discrimination'*) dat je stav, při kterém lze prostor hodnot regresoru rozdělit nadrovinou na část, kde pro všechna pozorování je $Y = 0$, a část, kde je pro ně $Y = 1$. To znamená, že můžeme

přesně předpovědět odezvu, pokud známe hodnotu regresoru (mimo hraniční místa). V tomto případě ale maximálně věrohodné odhady buď neexistují nebo jsou nekonečné. Odhady koeficientů, pokud je program vypočítá, jsou tedy zkreslené. Na obrázku 3.1 je graficky znázorněn jeden z výskytů tohoto jevu u našich dat.



Obrázek 3.1: Příklad dokonalého rozdělení dat pro $x_1 = 1$

Popsaný problém jsme vyřešili následovně. Při jeho vzniku jsme dokonale rozdělená data vyloučili, vrátili jsme se na začátek zdrojového kódu a nasimulovali jsme místo nich nové hodnoty. Postup jsme opakovali, dokud tvorba modelu neproběhla v pořádku.

Tímto způsobem může docházet ke změně rozložení dat. Naší prioritou je, aby data byla podobná reálným hodnotám, a proto nám malé změny nevadí. Chceme se ale vyvarovat větších odchylek od stanoveného rozdělení. Zaznamenávali jsme proto, kolik nasimulovaných dat bylo nutno nahradit (opakované „opravy“ jsme považovali za jedno nahrazení), a pokud procento takových dat přesahovalo 8,5 procenta, nepovažovali jsme je za důvěryhodné. Takové případy nastaly především při 80 pozorováních.

Další komplikace vznikaly při volbě hodnoty parametru q blízko 1 nebo 0. Stávalo se, že se nasimulovalo 80 pozorování, ve kterých se všechna x_1 (nebo

79 z nich) rovnala jedné nebo nule. Taková data jsme také vyloučili (v realitě bychom také nezkoumali rozdíly mezi pohlavími ve výběru se samými muži).

Poznámka:
Abychom urychlili výpočty jednotlivých statistik, počítali jsme inverze pozitivně definitních a tedy i symetrických matic pomocí metody Choleského dekompozice. Jednalo se o varianční matici var($\hat{\beta}$) a části Fisherovy informační matice. Ve srovnání s klasickou metodou Gaussovy eliminace tento postup zajistil několikanásobně rychlejší provedení výpočtů. (Pro matici 4×4 byl výpočet inverze asi 4,5 krát rychlejší při použití metody Choleského dekompozice než při LU rozkladu.)

Kapitola 4

Výsledky simulačních studií

Tato kapitola seznamuje s výsledky, které jsme získali simulováním dat za použití postupů uvedených v kapitole 3. Vybrané výstupy jsme znázornili pomocí tabulek a grafů. Na základě zjištěných výsledků popíšeme chování LR, Wald a Rao testu pro model s interakcemi zadaný vztahem (3.1) a pro jednoduchý model (3.2). Nadále budeme používat značení zavedené v minulé kapitole a příklad diagnózy nemoci pro popis některých vztahů.

4.1 Výsledky v modelu s interakcemi

Na základě principů faktoriálového designu jsme simulovali nezávislá pozorování pro model s interakcemi (3.1) s rozdělením regresorů popsané v části 3.1. V modelu jsme volili kombinace parametrů podle kapitoly 3.3.

Některé vlastnosti testů jsme znázornili pomocí obrázků složených ze čtyř grafů (viz 4.2, 4.1 a 4.3). Každý graf odhadů hladin významnosti testů (při použití kritických hodnot $\chi^2_{2\,(0,95)}$) má pod sebou graf odhadů „empiricky upravených“ sil testů, který s ním souvisí. Odhady hladin a empirické kritické hodnoty pro odvození sil byly získány za platnosti stejné nulové hypotézy.

Odhady hladin i sil jsou v grafech znázorněny v závislosti na rozsahu výběru (znač. n). Jednotlivé hodnoty jsou v nich pro větší názornost spojeny přímkami. Pro $n = 80$ jsme byli nuceni některé výsledky vyřadit, protože jejich rozdělení bylo příliš ovlivněno častým nahrazováním vygenerovaných dat (viz Kapitola 3.4).

Poznámka:
Hladiny a síly odhadujeme pomocí relativních četností. Při 5000 simulací měly naše výsledky maximální rozdíl mezi horním a dolním intervalem spolehlivosti odhadu hladiny roven 0,014, pro odhad síly byla tato vzdálenost 0,0278.

4.1.1 Vlastnosti hladin testů

Hladiny významnosti testů získané při použití kritických hodnot $\chi^2_{2\,(0,95)}$ budeme dále nazývat jen hladiny. Tuto sekci rozdělíme na dvě části podle zvolené velikosti parametru modelu O_Z.

Nejprve popíšeme výsledky při takové volbě koeficientu β_0, že

- $O_Z = 0{,}03$.

