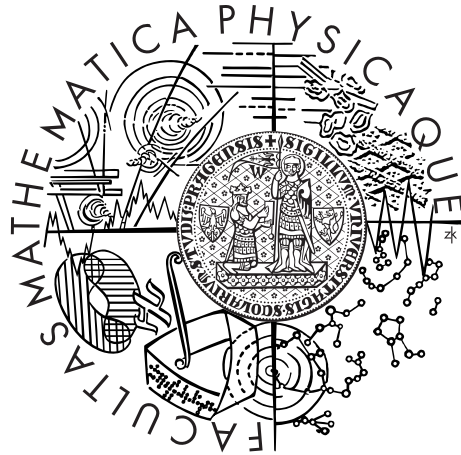


Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bc. Anežka Faltýnková

Modely s kategoriální odezvou

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jan Kalina, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční a pojistná matematika

Praha 2015

Ráda bych na tomto místě poděkovala vedoucímu své diplomové práce panu RNDr. Janu Kalinovi, Ph.D. za vstřícný přístup, cenné rady a čas, který věnoval konzultacím a četbě mé práce. Za cenné rady rovněž děkuji svému internímu konzultantovi doc. RNDr. Arnoštu Komárkovi, Ph.D.

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) 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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Název práce: Modely s kategoriální odezvou

Autor: Bc. Anežka Faltýnková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jan Kalina, Ph.D., Ústav informatiky AV ČR

Abstrakt: Tato práce studuje regresní modely s kategoriální odezvou. Zaměřuje se na logistickou regresi s binární odezvou a na její zobecnění v podobě multinomické regrese s odezvou multinomickou, u které se zabývá dvěma modely: s nominální a s ordinální odezvou. Práce dále obsahuje odvození Waldova testu a testu poměrem věrohodností pro všechny tři studované modely. Toto teoretické odvození je použito ke spočítání příslušných testových statistik u konkrétních příkladů v statistickém softwaru R. V práci uvedená teorie je ilustrována na příkladech s menším i větším počtem regresorů.

Klíčová slova: logistická regrese, multinomická regrese, teorie maximální věrohodnosti, Waldův test, test poměrem věrohodností

Title: Models with categorical response

Author: Bc. Anežka Faltýnková

Department: Department of Probability and Mathematical Statistic

Supervisor: RNDr. Jan Kalina, Ph.D., Institute of Computer Science AS CR

Abstract: This thesis concentrates on regression models with a categorical response. It focuses on the model of logistic regression with binary response and its generalization in which two models are distinguished: multinomial regression with nominal response and multinomial regression with ordinal response. For all three models separately, the Wald test and the likelihood ratio test are derived. These theoretical derivations are then used to calculate the test statistics for specific examples in statistical software R. The theory described in the thesis is illustrated by examples with small and large number of explanatory variables.

Keywords: logistic regression, multinomial regression, maximum likelihood theory, Wald test, likelihood ratio test

Obsah

Úvod	2
1 Logistická regrese	3
1.1 Teorie maximální věrohodnosti	3
1.2 Odhad parametrů v modelu logistické regrese	4
1.3 Dummy proměnné	6
1.4 Logistická regrese-příklad	6
2 Multinomická regrese	9
2.1 Logitové modely pro nominální odezvu	9
2.2 Logitové modely pro nominální odezvu-příklad	11
2.3 Kumulativní logitové modely s ordinální odezvou	12
2.4 Kumulativní logitové modely s ordinální odezvou-příklad	13
2.5 Hledání vhodného modelu	14
2.6 Hledání vhodného modelu-příklad 1	15
2.7 Hledání vhodného modelu-příklad 2	16
2.8 Lineární regrese versus multinomická regrese s ordinální odezvou .	19
2.9 Výpočetní aspekty iterativního hledání odhadů parametrů β . . .	21
2.10 Grafické hledání odhadů parametrů β	22
2.11 Naivní hledání odhadů parametrů β	28
2.12 Generování dat pro multinomickou regresi s nominální a ordinální odezvou	31
3 Asymptotické testy	39
3.1 Test poměrem věrohodností pro logistickou regresi	39
3.2 Test poměrem věrohodností pro multinomickou regresi s nominální odezvou	39
3.3 Test poměrem věrohodností pro multinomickou regresi s ordinální odezvou	40
3.4 Waldův test pro logistickou regresi	41
3.5 Waldův test pro multinomickou regresi s nominální odezvou . . .	43
3.6 Waldův test pro multinomickou regresi s ordinální odezvou	45
3.7 Ekvivalence testů pro dvě kategorie	52
4 Kvalita vína	56
4.1 Model multinomické regrese s nominální odezvou pro bílé víno . .	56
4.2 Model multinomické regrese s ordinální odezvou pro bílé víno . . .	60
4.3 Srovnání modelů a jejich interpretace	62
4.4 Hledání vhodného modelu pro data o červeném víně	64
Závěr	69
Seznam použité literatury	70
Přílohy	74

Úvod

Cílem této diplomové práce je pojednat o regresních modelech, které jsou vhodné pro kategoriální odezvu. Jako kategoriální odezvu označujeme kvalitativní znak, u něhož nemůžeme zjistit měřitelné hodnoty, ale určujeme pouze rovnost či různost (tj. zda jedinec danou kvalitu splňuje nebo ne). Kvalitativním znakem může být pohlaví, stupeň dosaženého vzdělání, povolání, apod.

Práce se zaměří na logistickou regresi pro binární odezvu, tj. odezvu, která nabývá pouze dvou kategorií, a multinomickou logistickou regresi, která představuje její zobecnění pro odezvu s multinomickým rozdělením. V tomto případě budou představeny dva modely. První se používá pro multinomickou odezvu, jejíž kategorie nelze uspořádat, a druhý se používá v případě, že kategorii lze uspořádat.

Model logistické regrese je v praxi široce využíván. Původně se aplikoval pro účely biomedicínských studií, ale za posledních 20 let se také používal ve výzkumu spojeném se sociálními vědami a marketingem [1]. V současné době je logistická regrese využívána v tzv. skóringových funkcích, což je model, který se v bankovníctví užívá k určení toho, s jakou pravděpodobností klient nesplatí úvěr. Dále se logistická regrese používá v klasifikační analýze v případě, že máme dvě skupiny pozorování, např. rizikový a důvěryhodný klient, a zajímá nás, do které skupiny patří nový klient.

Model multinomické logistické regrese se pak například aplikuje v případě, kdy chceme zjistit, jakým dopravním prostředkem se dopraví jedinec do zaměstnání.

První kapitola se bude zabývat logistickou regresí. Zaměří se na teorii maximální věrohodnosti a odhady parametrů. Dále bude vysvětlen pojem dummy proměnných a uvedenou teorii aplikuje na příkladě.

V druhé kapitole bude popsán model multinomické logistické regrese s nominální a ordinální odezvou. Oba modely budou ilustrovány na příkladech. Dále bude v této kapitole vysvětlena strategie hledání vhodného modelu. Na konkrétním příkladě bude srovnán přístup logistické regrese a multinomické regrese s ordinální odezvou. Zaměříme se také na výpočetní aspekty iterativního hledání odhadů parametrů. Pokusíme se nalézt graficky odhady parametrů a také jedním naivním přístupem. A na závěr uvedeme, jak lze v statistickém softwaru R vygenerovat data vhodná pro multinomickou regresi s nominální a s ordinální odezvou.

Třetí kapitola se zaměří na odvození testu poměrem věrohodností a Waldova testu pro všechny tři v práci popsané modely a tyto testy pro konkrétní příklady spočítáme v R. Dále si dokážeme, že v případě binární odezvy jsou příslušné testy pro všechny tři modely ekvivalentní.

Ve čtvrté kapitole budeme hledat vhodný model pro data týkající se kvality červeného a bílého vína. Na data o bílém víně aplikujeme oba multinomické regresní modely a zvolíme ten, který se pro danou situaci lépe hodí a tento model použijeme pro data o červeném víně, následně srovnáme, zda jsou v obou případech významné stejné vysvětlující proměnné.

1. Logistická regrese

Úkolem této kapitoly je představit logistickou regresi, k čemuž je nezbytné popsat teorii maximální věrohodnosti.

Klíčovým faktorem, kterým se logistická regrese liší od lineární regrese, je, že vysvětlovaná proměnná je binární, značíme Y_i . U konkrétních dat pak označujeme vysvětlující proměnnou y_1, \dots, y_n . Vysvětlující proměnné označíme \mathbf{X}_i , pro konkrétní data pak $\mathbf{x}_1, \dots, \mathbf{x}_n$, které považujeme za konstanty. Tyto vysvětlující proměnné mohou být jak spojité, tak kategoriální. Nechť $\pi(\mathbf{x}_i) = P(Y_i = 1, \mathbf{X}_i = \mathbf{x}_i)$. Odezva Y_i má potom alternativní rozdělení s parametrem $\pi(\mathbf{x}_i)$, tj. $Y_i \sim \text{Alt}(\pi(\mathbf{x}_i))$.

Model logistické regrese je

$$\pi(\mathbf{x}_i) = \frac{e^{\beta^T \mathbf{x}_i}}{1 + e^{\beta^T \mathbf{x}_i}},$$

kde β je vektor p parametrů.

Kromě již citovaného Agrestiho knihy se problematikou logistické regrese zabývá také Pekár a Brabec [10]. Tato kniha ovšem obsahuje jen malé množství teorie a zaměřuje se převážně na příklady a základy práce se statistickým softwarem R.

1.1 Teorie maximální věrohodnosti

V této podkapitole se budeme věnovat teorii maximální věrohodnosti, neboť pomocí této teorie se v modelu logistické regrese odhadují parametry. Teorie maximální věrohodnosti je součástí přednášek na Matematicko-fyzikální fakultě.

Uveďme několik definic a vět.

Definice 1. *Věrohodnost* definujeme jako

$$l(\boldsymbol{\theta}, \mathbf{Y}) = \prod_{i=1}^n f_i(\boldsymbol{\theta}, \mathbf{Y}),$$

kde $f_i(\boldsymbol{\theta}, \mathbf{Y})$ je hustota i -tého pozorování.