Pokud je šance na výskyt nemoci u v-leté ženy v našem příkladu rovna $0{,}03 \cdot \mathrm{e}^{v\beta_2}$, pro obě hodnoty parametru π se testy chovají téměř stejně a zachovávají totožné rozdíly ve velikostech odhadů hladin. Chování hladin je velmi podobné i pokud změníme rozdělení regresoru věku X_2 z gamma (3.4) na lognormální (3.5) nebo normální (3.3). Na obrázku 4.1 uvádíme vybrané grafy znázorňující toto chování.

Waldův test se v těchto případech chová konzervativně. Pro malé výběry je konzervativní natolik, že odhady hladin při 150 pozorováních dosahují maximálně 2%. Pro větší výběry jeho hladina vzrůstá, přesto i pro $n = 850$ nepřekračují odhady v některých případech 3,5 %.

Test založený na věrohodnostním poměru (LR test) a Raův test mají velmi podobné chování. Při gamma a lognomálním rozdělení X_2 se odhady hladin většinou překrývají nebo se liší jen o jednu nebo dvě desetiny procenta. Při normálně rozděleném X_2 je rozdíl výraznější. Z našich výsledků usuzujeme, že oba dva jsou dobře aproximovány pomocí kritických hodnot rozdělení chí-kvadrát pro počet pozorování větší než 500.

Dále uvádíme výsledky zjištěné v modelech s šancí $1{,}05 \cdot \mathrm{e}^{v\beta_2}$, že žena ve věku v bude shledána nemocnou,

- $O_Z = 1{,}05$.

Pro tuto velikost parametru O_Z jsme na základě simulací zjistili následující rozdíly v chování testů při $\pi = 48\%$ a při pravděpodobnosti odezvy 15%.

$\pi = 48\%$

Stejně jako při nízké hodnotě π_Z je i v tomto případě chování testů při všech třech uvažovaných rozděleních podobné. Pro model s gamma rozdělením regresoru X_2 jsou odhady hladin při $\pi = 0{,}48$ znázorněny na obrázku 4.2. Z grafu je vidět, že testy zachovávají stejné pořadí jako na obr. 4.1. Liší se ve velikostech odhadnutých hladin.

LR test se při $\pi = 0{,}48$ chová mírně liberálně. Pro $n = 250$ překračuje 6%, pro velikost rozsahu výběru 500 a vyšší je už pod horní hranicí intervalu spolehlivosti odhadu 5% hladiny, která je rovna 5,6%.

Průběh odhadu hladiny Raova testu kopíruje průběh odhadu hladiny LR testu s tím, že Raův test je ve všech případech mírně konzervativnější. V tomto případě tedy hladina Raova testu, zvláště při nízkém rozsahu výběru, je blíže k 5% a Raův test se proto více hodí pro testování.

Poznámka:
Na tomto místě musíme poznamenat, že parametr β_1 pro výpočet hladin byl volen tak, aby bylo možno vypočítat a porovnávat i empiricky upravené síly testů příslušející k dané hladině. Nebylo možno volit jeho velikost stejnou pro každou kombinaci parametrů, a proto musíme počítat i s jeho vlivem na chování odhadů hladin. V obr. 4.2 pro případ $\pi = 0{,}48$ bylo β_1 voleno tak, že $OR_M = 0{,}3$. Dále v textu ukážeme, že tento parametr také velmi ovlivňuje velikosti hladin.

$\pi = 15\%$

Nejprve se zaměříme na model s gamma rozdělením. Na obrázku 4.2 jsou vidět rozdíly v chování testů při $\pi = 0{,}48$ a $\pi = 0{,}15$. Při volbě rozsahu výběru 250, 500 a 700 pozorování došlo ke změně pořadí LR a Raova testu. Odhady hladiny Raova testu se ve srovnání s odhady hladiny testu poměrem věrohodností rychleji blíží k 5 %. Waldův test je stejně jako v případě nízké hodnoty π_Z velmi konzervativní. Pro malé počty pozorování jsou konzervativní i LR a Raův test. Je třeba vzít v úvahu i to, že v případě $\pi = 0{,}15$ je $OR_M = 0{,}013$ a tedy poměr šancí na nemoc u muže oproti ženě ve věku v je o dost nižší než při $\pi = 0{,}48$, stejném věku a stejném interakčním koeficientu β_3.

Chování hladin testů při lognormálním a normálním rozdělení regresoru X_2 je stejné. Dále proto popisujeme výsledky jen pro model s normálně rozděleným regresorem věku.

Pro $\pi = 0{,}15$ jsme v modelu volili parametr $\beta_1 = -20$. Parametr OR_M byl v tomto případě tedy téměř nulový a proto i pravděpodobnost, že se například u 40-letého muže vyskytne daná nemoc (při volbě $\beta_2 = 0{,}044$, $\beta_3 = 0{,}2$), byla velmi blízká nule. Taková volba velmi ovlivnila naše výsledky. Všechny testy se v tomto případě chovají vysoce konzervativně a pro sledované velikosti výběru zůstávají na 1,5%.

Vliv OR_M:

Zkoumali jsme dále, zda a jak rychle v tomto případě konvergují experimentální odhady hladiny k žádaným pěti procentům. Odhady byly i zde provedeny na základě 5000 opakování simulací a jsou znázorněny v tabulce 4.1.

Protože hladiny za takové volby OR_M nekonvergovaly ani při 7000 pozorování (k zjištění odhadů hladin pro větší hodnoty rozsahu jsme neměli výpočetní prostředky), zvýšili jsme velikost β_1 na -8 ($OR_M = 3{,}4 \cdot 10^{-4}$). Je vidět, že velikosti odhadů hladin testů zachovávají stejné pořadí jako v ostatních případech, ale konvergence je i v tomto případě opravdu velmi pozvolná.

Tabulka 4.1: Odhady hladin významnosti testů při použití kritických hodnot $\chi^2_{2\,(0,95)}$ pro testování $\beta_2 = \beta_3 = 0$ v modelu $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$ s alternativním rozdělením X_1 a normálním rozdělením X_2, kde g je linková funkce, $\mathrm{E}Y = 0{,}15$, $\mathrm{e}^{\beta_0} = 1{,}05$ a e^{β_1} značíme OR_M.

	$OR_M = 2 \cdot 10^{-9}$			$OR_M = 3{,}4 \cdot 10^{-4}$		
Rozsah výběru	**LR**	**Wald**	**Rao**	**LR**	**Wald**	**Rao**
2000	0,0172	0,0168	0,0172	0,0190	0,0154	0,0190
4000	0,0136	0,0132	0,0132	0,0236	0,0192	0,0224
6000	0,0154	0,0152	0,0154	0,0270	0,0212	0,0256
7000	0,0160	0,0160	0,0160	0,0348	0,0268	0,0340

Dále jsme zvyšovali velikost parametru β_1. Na obrázku 4.3 jsme znázornili,

jak při stejné volbě O_Z a π a rostoucí hodnotě β_1 (tedy i rostoucí šanci muže proti ženě na výskyt nemoci v daném věku) se mění velikost hladin jednotlivých testů. Například LR test je pro $OR_M = 0{,}013$ konzervativní, pro $OR_M = 0{,}135$ je mírně liberální.

Na základě dalších simulací v modelech s lognormálním a gamma rozdělením X_2 při $O_Z = 1{,}05$ jsme došli k závěru, že i při těchto rozděleních platí, že při rostoucí velikosti parametru OR_M rostou velikosti odhadů hladin testů.

Rozdíly ve velikostech hladin jednotlivých testů při volbě parametru $\pi = 0{,}15$ a $\pi = 0{,}48$ při gamma, lognormálním nebo normálním rozdělení regresoru X_2 jsou podle tohoto zjištění způsobeny odlišnou volbou parametru OR_M.

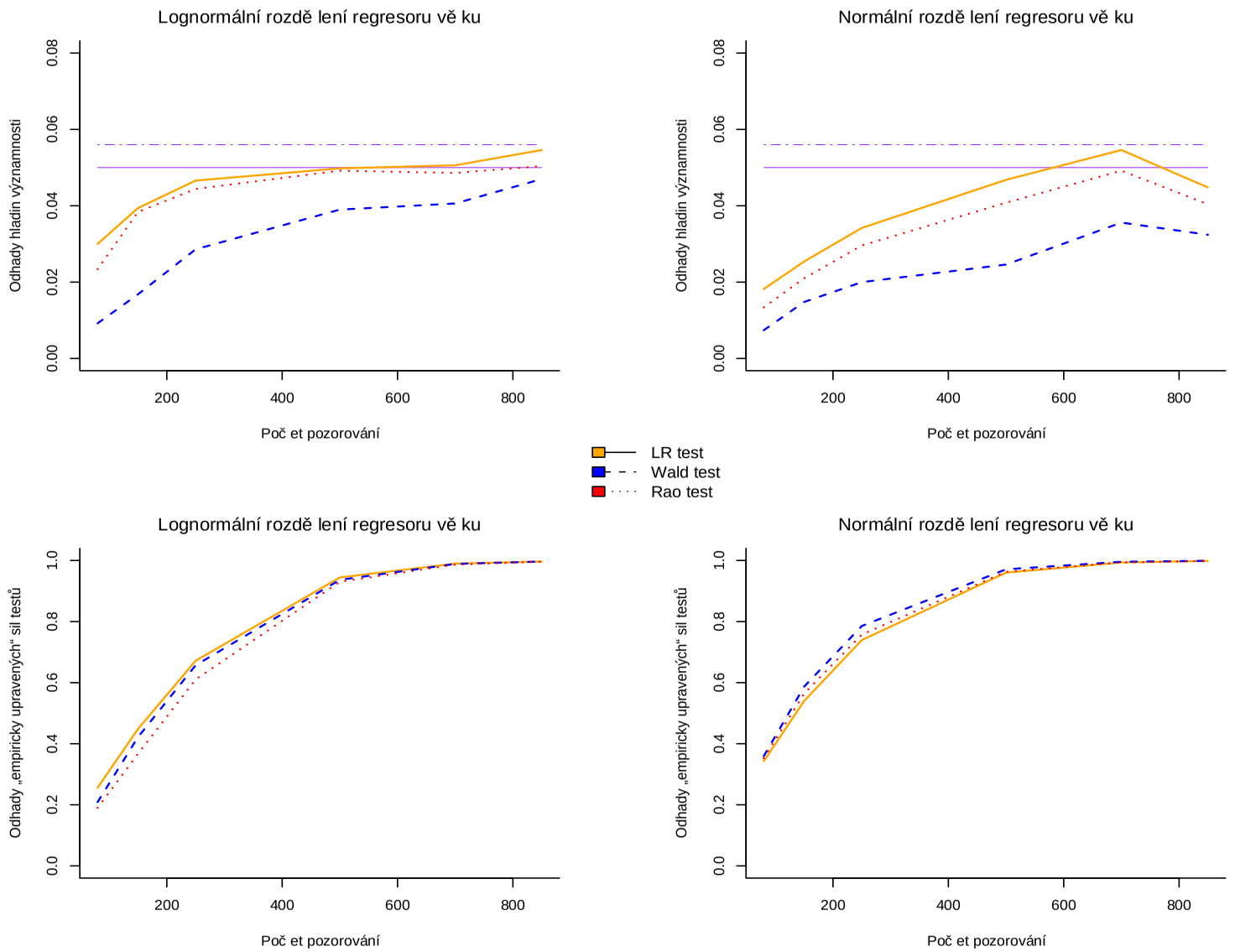
Obecně platilo, při $O_Z = 1{,}05$ a velmi nízké volbě velikosti OR_M a při $O_Z = 0{,}03$ a velmi vysoké volbě velikosti OR_M se všechny testy chovají vysoce konzervativně i při rozsahu výběru v řádu tisíců.