Definice 2. *Logaritmická věrohodnost* je definována vztahem

$$L(\boldsymbol{\theta}, \mathbf{Y}) = \log \{l(\boldsymbol{\theta}, \mathbf{Y})\} = \sum_{i=1}^n \log f_i(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{i=1}^n L_i(\boldsymbol{\theta}, \mathbf{Y}).$$

Definice 3. Jako *skórový vektor* se označuje

$$U(\boldsymbol{\theta}, \mathbf{Y}) = \frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{i=1}^n \frac{\partial L_i}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{i=1}^n U_i(\boldsymbol{\theta}, \mathbf{Y}).$$

Definice 4. *Věrohodnostními rovnicemi* označujeme vztah

$$U(\boldsymbol{\theta}, \mathbf{Y}) = \mathbf{0}. \tag{1.1}$$

Definice 5. Matici

$$I(\boldsymbol{\theta}, \mathbf{Y}) = -\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{i=1}^n I_i(\boldsymbol{\theta}, Y_i) = \sum_{i=1}^n -\frac{\partial^2 L_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}, Y_i)$$

označujeme jako *pozorovanou informační matici*.

Definice 6. *Pozorovaná matice vyčíslená v $\hat{\boldsymbol{\theta}}_n$ je*

$$\hat{I}(\mathbf{Y}) = I(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}).$$

Definice 7. *Fisherovu (očekávanou) informační matici definujeme jako*

$$J(\boldsymbol{\theta}) = \mathbb{E} \{U(\boldsymbol{\theta}, \mathbf{Y})U^T(\boldsymbol{\theta}, \mathbf{Y})\}.$$

Věta 8. *Za slabých podmínek regularity platí vztah*

$$J(\boldsymbol{\theta}) = EI(\boldsymbol{\theta}, \mathbf{Y}).$$

Závislost vybraných definic na velikosti vzorku zdůrazníme indexem a budeme psát $L^{(n)}(\boldsymbol{\theta}, \mathbf{Y}), U^{(n)}(\boldsymbol{\theta}, \mathbf{Y}), I^{(n)}(\boldsymbol{\theta}, \mathbf{Y})$ a $J^{(n)}(\boldsymbol{\theta})$.

Věta 9. *Při splnění podmínek regularity a pokud platí, že $\hat{\boldsymbol{\theta}}_n : n = 1, 2, \dots$ je taková posloupnost řešení věrohodnostních rovnic $U^{(n)}(\boldsymbol{\theta}, \mathbf{Y}) = \mathbf{0}, n = 1, 2, \dots$ taková, že*

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}^0 \text{ pro } n \rightarrow \infty,$$

kde $\boldsymbol{\theta}^0 \in \Theta$ je skutečná hodnota parametru ležící ve vnitřku Θ , platí

$$\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)^T F^{(n)}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) \xrightarrow{D} \chi_p^2 \text{ pro } n \rightarrow \infty, \quad (1.2)$$

kde $F^{(n)}$ může být jakákoliv z matic $J^{(n)}(\boldsymbol{\theta}^0), J^{(n)}(\hat{\boldsymbol{\theta}}_n), I^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}), I^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y})$.

Z tohoto vztahu je pak odvozen Waldův test, který testuje nulovou hypotézu $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ pro zvolené $\boldsymbol{\theta}^0 \in \Theta$.

Věta 10. *Za podmínek věty 9 také platí*

$$-2 \left\{ L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) - L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) \right\} \xrightarrow{D} \chi_p^2 \text{ pro } n \rightarrow \infty. \quad (1.3)$$

Z tohoto vztahu se pak odvozuje test poměrem věrohodností, který testuje nulovou hypotézu $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ pro zvolené $\boldsymbol{\theta}^0 \in \Theta$. Zdůrazněme, že v obou případech se jedná o asymptotické testy, které jsou za platnosti H_0 asymptoticky ekvivalentní.

1.2 Odhad parametrů v modelu logistické regrese

Cílem této podkapitoly bude aplikovat teorii maximální věrohodnosti, která byla uvedena v podkapitole 1.1, na model logistické regrese.

Věta 11. Pro model logistické regrese jsou věrohodnostní rovnice obecně vyjádřené vztahem (1.1) následující

$$\sum_{i=1}^n x_i^{(j)} (y_i - \pi(\mathbf{x}_i)) = 0, j = 1, \dots, p. \quad (1.4)$$

Důkaz. Když $y_i = 1$, pak pravděpodobnost je $\pi(\mathbf{x}_i)$. Pokud je $y_i = 0$, pak pro pravděpodobnost platí $1 - \pi(\mathbf{x}_i)$.

Pro i -té pozorování je hustota následující

$$\pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}.$$

Z těchto hustot pak dostaneme věrohodnost, která je

$$l(\boldsymbol{\beta}) = \prod_{i=1}^n \pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}.$$

Logaritmická věrohodnost potom přirozeně bude

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n \{y_i \ln [\pi(\mathbf{x}_i)] + (1 - y_i) \ln [1 - \pi(\mathbf{x}_i)]\}. \quad (1.5)$$

Abychom našli maximálně věrohodný odhad $\boldsymbol{\beta}$, derivujeme logaritmickou věrohodnost podle β_i a výsledek položíme rovný 0.

Jestliže $\mathbf{x}_i = (x_i^{(1)}, \dots, x_i^{(p)})$, pak dostaneme věrohodnostní rovnice ve tvaru

$$\sum_{i=1}^n x_i^{(j)} (y_i - \pi(\mathbf{x}_i)) = 0, j = 1, \dots, p.$$

[2]

□

V lineární regresi jsou věrohodnostní rovnice lineární vzhledem k neznámým parametrům, a tudíž jsou snadno řešitelné. Rovnice (1.4) jsou naproti tomu nelineární, a proto vyžadují speciální metody k jejich řešení. Tato metoda je iterativní a nazývá se Newton-Raphsonův algoritmus.

1. Zvolíme $\boldsymbol{\beta}_0$

2. $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbf{H}^{(t)})^{-1} \cdot \mathbf{u}^{(t)}$, kde $\mathbf{H}^{(t)} = (h_{ab}^{(t)})$

$$u_j^{(t)} = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \Big|_{\boldsymbol{\beta}^{(t)}} = \sum_{i=1}^n (y_i - \pi_i^{(t)}) x_i^{(j)}$$

$$h_{ab}^{(t)} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b} \Big|_{\boldsymbol{\beta}^{(t)}} = - \sum_{i=1}^n x_i^{(a)} x_i^{(b)} \pi_i^{(t)} (1 - \pi_i^{(t)}),$$

kde

$$\pi_i^{(t)} = \frac{\exp(\sum_{j=1}^p \beta_j^{(t)} x_i^{(j)})}{1 + \exp(\sum_{j=1}^p \beta_j^{(t)} x_i^{(j)})}.$$

[1]

1.3 Dummy proměnné

Tato kapitola se zabývá tím, jak vypadají vstupní data v případě, že je alespoň jeden z regresorů kategoriální.

Klíčový vztah je tvaru

$$\frac{e^{\beta^T \mathbf{x}}}{1 + e^{\beta^T \mathbf{x}}}, \quad (1.6)$$

kde \mathbf{x} je vektor nezávisle proměnných. Tento vektor si lze jednoduše představit, pokud jsou všechny nezávisle proměnné spojité, pak tento vektor vypadá například následovně (1, 4, 8, 6, 7). První je jednička představující konstantní člen a následují hodnoty spojitých proměnných. Ale jak vypadá vektor \mathbf{x} , když je jedna proměnná kategoriální o k kategoriích? Nabízelo by se vytvořit systém k pomocných proměnných, které budou náležet k kategoriím. Daná pomocná proměnná bude nabývat hodnoty 1, pokud bude platit příslušná kategorie, a 0 v opačném případě. Představme si nyní matici \mathbf{X} o n řádcích, kde n je počet pozorování. Je zřejmé, že pokud je mezi nezávisle proměnnými kategoriální proměnná, tak tato matice nemá nezávislé sloupce, protože sečteme-li sloupce příslušící kategoriální proměnné, dostaneme sloupec samých jedniček. Tento sloupec bude pak totožný s prvním sloupcem matice \mathbf{X} , který představuje konstantní člen. Tato situace má však řešení. Místo k pomocných proměnných, tzv. dummy proměnných, budeme uvažovat jen $k - 1$ těchto proměnných, tj. prakticky jeden sloupec v matici \mathbf{X} vyškrtáme. Důležité je si všimnout, že tímto postupem neztratíme, žádnou informaci, protože pozorování, které nabývalo hodnoty vyřazené kategorie, poznáme tak, že všechny jeho dummy proměnné nabývají hodnoty 0.

Jak je uvedeno v [2] příslušná mocnina v (1.6) bude vypadat následovně

$$\beta^T \mathbf{x} = \beta_0 + \beta_1 x_1 + \dots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p,$$

kde D_{jl} jsou dummy proměnné a β_{jl} jsou k nim příslušné koeficienty.

1.4 Logistická regrese-příklad

Cílem této podkapitoly bude na jednoduchém příkladě aplikovat teorii k modelu logistické regrese uvedenou v podkapitolách 1.1 a 1.2. Úkolem bude tedy odhadnout parametry pomocí teorie maximální věrohodnosti a odhadnuté parametry následně otestovat pomocí testu poměrem věrohodností. Popíšme si nyní použitá data, která byla převzata z knihy [3].

Let	Teplota	Problém	Let	Teplota	Problém
1	66	0	13	67	0
2	70	1	14	53	1
3	69	0	15	67	0
4	68	0	16	75	0
5	67	0	17	70	0
6	72	0	18	81	0
7	73	0	19	76	0
8	70	0	20	79	0
9	57	1	21	75	1
10	63	1	22	76	0
11	70	1	23	58	1
12	78	0			

Tabulka 1.1: Data o problému s O-ringem v závislosti na teplotě

Pro 23 letů vesmírného raketoplánu výše uvedená tabulka 1.1 ukazuje teplotu (°F) v době letu a zda alespoň jedna primární součástka anglicky nazývaná O-ring měla problém s teplotou. Jak již bylo uvedeno, příklad budeme modelovat pomocí logistické regrese, kde vysvětlující proměnná bude teplota při letu. Teplotu budeme považovat za spojitou náhodnou veličinu, protože nabývá velkého množství různých hodnot. Rovnice příslušící k tomuto modelu je

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_i,$$

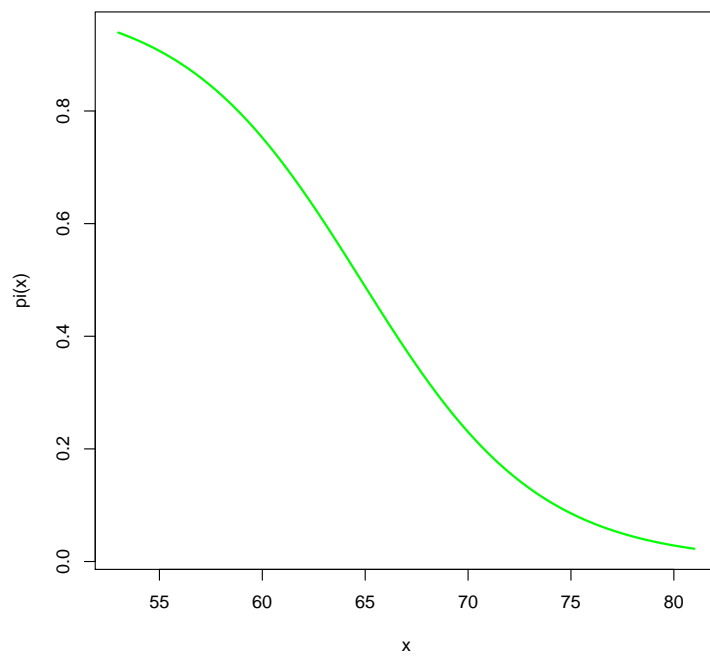
kde π_i je pravděpodobnost zhadu pro i -té pozorování. Data budeme zpracovávat v softwarovém balíku R pomocí funkce `glm`. Příslušný zdrojový kód je uveden v příloze. Spočítané odhady a další charakteristiky jsou uvedeny v tabulce 1.2, z nichž směrodatná odchylka l -tého parametru je dána vztahem $\sigma = \sqrt{\text{var}\hat{\beta}_l}$

Parametr	Odhad	Směr. odchylka	p-hodnota
β_0	15,043	7,3786	
β_1	-0,232	0.1082	0.004804

Tabulka 1.2: Odhady parametrů a další charakteristiky pro model logistické regrese pro data týkající se problému s O-ringem v závislosti na teplotě

Záporná hodnota koeficientu $\hat{\beta}_1$ indikuje, že nízké hodnoty teploty při letu zvyšují pravděpodobnost problému s teplotou. Nyní je třeba provést test významnosti regresního parametru β_1 , který odpovídá teplotě při letu. Budeme testovat nulovou hypotézu $H_0 : \beta_1 = 0$ proti alternativě $H_1 : \beta_1 \neq 0$. Tento test vychází významně s p -hodnotou 0,005. Můžeme tedy říci, že na hladině 5% se podařilo prokázat vliv teploty při letu na existenci problému s teplotou O-ringem.

Na závěr ještě uvedeme grafické zobrazení závislosti pravděpodobnosti π na vysvětlující proměnné x , které představuje obrázek 1.1.



Obrázek 1.1: Hodnota pravděpodobnosti $\pi(x)$ v závislosti na teplotě při letu

2. Multinomická regrese

Logistická regrese se používá k modelování binární odezvy, tj. odezvy, která nabývá hodnoty 0 nebo 1. Tato kapitola se bude zabývat zobecněním tohoto modelu, neboť multinomická regrese se používá pro odezvy s více jak dvěma kategoriemi. Tyto kategorie mohou být dvojího typu. Prvním typem jsou nominální vysvětlované proměnné. To nastává v případě, že kategorie nelze nijak uspořádat. Druhým typem jsou ordinální vysvětlované proměnné. O tomto typu proměnných mluvíme v případě, že kategorie lze uspořádat.

2.1 Logitové modely pro nominální odezvu

Nechť Z je vysvětlovaná proměnná, J je počet kategorií, které tato proměnná nabývá, a $\{\pi_1, \dots, \pi_J\}$ označují pravděpodobnosti jednotlivých kategorií. Platí $\sum_{j=1}^J \pi_j = 1$. Je zřejmé, že Z má multinomické rozdělení a nabývá hodnot $1, 2, \dots, J$, píšeme $Z \sim \text{Multinom}(n, \pi_1, \dots, \pi_J)$, kde n je počet pozorování. Tuto náhodnou veličinu budeme reprezentovat pomocí $\mathbf{Y} = \{y_1, y_2, \dots, y_J\}$, kde y_j nabývá hodnoty 1, když $Z = j$, a hodnoty 0 jinak. Lze si všimnout, že nezáleží na pořadí kategorií, protože odezva je nominální, tedy kategorie jsou neuspořádané.

Logitový model pro nominální odezvu sestavíme pomocí logaritmu poměru pravděpodobností. J -tá kategorie se považuje za referenční. Uvažované rovnice jsou následující

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j^T \mathbf{x}, \quad j = 1, \dots, J-1. \quad (2.1)$$

Těchto rovnic je $J-1$, kde \mathbf{x} je vektor p regresorů s tím, že těchto vektorů je n , a kde α_j a β_j jsou parametry, které vidíme, že jsou odlišné pro každou rovnici. [3]

Věta 12. *Pravděpodobnosti jednotlivých kategorií lze vyjádřit ve tvaru*

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \beta_j^T \mathbf{x})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \beta_h^T \mathbf{x})} \quad (2.2)$$

Důkaz. Rovnice (2.1) přepíšeme na tvar

$$\pi_j = \pi_J \exp(\alpha_j + \beta_j^T \mathbf{x}), \quad j = 1, \dots, J-1.$$

Nyní sečteme všechny pravděpodobnosti

$$1 = \pi_1 + \dots + \pi_J = \pi_J \exp(\alpha_1 + \beta_1^T \mathbf{x}) + \dots + \pi_J \exp(\alpha_{j-1} + \beta_{j-1}^T \mathbf{x}) + \pi_J.$$

Vyjádříme π_J

$$\pi_J = \frac{1}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \beta_h^T \mathbf{x})}. \quad (2.3)$$

Nyní už jen tento vztah dosadíme do π_j a získáme výsledek

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \beta_j^T \mathbf{x})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \beta_h^T \mathbf{x})}.$$

□

Nyní se zabýváme odhadem parametrů v logitovém modelu s nominální odezvou. Při odhadu parametrů se používá stejně jako v logistické regresi metoda maximální věrohodnosti. Odhad se provádí tak, aby bylo současně splněno všech $J - 1$ rovnic, které určují model. Pro $i = 1, \dots, n$, nechť $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$ reprezentuje multinomickou závisle proměnnou. Platí $y_{ij} = 1$, když odezva nabývá kategorie j a $y_{ij} = 0$ jinak. Tudíž platí $\sum_{j=1}^J y_{ij} = 1$. Nechť dále $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ označuje i -tou vysvětlující proměnnou. A nechť $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jp})^T$ značí parametry pro j -tou rovnici.

Věta 13. *Logaritmická věrohodnostní funkce pro multinomickou regresi s nominální odezvou je rovna výrazu*

$$\sum_{j=1}^{J-1} \left[\alpha_j \left(\sum_{i=1}^n y_{ij} \right) + \sum_{k=1}^p \beta_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right].$$

Důkaz. Protože platí $\pi_J = 1 - (\pi_1 + \dots + \pi_{J-1})$ a $y_{iJ} = 1 - (y_{i1} + \dots + y_{i,J-1})$, příspěvek do logaritmické věrohodnostní funkce pro pozorování i je

$$\begin{aligned} \log \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{j=1}^{J-1} y_{ij} \log \pi_j(\mathbf{x}_i) + \left(1 - \sum_{j=1}^{J-1} y_{ij} \right) \log \left[1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i) \right] = \\ &= \sum_{j=1}^{J-1} y_{ij} \log \frac{\pi_j(\mathbf{x}_i)}{1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i)} + \log \left[1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i) \right] \end{aligned}$$

Nyní předpokládejme n nezávislých pozorování. V prvním členu provedeme substituci logitu za $\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i$ a ve druhém členu $\pi_J(\mathbf{x}_i) = 1 / \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right]$. Odtud dostáváme logaritmickou věrohodnost

$$\begin{aligned} \log \prod_{i=1}^n \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{i=1}^n \left\{ \sum_{j=1}^{J-1} y_{ij} (\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) - \right. \\ &\quad \left. - \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right] \right\} = \\ &= \sum_{j=1}^{J-1} \left[\alpha_j \left(\sum_{i=1}^n y_{ij} \right) + \sum_{k=1}^p \beta_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] - \\ &\quad - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right]. \end{aligned} \tag{2.4}$$

Odhad se dále získá pomocí Newton-Raphsonovy metody.[1] □

2.2 Logitové modely pro nominální odezvu-příklad

Pohlaví	Rasa	Demokrat	Republikán	Nezávislý
Muž	Bílý	132	176	127
	Černý	42	6	12
Žena	Bílá	172	129	130
	Černá	56	4	15

Tabulka 2.1: Data týkající se příslušnosti k politické straně v závislosti na pohlaví a rase

Cílem této podkapitoly je teorii o logitovém modelu pro nominální odezvu představenou v 2.1 aplikovat na výše uvedená data z tabulky 2.1, která byla převzata z knihy [1].