Souhrnem, testujeme-li hypotézu o nulovosti parametrů β_2 a β_3 v modelu zadaném vztahem (3.1) při alternativním rozdělením regresoru X_1 a normálním rozdělení X_2 (viz (3.3)), je vzhledem k aproximaci kritických hodnot $\chi^2_{2\,(0{,}95)}$ kvantilem ve většině případů nejvhodnější použít test poměrem věrohodností. Dosahuje totiž 5% při nižších hodnotách velikosti výběru než Raův nebo Waldův test.

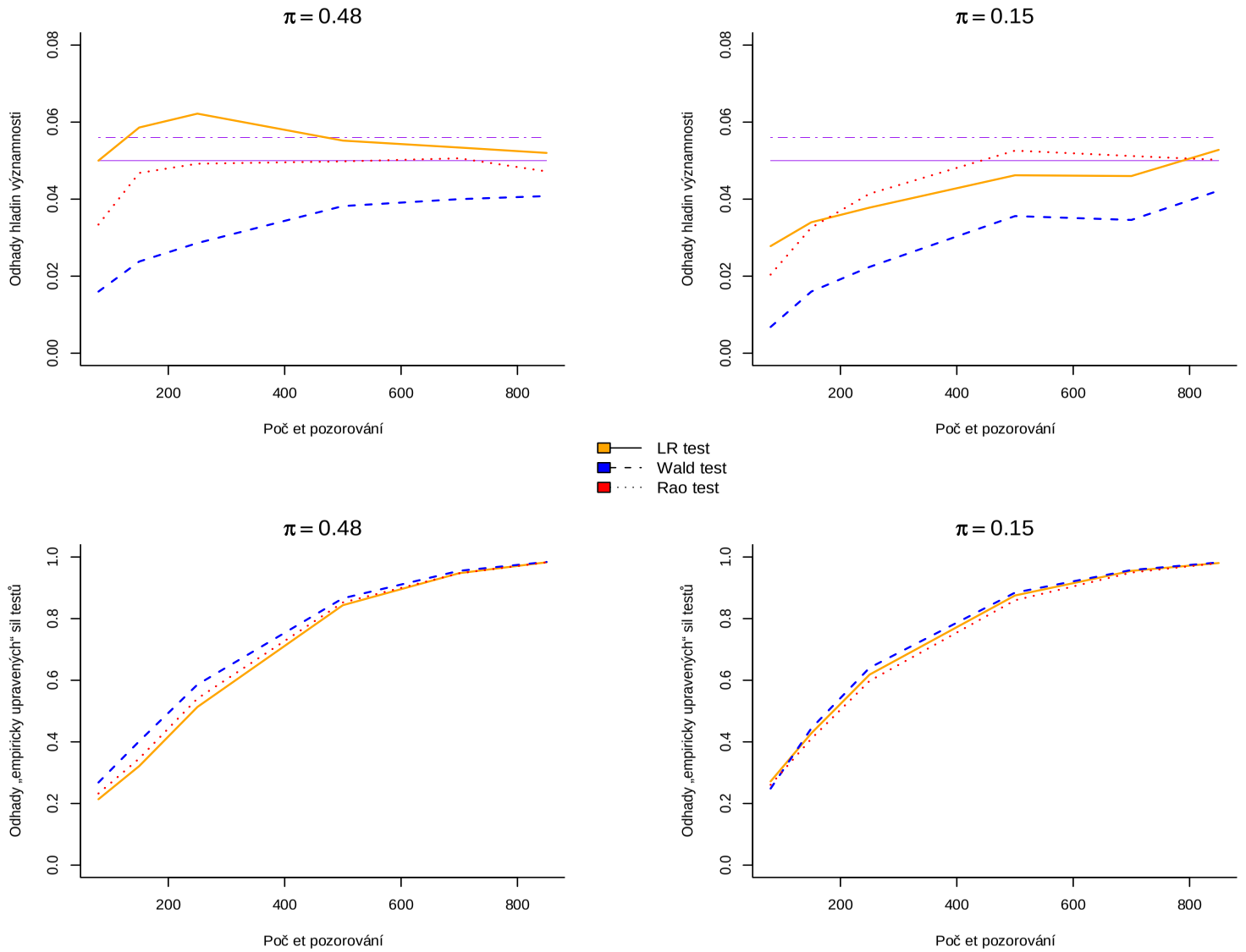
Pouze v některých případech LR test zamítá hypotézu častěji, než by měl. Tyto situace nastávají při $O_Z = 1{,}05$ a velikosti výběru menší než 500 a na jejich vznik má vliv parametr β_1. V těchto případech je vhodné použít hodnoty skórového testu.