Příklad se zabývá vztahem mezi příslušností k politické straně, která představuje multinomickou odezvu, a pohlavím a rasou, které reprezentují regresory, proto má smysl použít teorii z předchozí kapitoly. Příslušné rovnice pro logitový model s nominální odezvou jsou následující

$$\log \left(\frac{P(\text{strana}=\text{demokrat})}{P(\text{strana}=\text{nezavisly})} \right) = \alpha_1 + \beta_{11}(\text{pohlavi}=\text{zena}) + \beta_{12}(\text{rasa}=\text{cerny}), \quad (2.5)$$

$$\log \left(\frac{P(\text{strana}=\text{republikan})}{P(\text{strana}=\text{nezavisly})} \right) = \alpha_2 + \beta_{21}(\text{pohlavi}=\text{zena}) + \beta_{22}(\text{rasa}=\text{cerny}). \quad (2.6)$$

Pro tento příklad byl použit statistický software R, konkrétně balík `net` a funkce `multinom`. Zdrojový kód je přiložen v příloze. Výsledné odhady a další charakteristiky jsou uvedeny v tabulce 2.2:

Parametr	Odhad	e ^{odhad}	Směr. odchylka
α_1	0,050	1,051	0,120
α_2	0,335	1,398	0,115
β_{11}	0,220	1,246	0,158
β_{12}	1,118	3,060	0,234
β_{21}	-0,353	0,703	0,165
β_{22}	-1,160	0,314	0,380

Tabulka 2.2: Odhady parametrů a další charakteristiky pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na pohlaví a rase

Zamysleme se nyní nad interpretací odhadů vzešlých s multinomického modelu s nominální odezvou. Fakt, že odhad parametru β_{11} vyšel větší než 0, nám říká, že ženské pohlaví zvyšuje pravděpodobnost, že člověk bude demokrat oproti pravděpodobnosti, že bude nezávislý. To stejné vzhledem ke kladnosti koeficientu β_{12}

platí pro člověka s černou pletí. Záporný koeficient β_{21} značí, že fakt, že je člověk žena, snižuje pravděpodobnost, že bude republikán, oproti pravděpodobnosti, že bude nezávislý. Stejně tak tuto pravděpodobnost snižuje, když je člověk černé pleti, vzhledem k zápornosti koeficientu β_{22} .

2.3 Kumulativní logitové modely s ordinální odezvou

V této podkapitole představíme kumulativní logitový model s ordinální odezvou. Základní myšlenkou tohoto modelu je, že logity využijí uspořádanost kategorií závisle proměnné. Zavedme kumulativní pravděpodobnost, tj. pravděpodobnost, že vysvětlovaná proměnná Z padne pod určitý bod. Pro j -tou kategorii je kumulativní pravděpodobnost

$$P(Z \leq j) = \pi_1 + \dots + \pi_j, \quad j = 1, \dots, J,$$

kde $\{\pi_1, \dots, \pi_J\}$ označují pravděpodobnosti jednotlivých kategorií.

Platí $P(Z \leq 1) \leq P(Z \leq 2) \leq \dots \leq P(Z \leq J) = 1$. Modely však nevyužívají poslední kumulovanou pravděpodobnost z toho důvodu, že je nutně rovna 1.

Pro logit kumulované pravděpodobnosti platí

$$\begin{aligned} \text{logit}[P(Z \leq j)] &= \log \left[\frac{P(Z \leq j)}{1 - P(Z \leq j)} \right] = \\ &= \log \left[\frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J} \right], \quad j = 1, \dots, J - 1. \end{aligned}$$

Model pro kumulativní logit vypadá jako model logistické regrese, kde kategorie 1 až j tvoří jednu novou kategorii a kategorie $j + 1$ až J tvoří druhou novou kategorii. Rovnice pro tento typ modelu vypadají následovně

$$\text{logit}[P(Z \leq j)] = \alpha_j + \boldsymbol{\beta}^T \mathbf{x}, \quad j = 1, \dots, J - 1. \quad (2.7)$$

Jedná se o soustavu $J - 1$ podmínek pro J pravděpodobností, kde se vyskytuje vektor p parametrů $\boldsymbol{\beta}$ a rovněž vektor p vysvětlujících proměnných \mathbf{x} . Všimněme si, že parametry $\boldsymbol{\beta}$ nemají index j , tudíž jsou tyto parametry společné pro všech $J - 1$ rovnic. [3]

Odhad parametrů tohoto modelu se opět provede pomocí metody maximální věrohodnosti. Vysvětlovanou proměnnou Z však budeme stejně jako v předchozím případě reprezentovat pomocí $\mathbf{Y} = \{y_1, y_2, \dots, y_J\}$, kde y_j nabývá hodnoty 1 v případě, že $Z = j$, a hodnoty 0 jinak.

Věrohodnost kumulativního modelu je

$$\begin{aligned} \prod_{i=1}^n \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \prod_{i=1}^n \left[\prod_{j=1}^J (P(Z \leq j | \mathbf{x}_i) - P(Z \leq j - 1 | \mathbf{x}_i))^{y_{ij}} \right] = \\ &= \prod_{i=1}^n \left\{ \left(\frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right)^{y_{i1}} \left[\prod_{j=2}^{J-1} \left(\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \right. \right. \right. \\ &\quad \left. \left. \left. - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right)^{y_{ij}} \right] \left(1 - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right)^{y_{iJ}} \right\}. \quad (2.8) \end{aligned}$$

[1]

Ze vzorce (2.8) lze pak odvodit logaritmickou věrohodnost

$$\sum_{i=1}^n \left\{ y_{i1} \log \left(\frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) + \sum_{j=2}^{J-1} y_{ij} \log \left(\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) + y_{iJ} \log \left(1 - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \right\}. \quad (2.9)$$

2.4 Kumulativní logitové modely s ordinální odezvou-příklad

			Vážnost zranění				
Pohlaví	Místo	Bezpečnostní pás	1	2	3	4	5
Žena	Město	Ne	7287	175	720	91	10
		Ano	11587	126	577	48	8
	Vesnice	Ne	3246	73	710	159	31
		Ano	6134	94	564	82	17
Muž	Město	Ne	10381	136	566	96	14
		Ano	10969	83	259	37	1
	Vesnice	Ne	6123	141	710	188	45
		Ano	6693	74	353	74	12

Tabulka 2.3: Data týkající se závažnosti zranění při automobilových nehodách v závislosti na pohlaví, místě nehody a použití bezpečnostního pásu

Zde uvedeme příklad, který se zabývá závažností zranění při automobilových nehodách. Jako model budeme uvažovat kumulativní logitový model s ordinální odezvou, neboť závažnost zranění lze seřadit od nejlehčích zranění po nejtěžší. Jako vysvětlující proměnné zde budou figurovat pohlaví, místo nehody a užití bezpečnostního pásu. Data pro tento příklad byla převzata z knihy [3] a jsou uvedeny v tabulce 2.3.

Príslušné rovnice pro tento model vychází ze vztahu (2.7) a konkrétně si popíšeme jednu z nich

$$\log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5} \right) = \alpha_1 + \beta_1(\text{pohlavi}=\text{muz}) + \beta_2(\text{misto}=\text{vesnice}) + \beta_3(\text{bezpecnostni pas}=\text{ano}), \quad (2.10)$$

kde π_1, \dots, π_5 jsou parametry modelu z kapitoly 2.3.

Tento příklad byl řešen opět pomocí statistického softwaru R, v tomto případě pomocí balíku VGAM, z něhož byla využita funkce `vglm`. Příslušný zdrojový kód je opět uveden v příloze.

Výsledné odhady koeficientů a další charakteristiky jsou zaznamenány v tabulce 2.4.

Parametr	Odhad	e ^{odhad}	Směr. odchylka
α_1	1,977	7,218	0,025
α_2	2,151	8,593	0,026
α_3	4,018	55,588	0,040
α_4	5,925	374,192	0,088
β_1	0,545	1,724	0,027
β_2	-0,773	0,462	0,027
β_3	0,824	2,280	0,028

Tabulka 2.4: Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na pohlaví, místu nehody a použití bezpečnostního pásu

Zamysleme se nyní nad tím, co jednotlivé odhady koeficientů říkají. Abychom mohli lépe posoudit, co od jednotlivých koeficientů čekat, vykresleme si pomocnou tabulku 2.5, ve které bude uveden procentuální podíl závažnosti zranění pro jednotlivé řádky výše uvedené tabulky 2.3.

			Vážnost zranění				
Pohlaví	Místo	Bezp. pás	1	2	3	4	5
Žena	Město	Ne	87,98%	2,11%	8,69%	1,10%	0,12%
		Ano	93,85%	1,02%	4,67%	0,39%	0,06%
	Vesnice	Ne	76,94%	1,73%	16,83%	3,77%	0,73%
		Ano	89,01%	1,36%	8,18%	1,19%	0,25%
Muž	Město	Ne	92,75%	1,22%	5,06%	0,86%	0,13%
		Ano	96,65%	0,73%	2,28%	0,33%	0,01%
	Vesnice	Ne	84,96%	1,96%	9,85%	2,61%	0,62%
		Ano	92,88%	1,03%	4,90%	1,03%	0,17%

Tabulka 2.5: Procentuální podíl závažnosti zranění pro data týkající se zranění při automobilových nehodách v závislosti na pohlaví, místu nehody a použití bezpečnostního pásu

Nyní vidíme, že přítomnost bezpečnostního pásu jednoznačně zvyšuje pravděpodobnost lehkého zranění, což potvrzuje kladný koeficient β_3 . Čísla dále ukazují, že pokud se zranění stalo na vesnici, je menší pravděpodobnost, že bylo lehké. Toto potvrzuje záporný koeficient β_2 . Když nyní porovnáme první čtyři řádky prvního sloupce s druhými čtyřmi řádky prvního sloupce dané tabulky, vidíme, že pokud byl zraněným muž, pak je větší pravděpodobnost, že zranění bylo lehké. A zvyšování pravděpodobnosti lehkého zranění pro mužské pohlaví dokládá i kladný koeficient β_1 .

2.5 Hledání vhodného modelu

Cílem této podkapitoly je popsat strategii hledání vhodného modelu. Při hledání vhodného modelu zpravidla začínáme u nejširšího možného modelu, kterým je

saturovaný model, tj. model, který má tolik parametrů, jako je počet pozorování. Tento model označíme M_0 . Dále zvolíme podmodel modelu M_0 , který označíme M_1 . Poté stanovíme podmodel modelu M_1 označený jako M_2 , atd.

Nyní budeme testovat hypotézy. K tomu využijeme test poměrem věrohodností, který byl popsán v kapitole 1.1. Testovou statistiku tohoto testu nazýváme deviance. V kontextu hledání vhodného modelu lze pro testovou statistiku G psát

$$G = -2[\ln(\text{věrohodnost podmodelu}) - \ln(\text{věrohodnost širšího modelu})].$$

Lze říci, že deviance porovnává maximální hodnotu věrohodnostní funkce za předpokladu, že platí uvažovaný model, s její maximální možnou hodnotou, která je dosažitelná za platnosti saturovaného modelu.

Deviance má asymptoticky χ^2 rozdělení s tolika stupni volnosti, kolik je rozdíl v počtu stupňů volnosti jednotlivých modelů. Platí, že to se zároveň rovná rozdílu v počtu parametrů obou modelů. Pokud tedy testujeme nulovou hypotézu o nulovosti jednoho parametru, tak má testová statistika rozdělení χ_1^2 .