Poznamenejme ještě, že při použití Waldova testu je, podle našich výsledků, lepší zamítat hypotézy na základě kritických hodnot chí-kvadrát s vyšší hladinou významnosti než 5%, abychom dosáhli cílové 5% (například při 150 pozorováních až o tři procenta vyšší).

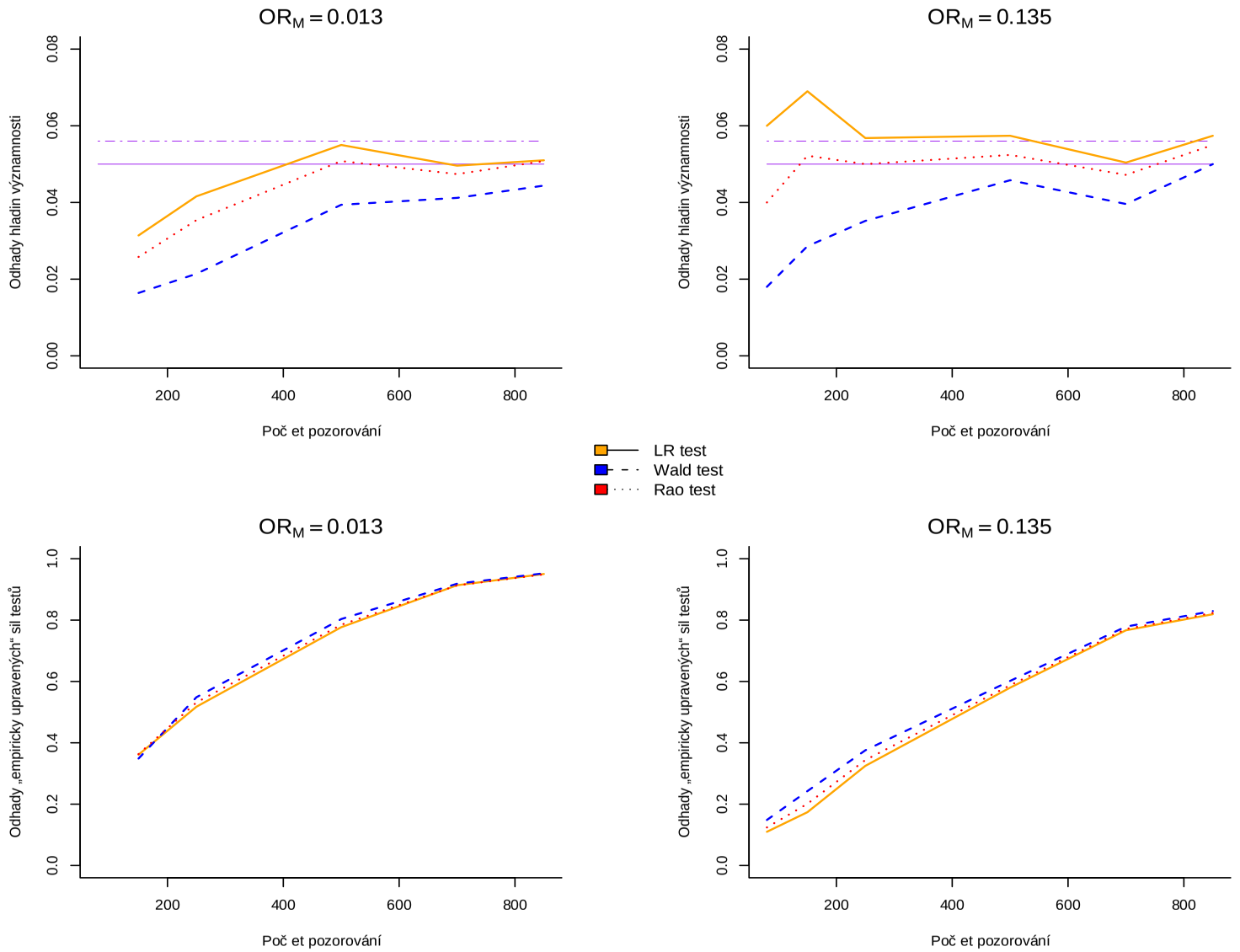
Pokud rozdělení X_2 změníme na lognormální nebo gamma (viz (3.5) a (3.4)), hladiny Raova a LR testu dosahují v některých případech stejných hodnot, výjimečně Raův test dosahuje 5 procent pro nižší rozsah výběru než test poměrem věrohodností.