Pokud testujeme model M_1 proti podmodelu M_0 a výsledkem je nezamítnutí modelu M_1 , můžeme proti zvyklostem matematické statistiky říci, že přijímáme model M_1 . Dále je otázkou, zda není lepší model M_2 než M_1 . K tomu můžeme provést test s nulovou hypotézou, že platí model M_2 , proti alternativě, že platí M_1 . Statistický software R příslušnou testovou statistiku nevypočítá přímo, ale my můžeme využít známého rozkladu testové statistiky G , která je uvedena například v [4]

$$G_{2,1} = G_2 - G_1,$$

tedy testová statistika při testování modelu M_2 proti M_1 je rovna rozdílu testové statistiky při testování modelu M_2 proti saturovanému modelu a testové statistiky při testování modelu M_1 proti saturovanému modelu. Připomeňme, že obecně modelem M_0 nemusí být saturovaný model, ale statistický balík R právě se saturovaným modelem počítá.

Testová statistika $G_{2,1}$ má asymptoticky χ^2 rozdělení za platnosti M_2 , kde počet stupňů volnosti odpovídá rozdílu v počtu parametrů mezi M_2 a M_1 .

2.6 Hledání vhodného modelu-příklad 1

V této podkapitole rozšíříme příklad z kapitoly 2.2, tj. příklad zabývající se vztahem mezi příslušností k politické straně na jedné straně a pohlavím a rasou na straně druhé, neboť již nebudeme považovat rovnice příslušící logitovému modelu za dané, ale budeme hledat vhodný model. Jako výchozí model použijeme model $Y \sim P * R$, který obsahuje interakci mezi P a R a interakce nižšího řádu. Tomuto modelu vyhovují následující rovnice:

$$\begin{aligned} \log \left(\frac{P(\text{strana}=\text{demokrat})}{P(\text{strana}=\text{nezavisly})} \right) &= \alpha_1 + \beta_{11}(\text{pohlavi}=\text{zena}) + \beta_{12}(\text{rasa}=\text{cerny}) + \\ &\quad + \beta_{13}(\text{pohlavi}=\text{zena}) \cdot (\text{rasa}=\text{cerny}), \\ \log \left(\frac{P(\text{strana}=\text{republikan})}{P(\text{strana}=\text{nezavisly})} \right) &= \alpha_2 + \beta_{21}(\text{pohlavi}=\text{zena}) + \beta_{22}(\text{rasa}=\text{cerny}) + \end{aligned}$$

$$+\beta_{23}(\text{pohlavi=zena}) \cdot (\text{rasa=cerny}).$$

Odhady příslušných koeficientů vycházejí $\hat{\alpha}_1 = 0,039$, $\hat{\alpha}_2 = 0,326$, $\hat{\beta}_{11} = 0,241$, $\hat{\beta}_{12} = 1,214$, $\hat{\beta}_{13} = -0,177$, $\hat{\beta}_{21} = -0,334$ a $\hat{\beta}_{22} = -1,019$, $\hat{\beta}_{23} = -0,294$.

Nyní budeme uvažovat podmodel výše uvedeného modelu. Tímto modelem bude model $Y \sim P + R$. Tento model byl již uvažován v kapitole 2.2.

Nyní testujme hypotézu $H_0 : Y \sim P + R$ proti alternativě $H_1 : Y \sim P * R$. Testovou statistiku příslušící tomuto testu nám R přímo nespočítá, ale jak bylo výše uvedeno, rovná se rozdílu deviance podmodelu a deviance širšího modelu. Nyní zbývá určit počet stupňů volnosti, ten se rovná rozdílu v počtu parametrů obou modelů. V našem případě je tedy počet stupňů volnosti 2. Výsledná p -hodnota je 0.906. To znamená, že nezamítáme model $Y \sim P + R$. Tedy můžeme říct, že tento model přijímáme.

Pokusíme se opět o zjednodušení modelu. Tentokrát v něm bude vystupovat jediná vysvětlující proměnná. V tomto případě máme ale dvě možnosti, jak sestavit model. Prvním modelem je $Y \sim P$. Tento model vystihují následující rovnice

$$\log \left(\frac{P(\text{strana=demokrat})}{P(\text{strana=nezavisly})} \right) = \alpha_1 + \beta_{11}(\text{pohlavi=zena}),$$

$$\log \left(\frac{P(\text{strana=republikan})}{P(\text{strana=nezavisly})} \right) = \alpha_2 + \beta_{21}(\text{pohlavi=zena}).$$

Odhady koeficientů vycházejí $\hat{\alpha}_1 = 0,225$, $\hat{\alpha}_2 = 0,270$, $\hat{\beta}_{11} = 0,228$, $\hat{\beta}_{21} = -0,356$. Deviance pro výše uvedený model je 2162,507.

Druhým modelem, který bychom mohli uvažovat je model $Y \sim R$. Tomuto modelu přísluší tyto rovnice

$$\log \left(\frac{P(\text{strana=demokrat})}{P(\text{strana=nezavisly})} \right) = \alpha_1 + \beta_{11}(\text{rasa=cerny}),$$

$$\log \left(\frac{P(\text{strana=republikan})}{P(\text{strana=nezavisly})} \right) = \alpha_2 + \beta_{21}(\text{rasa=cerny}).$$

Výsledné odhady koeficientů jsou $\hat{\alpha}_1 = 0,168$, $\hat{\alpha}_2 = 0,171$, $\hat{\beta}_{11} = 1,121$, $\hat{\beta}_{21} = -1,165$. Deviance tohoto modelu je 2099,224.

Nyní se dostáváme k otázce, který model vybrat jako podmodel modelu $Y \sim P + R$. Pravidlem je vybírat podmodel s nejmenší deviancí, aby se co nejméně lišil od saturovaného modelu. Tímto se vyhneme ztrátě některých klíčových proměnných. Aplikací tohoto pravidla tedy vybereme podmodel $Y \sim R$ a můžeme testovat hypotézu $H_0 : Y \sim R$ proti alternativě $H_1 : Y \sim P + R$. Počet stupňů volnosti je v tomto případě opět 2.

Výsledná p -hodnota je 0,001. To znamená, že zamítáme podmodel $Y \sim R$. Závěrem celého hledání vhodného podmodelu je přijetí $Y \sim P + R$ jako nejlepšího modelu.

2.7 Hledání vhodného modelu-příklad 2

V této podkapitole budeme hledat vhodný model pro příklad z podkapitoly 2.4. Tento příklad se zabývá závažností zranění při automobilových nehodách. Vý-

chozím modelem pro tento příklad bude model $Y \sim P * M * B$, který obsahuje interakci mezi P , M a B a všechny interakce nižšího řádu. K tomuto modelu náleží 4 rovnice. My zde uvedeme pouze první rovnici, zbytek rovnic lze analogicky odvodit od ní

$$\begin{aligned} \log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5} \right) &= \alpha_1 + \beta_1(\text{pohlavi=muz}) + \beta_2(\text{misto=vesnice}) + \\ &+ \beta_3(\text{bezpecnostni pas=ano}) + \beta_4(\text{pohlavi=muz})(\text{misto=vesnice}) + \\ &+ \beta_5(\text{pohlavi=muz})(\text{bezpecnostni pas=ano}) + \\ &+ \beta_6(\text{misto=vesnice})(\text{bezpecnostni pas=ano}) + \\ &+ \beta_7(\text{pohlavi=muz})(\text{misto=vesnice})(\text{bezpecnostni pas=ano}). \end{aligned}$$

Výsledné odhady koeficientů jsou $\hat{\alpha}_1 = 1,999$, $\hat{\alpha}_2 = 2,173$, $\hat{\alpha}_3 = 4,041$, $\hat{\alpha}_4 = 5,948$, $\hat{\beta}_1 = 0,551$, $\hat{\beta}_2 = -0,816$, $\hat{\beta}_3 = 0,731$, $\hat{\beta}_4 = -0,014$, $\hat{\beta}_5 = 0,085$, $\hat{\beta}_6 = 0,177$ a $\hat{\beta}_7 = -0,147$. Deviance pro tento model je 151,0362.

Podívejme se ještě, jaký model $Y \sim P * M * B$ obsahuje koeficienty v závislosti na hodnotách jednotlivých vysvětlujících proměnných. V níže uvedené tabulce 2.6 je přehledný souhrn těchto koeficientů.

Pohlaví	Místo	Bezp. pás	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Muž	Vesnice	Ano	+	+	+	+	+	+	+
Muž	Vesnice	Ne	+	+	-	+	-	-	-
Muž	Město	Ano	+	-	+	-	+	-	-
Muž	Město	Ne	+	-	-	-	-	-	-
Žena	Vesnice	Ano	-	+	+	-	-	+	-
Žena	Vesnice	Ne	-	+	-	-	-	-	-
Žena	Město	Ano	-	-	+	-	-	-	-
Žena	Město	Ne	-	-	-	-	-	-	-

Tabulka 2.6: Přehled koeficientů, které jsou zahrnuty v modelu $Y \sim P * M * B$ v závislosti na hodnotách jednotlivých vysvětlujících proměnných

Zvolme teď podmodel modelu $Y \sim P * M * B$. Tímto podmodelem bude model $Y \sim P * M + P * B + M * B$, který obsahuje interakce mezi P a M , P a B a mezi M a B a interakce nižšího řádu. Tento model reprezentují rovněž 4 rovnice, z nichž uvádíme první:

$$\begin{aligned} \log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5} \right) &= \alpha_1 + \beta_1(\text{pohlavi=muz}) + \beta_2(\text{misto=vesnice}) + \\ &+ \beta_3(\text{bezpecnostni pas=ano}) + \beta_4(\text{pohlavi=muz})(\text{misto=vesnice}) + \\ &+ \beta_5(\text{pohlavi=muz})(\text{bezpecnostni pas=ano}) + \\ &+ \beta_6(\text{misto=vesnice})(\text{bezpecnostni pas=ano}). \end{aligned}$$

Výsledné koeficienty pro tento model vychází $\hat{\alpha}_1 = 1,986$, $\hat{\alpha}_2 = 2,160$, $\hat{\alpha}_3 = 4,027$, $\hat{\alpha}_4 = 5,934$, $\hat{\beta}_1 = 0,579$, $\hat{\beta}_2 = -0,787$, $\hat{\beta}_3 = 0,761$, $\hat{\beta}_4 = -0,071$, $\hat{\beta}_5 = 0,008$ a $\hat{\beta}_6 = 0,115$. Deviance modelu $Y \sim P * M + P * B + M * B$ je 152,7545.

Zabýváme se teď testováním hypotézy $H_0 : Y \sim P * M + P * B + M * B$ proti alternativě $H_1 : Y \sim P * M * B$. Testová statistika tohoto testu je rozdílem deviancí obou modelů a počet stupňů volnosti je v tomto případě 1. Hledaná p -hodnota je 0,190, tudíž na hladině 5% nezamítáme model $Y \sim P * M + P * B + M * B$ a pokračujeme v hledání vhodného modelu dál.

Nyní bychom měli zvolit podmodel modelu $Y \sim P * M + P * B + M * B$, ale nabízí se nám tři možnosti. Můžeme zvolit podmodel $Y \sim P * M + P * B$, $Y \sim P * M + M * B$ nebo $Y \sim P * B + M * B$. Rozhodneme se podle deviancí jednotlivých podmodelů, které vycházejí po řadě 157,0586, 152,7761 a 154,438. Stejně jako v předchozím příkladě zvolíme model s nejmenší deviancí. Za podmodel vybereme tedy model $Y \sim P * M + M * B$. Opět uvedeme první ze čtyř rovnic, které reprezentují daný model

$$\begin{aligned} \log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5} \right) &= \alpha_1 + \beta_1(\text{pohlavi=muz}) + \beta_2(\text{misto=vesnice}) + \\ &+ \beta_3(\text{bezpecnostni pas=ano}) + \beta_4(\text{pohlavi=muz})(\text{misto=vesnice}) + \\ &+ \beta_5(\text{misto=vesnice})(\text{bezpecnostni pas=ano}). \end{aligned}$$

Odhadnuté koeficienty výše uvedeného modelu jsou $\hat{\alpha}_1 = 1,984$, $\hat{\alpha}_2 = 2,159$, $\hat{\alpha}_3 = 4,026$, $\hat{\alpha}_4 = 5,933$, $\hat{\beta}_1 = 0,583$, $\hat{\beta}_2 = -0,788$, $\hat{\beta}_3 = 0,764$, $\hat{\beta}_4 = -0,071$ a $\hat{\beta}_5 = 0,116$.

Testujme teď hypotézu $H_0 : Y \sim P * M + M * B$ proti alternativě $H_1 : Y \sim P * M + P * B + M * B$. Testová statistika je opět rozdíl deviancí a rozdíl v počtu parametrů je zase 1, což udává rovněž počet stupňů volnosti. Výsledná p -hodnota je 0,883, takže nezamítáme podmodel $Y \sim P * M + M * B$.

Dalším krokem bude volba podmodelu modelu $Y \sim P * M + M * B$. Tentokrát máme na výběr ze dvou možností. Kandidáty na vhodný podmodel jsou modely $Y \sim P * M + B$ a $Y \sim M * B + P$. Tyto modely mají devianci 157,1556 a 154,4522. Vybereme si tedy druhý model s nižší deviancí. Uvedeme opět první ze čtyř rovnic, které tomuto modelu vyhovují

$$\begin{aligned} \log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5} \right) &= \alpha_1 + \beta_1(\text{pohlavi=muz}) + \beta_2(\text{misto=vesnice}) + \\ &+ \beta_3(\text{bezpecnostni pas=ano}) + \\ &+ \beta_4(\text{misto=vesnice})(\text{bezpecnostni pas=ano}). \end{aligned}$$

Odhadnuté koeficienty pro model $Y \sim M * B + P$ jsou $\hat{\alpha}_1 = 2,001$, $\hat{\alpha}_2 = 2,175$, $\hat{\alpha}_3 = 4,042$, $\hat{\alpha}_4 = 5,950$, $\hat{\beta}_1 = 0,546$, $\hat{\beta}_2 = -0,823$, $\hat{\beta}_3 = 0,760$ a $\hat{\beta}_4 = 1,124$.

Testujeme-li hypotézu $H_0 : Y \sim M * B + P$ proti alternativě $H_1 : Y \sim P * M + M * B$, musíme určit testovou statistiku a počet stupňů volnosti. Počet stupňů volnosti je 1 a testová statistika je opět rozdílem deviancí obou modelů. Hledaná p -hodnota je 0,195. Na hladině 5% tedy nezamítáme podmodel $Y \sim M * B + P$.

V dalším kroku zvolíme podmodel modelu $Y \sim M * B + P$. Volba podmodelu je tentokrát jednoznačná a tímto podmodelem je model $Y \sim P + M + B$. Tento model byl popsán v kapitole 2.4, takže tady jen doplníme devianci, která je 159,6062.

Nyní testujeme hypotézu $H_0 : Y \sim P + M + B$ proti alternativě $H_1 : Y \sim M * B + P$. Počet stupňů volnosti je 1. Výsledná p -hodnota je 0.023, takže na hladině 5% zamítáme model $Y \sim P + M + B$.

Nejvhodnějším modelem pro příklad studující závažnost zranění při automobilových nehodách je model $Y \sim M * B + P$.

2.8 Lineární regrese versus multinomická regrese s ordinální odezvou

Úkolem této podkapitoly bude porovnat lineární regresi s multinomickou regresí s ordinální odezvou na praktické příkladě. Toto srovnání provedeme na příkladě s daty se spojitou odezvou, kterou buď takto necháme a použijeme lineární regresi, nebo ji rozdělíme do několika intervalů a aplikujeme multinomickou regresi s ordinální odezvou.

Popišme si konkrétní data, která byla čerpána z webové stránky [6]. Vysvětlovanou proměnnou je v tomto případě míra úmrtnosti na cirhózu jater v jednotlivých státech, která se pohybuje v intervalu [28; 129, 9]. Vysvětlujícími proměnnými jsou pak x_{i1} , která odpovídá velikosti městské populace, x_{i2} představující počet porodů u žen mezi věky 45 až 49 let, x_{i3} reprezentující konzumaci vína v přepočtu na hlavu a x_{i4} , která odpovídá konzumaci tvrdého alkoholu v přepočtu na hlavu.

Sestavme si model lineární regrese vzhledem k uvedeným datům. Rovnice tohoto modelu vypadá následovně

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4},$$

kde regresory nabývají po řadě toho významu, jaký byl uveden výše.

Model odhadneme pomocí statistického softwaru R, konkrétně použitím zabudované funkce `lm`. Výsledné odhady spolu s dalšími charakteristikami jsou uvedeny v tabulce 2.7.

Parametr	Odhad	Směr. odchylka	p-hodnota
α	-13,963	11,4004	0,2276
β_1	0,098	0,2441	0,6893
β_2	1,148	0,5830	0,0556
β_3	1,858	0,4010	0,0000
β_4	0,048	0,1334	0,7198

Tabulka 2.7: Odhady parametrů a další charakteristiky pro model lineární regrese pro data týkající se míry úmrtnosti na cirhózu jater

Z tabulky 2.7 je vidět, že na hladině 5% zamítáme nulovost parametru β_3 , z čehož vyplývá, že konzumace vína na hlavu má vliv na míru úmrtí na cirhózu jater. Zdálo by se zvláštní, že stejný vliv nebyl zjištěn u konzumace tvrdého alkoholu. To je ovšem možné vysvětlit tak, že mezi oběma veličinami existuje závislost, což by prokazoval fakt, že když jsme vynechali proměnnou týkající se konzumace vína a takto získaný model odhadli, byl zjištěn vliv konzumace tvrdého alkoholu

na míru umrtnosti na cirhózu jater. Dále je možné si všimnout, že hranici pěti procent jen o málo překročil parametr určující vliv počtu porodů u žen ve věku 45 až 49 let.

Nyní se pokusme použít stejná data pro multinomickou regresi s ordinální odezvou. Je zřejmé, že bude třeba upravit vysvětlovanou proměnnou. Tuto proměnnou tedy rozdělíme do 5 kategorií. Pokud proměnná původně nabývala hodnoty menší než 40, bude patřit do kategorie 1. Pokud byla menší než 60 a větší nebo rovno než 40, bude v kategorii 2. Jestliže byla proměnná menší než 80 a větší nebo rovno než 60, bude v kategorii 3. V případě, že byla proměnná menší než 100 a větší nebo rovno než 80, bude patřit do kategorie 4. Ve zbylých případech, tj. pokud byla proměnná větší než 100, bude zařazena do kategorie 5.

Sestavme rovnice pro multinomickou regresi s ordinální odezvou pro tento příklad. Zde uvedeme jen první rovnici, zbylé rovnice si lze jednoduše odvodit

$$\log\left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5}\right) = \alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4},$$

kde význam regresorů je shodný s významem v případě lineárního modelu.

Model odhadneme opět pomocí statistického softwaru R, použijeme k tomu balík VGAM a funkci vglm. Výsledné odhady spolu s dalšími charakteristikami jsou uvedeny v tabulce 2.8. Poznamenejme, že v této tabulce nejsou uvedeny p -hodnoty odpovídající konstantním členům, neboť ty jsme neměli zájem spočítat.

Parametr	Odhad	e ^{odhad}	Směr. odchylka	p-hodnota
α_1	9,779	17652	2,783	
α_2	13,929	1120665	3,222	
α_3	16,593	16081241	3,537	
α_4	20,527	822023055	4,259	
β_1	-0,042	0,959	0,052	0,439
β_2	-0,182	0,833	0,123	0,170
β_3	-0,230	0,795	0,089	0,006
β_4	-0,020	0,980	0,027	0,462

Tabulka 2.8: Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se míry úmrtnosti na cirhózu jater

Z tabulky 2.8 vyplývá, že zamítneme hypotézu o nulovosti koeficientu β_3 , tj. stejně jako v případě lineární regrese má konzumace vína na osobu vliv na míru úmrtnosti na cirhózu jater.