Obrázek 4.1: Empirické chování hladin testů při použití kritických hodnot $\chi^2_{2\ (0,95)}$ pro testování $\beta_2 = \beta_3 = 0$ a síly testů odvozených na základě empirických kritických hodnot za stejné hypotézy. Odhady jsou založeny na 5000 simulací pro model $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$ s alternativním rozdělením X_1 a normálním nebo lognormálním rozdělením regresoru věku X_2, kde g je linková funkce, $EY = 0{,}15$, $e^{\beta_0} = 0{,}027$, $e^{\beta_3} = 1{,}22$. Při normálním rozdělení X_2 je $\beta_1 = 2{,}2$, $\beta_2 = -0{,}15$, při lognormálním $\beta_1 = 2{,}8$, $\beta_2 = -0{,}09$. Vyznačena horní mez intervalu spolehlivosti odhadu 5% (rovna 5,6 %).



Obrázek 4.2: Empirické chování hladin testů při použití kritických hodnot $\chi^2_{2\,(0,95)}$ pro testování $\beta_2 = \beta_3 = 0$ a sil testů odvozených na základě empirických kritických hodnot za stejné hypotézy. Odhady jsou založeny na 5000 simulací pro model $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$ s alternativním rozdělením X_1 a gamma rozdělením X_2, kde g je linková funkce, $e^{\beta_0} = 1{,}05$, $e^{\beta_3} = 1{,}005$. Při $EY = 0{,}48$, je $\beta_1 = -1{,}2$, $\beta_2 = 0{,}1$, při $EY = 0{,}15$ je $\beta_1 = -4{,}35$, $\beta_2 = 0{,}15$. Vyznačena horní mez intervalu spolehlivosti odhadu 5% (rovna 5,6 %).



Obrázek 4.3: Empirické chování hladin testů při použití kritických hodnot $\chi^2_{2\,(0,95)}$ pro testování $\beta_2 = \beta_3 = 0$ a sil testů odvozených na základě empirických kritických hodnot za stejné hypotézy. Odhady jsou založeny na 5000 simulací pro model $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$ s alternativním rozdělením X_1 a normálním rozdělením X_2, kde g je linková funkce, $EY = 0{,}15$, $e^{\beta_0} = 1{,}05$, $e^{\beta_3} = 1{,}005$. Při $e^{\beta_1} = OR_M = 0{,}013$, je $\beta_2 = 0{,}044$, při $OR_M = 0{,}135$ je $\beta_2 = -0{,}025$. Vyznačena horní mez intervalu spolehlivosti odhadu 5% (rovna 5,6 %).

4.1.2 Porovnání sil testů

„Empiricky upravené“ síly sledovaných testů (dále jen síly) jsme získali použitím empirických kritických hodnot vypočtených na základě 5000 hodnot testových statistik za nulové hypotézy. Pro velikost výběru 250 uvádíme v tabulce 4.2 síly při gamma rozdělení regresoru X_2 a v tabulce 4.3 síly při lognormálním rozdělení X_2 spolu s kritickými hodnotami použitými k určení sil. Rozdíly v silách testů při normálním rozdělení X_2 byly stejné jako při lognormálním.

Tabulka 4.2: Síly testů odvozené na základě empirických kritických hodnot založených na 5000 simulací pro testování $\beta_2 = \beta_3 = 0$ na 5% hladině v modelu $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$ s alternativním rozdělením X_1 a **gamma** rozdělením X_2, kde g je linková funkce, $O_Z = \mathrm{e}^{\beta_0}$, $\pi = EY$ a $\frac{or_M}{or_Z} = \mathrm{e}^{\beta_3}$.