Zamysleme se nyní nad odhadem parametru β_3 , neboť ten vyšel záporný narozdíl od lineární regrese, což by se možná na první pohled mohlo zdát v rozporu. V lineární regresi větší konzumace vína na hlavu znamenala vyšší míru úmrtnosti na cirhózu jater. V případě multinomické regrese s ordinální odezvou záporný odhad parametru β_3 znamená, že vyšší hodnoty konzumace vína na osobu snižují pravděpodobnost nižší míry úmrtnosti na cirhózu jater, protože referenční kategorií je kategorie 5, která představuje nejvyšší míru úmrtnosti. To ovšem ale jinými slovy také znamená, že vyšší hodnoty konzumace vína na osobu zvyšují

pravděpodobnost vyšší míry úmrtnosti na cirhózu jater, takže mezi oběma modely není žádný rozpor.

Dále si lze všimnout, že se zcela ztratil vliv proměnné představující počet porodů u žen mezi lety 45 až 49. Z toho plyne, že při použití modelu multinomické regrese s ordinální odezvou došlo k určité ztrátě informace, což lze přičítat tomu, že při úpravě dat došlo ke zjednodušení situace, což mělo za následek právě onu ztrátu informace.

V tomto případě lze učinit závěr, že pokud by to bylo opravdu nutné, lze model lineární regrese nahradit modelem multinomické regrese s ordinální odezvou. Nicméně tento postup, kde jsme rozdělili spojitou odezvu do kategorií, ztrácí informaci a obecně jej nelze doporučit.

2.9 Výpočetní aspekty iterativního hledání odhadů parametrů β

Cílem této podkapitoly je prostudovat hledání iterativního odhadu parametrů β . Zaměříme se na multinomickou regresi s nominální a s ordinální odezvou. Bude nás hlavně zajímat, kolik iterací je potřeba, abychom dostali již relativně rozumné odhady.

Pro případ multinomické regrese s nominální odezvou využijeme data k příkladu z podkapitoly 2.2. Jedná se o studium závislosti příslušnosti k politické straně na pohlaví a rase. Příklad bude stejně jako v podkapitole 2.2 řešen pomocí statistického softwaru R s využitím balíku `nnet` a funkce `multinom`. Navíc ovšem použijeme parametr `maxit`, který určuje maximální počet iterací. Rovnice příslušící k tomuto modelu jsou (2.5) a (2.6). V tabulce 2.9 jsou uvedeny výsledné maximálně věrohodné odhady pro jednotlivé počty iterací.

Iterace	α_1	α_2	β_{11}	β_{12}	β_{21}	β_{22}
2.	0,109	-0,030	0,095	0,085	-0,057	-0,056
3.	0,109	0,288	0,240	0,492	-0,080	-0,312
4.	-0,030	0,233	0,185	0,727	-0,208	-0,495
5.	0,064	0,202	0,040	0,840	-0,124	-0,537
6.	0,147	0,333	-0,055	1,023	-0,368	-0,467
7.	-0,150	0,131	0,574	1,348	-0,009	-0,885
8.	-0,034	0,271	0,478	0,997	-0,135	-1,460
9.	0,075	0,357	0,152	1,116	-0,406	-1,098
10.	0,050	0,335	0,220	1,118	-0,353	-1,160

Tabulka 2.9: Odhady parametrů pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na počtu iterací

Z tabulky 2.9 vidíme, že se odhady v průběhu iterací relativně dost mění, takže 10 iterací je bezpochyby oprávněných. Je však možné podotknout, že už 9. iterace je natolik přesná, že by se jí dalo skončit.

Nyní se zaměříme na multinomickou regresi s ordinální odezvou. Pro tento případ využijeme data z podkapitoly 2.4. Tato data se zabývají vztahem mezi

závažností zranění při automobilových nehodách a pohlavím, místem nehody a užitím bezpečnostního pásu. Maximálně věrohodné odhady získáme stejně jako v podkapitole 2.4 pomocí statistického softwaru R, konkrétně využijeme balík `VGAM` a funkci `vglm`. Oproti kapitole 2.4 ještě použijeme parametr `maxit`, který stejně jako v případě multinomické regrese s nominální odezvou určuje maximální počet iterací. Studovaný model představuje rovnice (2.10). Další tři rovnice popisující tento model si lze z ní snadno odvodit. V tabulce 2.10 jsou uvedeny výsledné maximálně věrohodné odhady pro jednotlivé počty iterací.

Iterace	α_1	α_2	α_3	α_4	β_1	β_2	β_3
1.	1,977	2,141	3,956	5,818	0,545	-0,771	0,822
2.	1,977	2,152	4,016	5,919	0,545	-0,773	0,824
3.	1,977	2,151	4,018	5,925	0,545	-0,773	0,824
4.	1,977	2,151	4,018	5,925	0,545	-0,773	0,824

Tabulka 2.10: Odhady parametrů pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na počtu iterací

Z tabulky 2.10 vyplývá, že pokud uvažujeme přesnost na tři desetinná místa, tak 4 iterace jsou zbytečné. Už 3. iterace dává stejné odhady jako 4. Navíc si lze všimnout, že již 2. iterace je natolik přesná, že by se s ní dalo skončit.

Z uvedených příkladů by se dalo usuzovat, že iterativní postup hledání odhadů parametrů v případě multinomické regrese s nominální odezvou konverguje pomalu a jsou nutné všechny iterace, které R provede. V případě multinomické regrese s ordinální odezvou iterativní postup konverguje rychle s tím, že ani není nutné provést všechny iterace, které R provádí. Aby takový závěr byl opravdu podložený, bylo by nutné otestovat více případů.

2.10 Grafické hledání odhadů parametrů β

Úkolem této podkapitoly bude blíže prozkoumat logaritmickou věrohodnostní funkci. Pokusíme se zjistit, zda by bylo možné pro dva parametry β najít maximum logaritmické věrohodnosti graficky. Tuto problematiku budeme studovat na všech třech představených modelech, tj. na logistické regresi, multinomické regresi s nominální odezvou a na multinomické regresi s ordinální odezvou.

Zaměříme se nyní na případ logistické regrese, využijeme k tomu data z webové stránky [7], jedná se o data související s rakovinou prostaty. Popíšeme si nyní tato data. Jedná se o 380 pozorování. Jako vysvětlovaná proměnná v těchto datech figuruje, zda nádor pronikl do obalu prostaty. Tato proměnná má dva stavy pronikl a nepronikl, což vysvětluje, proč jsou tato data vhodná pro model logistické regrese. V datech se nachází 7 vysvětlujících proměnných, kterými jsou věk a rasa, která nabývá dvou hodnot bílá a černá. Další vysvětlující proměnnou je výsledek rektálního vyšetření, kterým může být žádná uzlovitost prostaty, uzlovitost v levém laloku prostaty, uzlovitost v pravém laloku prostaty a uzlovitost v obou lalocích. Následuje proměnná, která určuje, zda při rektálním vyšetření bylo objeveno, že rakovina prorostla pouzdro prostaty. Tato proměnná nabývá

hodnoty ne a ano. Dále je zahrnuta hodnota PSA (Prostatic Specific Antigen value), objem nádoru získaný z ultrazvuku a hodnota Total Gleason score, kterým se určuje prognóza rakoviny a nabývá hodnot 0 až 10.

Rovnice modelu pro tento případ vypadá následovně

$$\log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \alpha + \beta_1 x_{i1} + \beta_{21} D_{21}^{(i)} + \beta_{31} D_{31}^{(i)} + \beta_{32} D_{32}^{(i)} + \beta_{33} D_{33}^{(i)} + \beta_{41} D_{41}^{(i)} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}, \quad (2.11)$$

kde $D_{jl}^{(i)}$ mají význam dummy proměnných, jak jsou popsány v kapitole 1.3 a spolu s proměnnými x_{ij} reprezentují vysvětlující proměnné v tom pořadí, jak byly popsány výše.

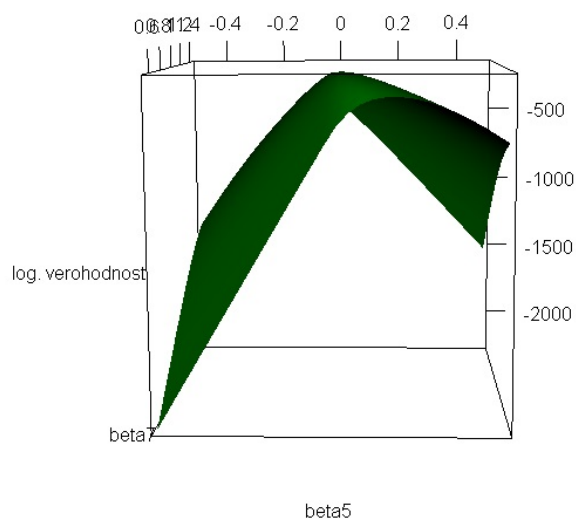
Nyní bychom chtěli maximalizovat logaritmicou věrohodnost, která je pro tento model reprezentována vzorcem (1.5). Do tohoto vzorce bude nutné za parametry, které nebudeme zkoumat, dosadit maximálně věrohodné odhady. Bude proto třeba model odhadnout. To provedeme pomocí statistického softwaru R, kde použijeme funkci `glm`, dále bude třeba využít funkce `factor`, poněvadž se v modelu vyskytují kategoriální vysvětlující proměnné. Výsledné odhady parametrů a další charakteristiky uvádí tabulka 2.11.

Parametr	Odhad	e ^{odhad}	Směr. odchylka	p-hodnota
α	-6,967	0,001	1,619	
β_1	-0,012	0,988	0,020	0,434
β_{21}	-0,651	0,521	0,472	0,842
β_{31}	0,730	2,075	0,360	0,000
β_{32}	1,509	4,524	0,377	0,000
β_{33}	1,387	4,004	0,462	0,000
β_{41}	0,492	1,636	0,464	0,000
β_5	0,030	1,030	0,010	0,000
β_6	-0,011	0,989	0,008	0,060
β_7	0,963	2,618	0,169	0,000

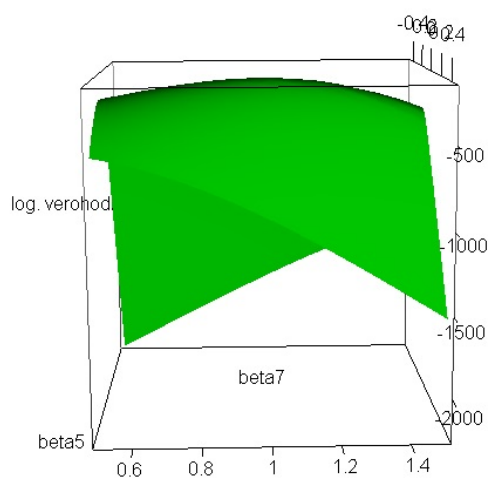
Tabulka 2.11: Odhady parametrů a další charakteristiky pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty

Poznamenejme, že uvedená p -hodnota vychází z testu poměrem věrohodností (konkrétně v R byla získána funkcí `anova`).