Rozsah výběru: **250**			Síly testů			Empir. krit. hodnoty		
O_Z	π	$\frac{or_M}{or_Z}$	**LR**	**Wald**	**Rao**	**LR**	**Wald**	**Rao**
1,05	0,48	1,22	0,535	0,535	0,526	6,119	5,571	5,915
1,05	0,48	1,00	0,513	0,586	0,541	6,464	5,189	5,968
1,05	0,15	1,22	0,595	0,657	0,652	6,128	5,429	5,939
1,05	0,15	1,00	0,618	0,640	0,598	5,472	4,752	5,522
0,03	0,48	1,22	0,553	0,560	0,497	5,415	4,509	5,289
0,03	0,48	1,00	0,594	0,646	0,600	5,494	4,502	5,261
0,03	0,15	1,00	0,696	0,740	0,680	5,797	4,871	5,669
0,03	0,15	1,22	0,656	0,727	0,688	5,918	4,832	5,790

Nezjistili jsme žádné výraznější rozdíly v silách testů. Při gamma rozdělení X_2 je Waldův test ve všech případech nejsilnější. Při lognormálním a tedy i normálním rozdělení je v některých případech silnější test poměrem věrohodností. Další pořadí sil se pro různé parametry modelu liší. Při stejných parametrech a rostoucím rozsahu zůstává pořadí stejné a rozdíly mezi silami se zmenšují (viz například obr. 4.1).

4.2 Výsledky v jednoduchém modelu

V modelu 3.2 nebyl interakční koeficient a testovali jsme tedy pouze jednoduchou hypotézu o nulovosti β_2. Chování testů nebylo závislé na hodnotách

Tabulka 4.3: Síly testů odvozené na základě empirických kritických hodnot založených na 5000 simulací pro testování $\beta_2 = \beta_3 = 0$ na 5% hladině v modelu $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$ s alternativním rozdělením X_1 a **lognormálním** rozdělením X_2, kde g je linková funkce, $O_Z = \mathrm{e}^{\beta_0}$, $\pi = EY$ a $\frac{or_M}{or_Z} = \mathrm{e}^{\beta_3}$.

Rozsah výběru: **250**			Síly testů			Empir. krit. hodnoty		
O_Z	π	$\frac{or_M}{or_Z}$	**LR**	**Wald**	**Rao**	**LR**	**Wald**	**Rao**
1,05	0,48	1,22	0,679	0,663	0,640	6,313	5,422	5,913
1,05	0,48	1,00	0,592	0,687	0,629	6,465	5,035	5,876
1,05	0,15	1,22	0,662	0,652	0,595	6,473	5,332	6,040
1,05	0,15	1,00	0,703	0,756	0,716	5,643	4,769	5,626
0,03	0,48	1,22	0,631	0,709	0,684	5,292	4,378	5,120
0,03	0,48	1,00	0,648	0,710	0,672	5,477	4,481	5,194
0,03	0,15	1,22	0,672	0,657	0,611	5,814	5,067	5,801
0,03	0,15	1,00	0,553	0,592	0,564	5,814	5,011	5,702

sledovaných parametrů modelu. Při všech třech uvažovaných rozděleních X_2 se testy chovaly velmi podobně. V tabulce 4.4 jsou výsledky odhadů hladin pro nízké rozsahy výběru ($n = 80$ a $n = 150$). Pro rozsah výběru velikosti 30 byly rozdíly mezi testy ještě extrémnější, ale docházelo k častým případům dokonalého rozdělení dat. Při počtu pozorování 200 a vyšší byly testy záměnné, jejich odhady téměř přesně kopírovaly 5 procent.

Porovnáním „empiricky upravených“ sil testů jsme došli k závěru, že neexistují žádné rozdíly v síle testů.

Poznámka:
Při 5000 simulací měly výsledné odhady pomocí relativních četností maximální rozdíl mezi horním a dolním intervalem spolehlivosti odhadu hladiny významnosti roven 0,0146, při odhadu síly byla tato vzdálenost 0,0278.

Tabulka 4.4: Empirické chování hladin testů při použití kritických hodnot $\chi^2_{2\,(0,95)}$ pro testování $\beta_2 = 0$. Odhady jsou založeny na 5000 simulací pro model $g(EY) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ s alternativním rozdělením X_1 a normálním rozdělením X_2, kde g je linková funkce, n je rozsah výběru, $O_Z = \mathrm{e}^{\beta_0}$, $\pi = EY$ a $OR_M = \mathrm{e}^{\beta_1}$.

$\boldsymbol{n = 80}$			Test		
O_Z	π	OR_M	**LR**	**Wald**	**Rao**
1,05	0,48	0,05	0,0520	0,0442	0,0494
1,05	0,15	0,05	0,0588	0,0426	0,0536
0,03	0,48	244,7	0,0556	0,0398	0,0494
0,03	0,15	244,7	0,0544	0,0294	0,0452
$\boldsymbol{n = 150}$			Test		
π_Z	π	OR_M	**LR**	**Wald**	**Rao**
1,05	0,48	0,05	0,0460	0,0410	0,0450
1,05	0,15	0,05	0,0488	0,0418	0,0468
0,03	0,48	244,7	0,0572	0,0502	0,0550
0,03	0,15	244,7	0,0524	0,0408	0,0486

Kapitola 5

Závěr

V této práci jsme zkoumali chování tří asymptotických testů (Raova, Waldova a poměrem věrohodností) na základě simulační studie. Zaměřili jsme se na logistické modely, které jsou jedny z prakticky velmi používaných regresních modelů.

Otázkou chování asymptotických testů se zabývá více článků, které tuto problematiku zkoumají pomocí různých simulačních přístupů. Tyto studie jsou však schopny popsat jen některé konkrétní modely s konkrétními parametry, proto u mnoha modelů nebyly vlastnosti těchto testů dosud zjištěny. Touto prací jsme chtěli přispět k prozkoumání chování asymptotických testů ve dvou modelech logistické regrese.

Uvažovali jsme takový model, kde je binární odezva závislá na regresoru s alternativním rozdělením, na regresoru s jistým spojitým rozdělením a jejich interakci. Také jsme zkoumali model bez interakcí.

Volba modelu byla motivována situací, kdy zkoumáme výskyt jistého jevu (například nemoci) v závislosti na pohlaví, věku a jejich interakci a zajímá nás, zda věk a interakce věku a pohlaví významně ovlivňuje daný jev. Jednalo se tedy o test hypotézy nulovosti dvou koeficientů nebo jednoho v případě modelu bez interakcí. Cílem bylo zjistit, jak se testy liší v odhadech hladiny významnosti získané aproximací pomocí kritických hodnot chí-kvadrát rozdělení ($\alpha = 5\%$), a co nejlépe porovnat, jak jsou při těchto odhadech silné. Zajímalo nás, jak se toto chování mění při rostoucím počtu pozorování a jak ho ovlivňuje volba některých parametrů regresního modelu.

Zjistili jsme, že v modelu s interakcemi, kde regresor značící věk měl normální, gamma nebo lognormální rozdělení, je chování testů pro různé parametry modelu velmi podobné.

Test poměrem věrohodností vedl v některých případech k zvýšené hladině významnosti, zvláště při rozsahu výběru menším než 300 a při volbě takového absolutního členu β_0, že e^{β_0} bylo rovno přibližně jedné (viz Kapitola 4.1.1). Vliv na liberální chování měl v těchto situacích parametr určující poměr šancí na výskyt nemoci mezi pohlavími. V ostatních případech byl test pro malou velikost výběru mírně konzervativní, při větších výběrech byl již v mezích intervalu spolehlivosti pětiprocentní hladiny, a tedy velmi vhodný.

Raův (skórový) test se choval většinou o něco konzervativněji než test poměrem věrohodností, v některých případech (při modelaci X_2 sešikmenými rozděleními) s ním byl záměnný, ale nikdy nepřekročil horní odhad pětiprocentní hladiny.

Waldův test byl z testů nejkonzervativnější. Při nejvyšším sledovaném rozsahu výběru (850) dával hladinu významnosti kolem čtyř procent.

Síly testů jsme zjišťovali pomocí empirických odhadů kritických hladin, aby se daly porovnávat. Při gamma rozdělení regresoru věku byl Waldův test ve všech případech nepatrně silnější, při dalších dvou uvažovaných rozděleních měl při některé volbě parametrů větší sílu test poměrem věrohodností. Mezi testy však neexistovaly výraznější rozdíly.

V modelu bez interakcí, kde se testovala jednoduchá hypotéza, měly všechny testy velmi dobré vlastnosti vzhledem k aproximaci kritických hodnot rozdělením chí-kvadrát. Už při rozsahu výběru 150 byla jejich odhadnutá hladina významnosti ve většině případů pětiprocentí. Jejich „empiricky upravené“ síly byly záměnné.

Pro zvolené modely logistické regrese jsme ověřili domněnku, která se vyskytuje v literatuře, že Waldova statistika aproximovaná chí-kvadrát rozdělením má pro menší výběry ze všech tří statistik nejhorší vlastnosti. Pro testování hypotéz v těchto modelech je vhodné používat test založený na statistice poměrem věrohodností, pokud se chceme vyhnout případům mírně liberálněj- šího odhadu hladiny, je dobré zvolit test Raův.

Literatura

[1] Alan Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 709 p. , 2nd edition, 2002.

[2] J. Anděl. *Základy matematické statistiky*. Preprint. Matematicko-fyzikální fakulta Univerzity Karlovy, Praha, 2002.

[3] M. Kulich. Asymptotické testy hypotéz v modelech s rušivými parametry. *Antoch, J., Dohnal, G. (Eds.) ROBUST 2000, Sborník 11. letní školy JČMF*, 125–134, 2001.

[4] P. McCullagh and J. A. Nelder. *Generalized linear models*. London : Chapman & Hall/ CRC Press. 511 p., 2nd edition, 1999.

[5] B.C. Sutradhar and R.F. Bartlett. Monte Carlo comparison of Wald's, likelihood ratio and Rao's tests. *J. Statist. Comput. Simul.*, 46:23–33, 1993.