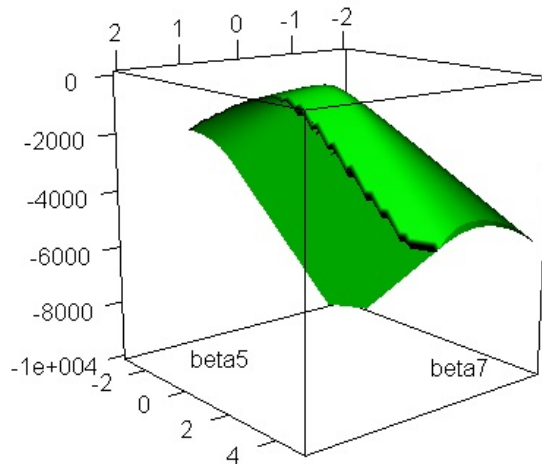
Nyní budeme graficky zkoumat závislost logaritmicke věrohodnosti na hodnotách parametrů β_5 a β_7 , což jsou parametry vyjadřující hodnotu PSA a hodnotu Total Gleason Score. Ke grafickému vykreslení logaritmicke věrohodnosti bylo použito funkce `persp3d`. Zde uvádíme nejprve dva grafy 2.1 a 2.2, z prvního lze vyčíst hodnotu koeficientu β_5 , při kterém nabývá logaritmicke věrohodnost svého maxima, a z druhého lze zase vyčíst hodnotu koeficientu β_7 , při kterém nabývá logaritmicke věrohodnost svého maxima. Je tedy vidět, že logaritmicke věrohodnost nabývá svého maxima přibližně pro hodnoty $\beta_5 = 0$ a $\beta_7 = 1$, což odpovídá maximálně věrohodným odhadům, které jsou uvedeny v tabulce 2.11, tj. $\hat{\beta}_5 = 0,030$ a $\hat{\beta}_7 = 0,963$.



Obrázek 2.1: Hodnota logaritmicke věrohodnosti pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty v závislosti na parametru β_5 a β_7 z pohledu parametru β_5



Obrázek 2.2: Hodnota logaritmicke věrohodnosti pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty v závislosti na parametru β_5 a β_7 z pohledu parametru β_7



Obrázek 2.3: Širší pohled oproti obrázku 2.1, resp. 2.2 na logaritmicou věrohodnostní funkci pro data týkající se rakoviny prostaty

Zajímavý je pohled na danou logaritmicou věrohodnostní funkci z širší perspektivy. Tento pohled představuje obrázek 2.3, na kterém je vidět, že logaritmicá funkce na jedné straně padá do mínus nekonečna daleko rychleji.

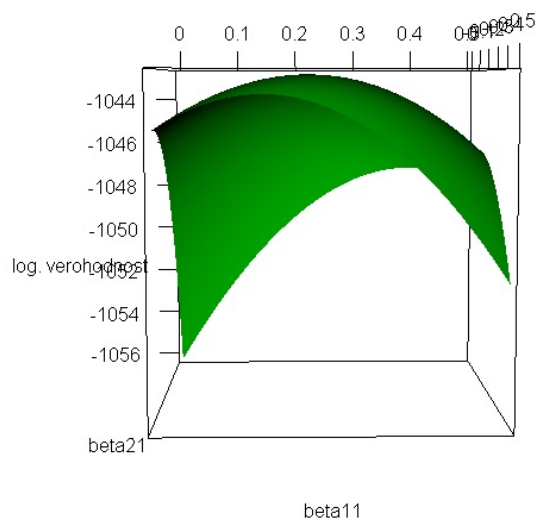
Prostudujme nyní případ multinomické regrese s nominální odezvou. Použijeme k tomu data z podkapitoly 2.2. Pomocí těchto dat lze studovat vztah mezi příslušností k politické straně a pohlavím a rasou. Zopakujme na tomto místě rovnice (2.5) a (2.6) představující model multinomické regrese s nominální odezvou pro daná data.

$$\log \left(\frac{P(\text{strana}=\text{demokrat})}{P(\text{strana}=\text{nezavisly})} \right) = \alpha_1 + \beta_{11}(\text{pohlavi}=\text{zena}) + \beta_{12}(\text{rasa}=\text{cerny}),$$

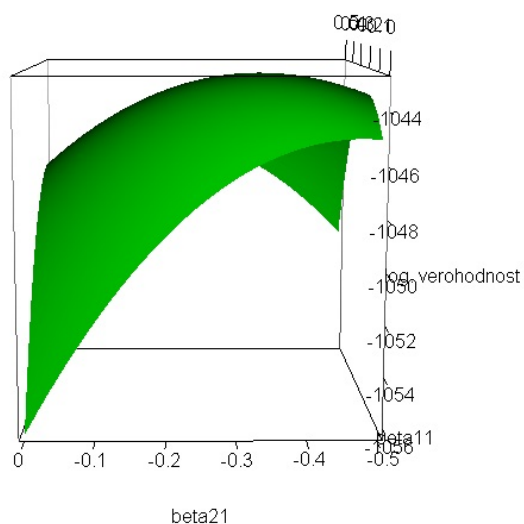
$$\log \left(\frac{P(\text{strana}=\text{republikan})}{P(\text{strana}=\text{nezavisly})} \right) = \alpha_2 + \beta_{21}(\text{pohlavi}=\text{zena}) + \beta_{22}(\text{rasa}=\text{cerny}).$$

Logaritmicá věrohodnost, pro kterou budeme hledat maximum, je dána vzorcem (2.4). Toto maximum budeme zkoumat v závislosti na hodnotách parametrů β_{11} a β_{21} , což jsou parametry představující vliv ženského pohlaví na pravděpodobnost, že člověk přísluší k demokratům oproti pravděpodobnosti, že je nezávislý, a na pravděpodobnost, že člověk přísluší k republikánům oproti pravděpodobnosti, že je nezávislý. Za zbylé parametry do logaritmicé věrohodnosti dosadíme maximálně věrohodné odhady, které jsme už pro tento případ spočítali a jsou uvedeny v tabulce 2.2. K vykreslení grafů použijeme stejně jako v předešlém případě funkci `persp3d`. Uvedeme nejprve dva grafy 2.4 a 2.5, z prvního lze vyčíst přibližnou hodnotu parametru β_{11} , pro niž nabývá logaritmicá funkce maxima, a z druhého hodnotu parametru β_{21} .

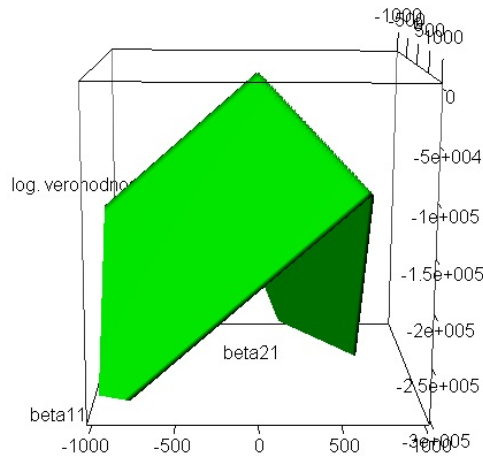
Z grafu 2.4 se zdá, že odhad parametru β_{11} by měl být přibližně 0,22 a také přesně vychází 0,220. Odhad parametru β_{21} by měl podle grafu 2.5 být -0,35 a



Obrázek 2.4: Hodnota logaritmicke věrohodnosti pro model multinomicke regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na parametru β_{11} a β_{21} z pohledu parametru β_{11}



Obrázek 2.5: Hodnota logaritmicke věrohodnosti pro model multinomicke regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na parametru β_{11} a β_{21} z pohledu parametru β_{21}



Obrázek 2.6: Širší pohled oproti obrázku 2.4, resp. 2.5 na logaritmickou věrohodnostní funkci pro data týkající se příslušnosti k politické straně

jeho přesná hodnota je -0,353. Je tedy vidět, že grafický způsob nalezení odhadů zde zafungoval velmi dobře.

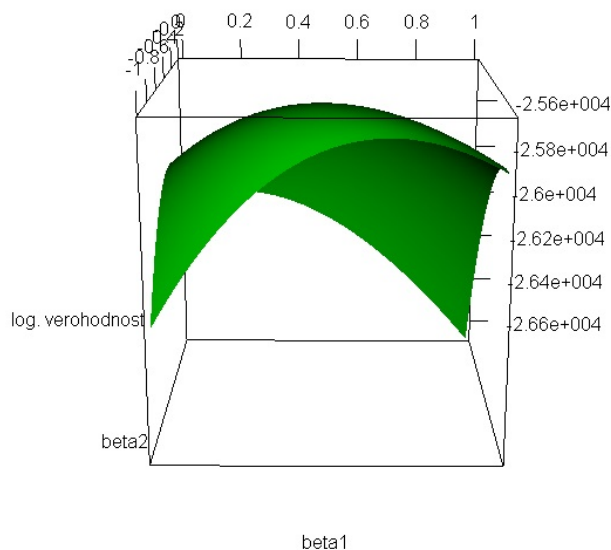
Když se podíváme na graf logaritmické věrohodnostní funkce z širší perspektivy, tak vidíme, že na jedné straně hodnota logaritmické funkce nabývá minus nekonečna rychleji, což zobrazuje obrázek 2.6.

Posledním případem, kterým se budeme zabývat, je případ multinomické regrese s ordinální odezvou. Použijeme data z podkapitoly 2.4, která se zabývají vztahem mezi závažností zranění při automobilové nehodě na jedné straně a pohlavím, místem nehody a užitím bezpečnostního pásu na druhé straně. Model multinomické regrese s ordinální odezvou pro tato data představují čtyři rovnice, my z nich tady zopakujeme první rovnici (2.10) a zbylé rovnice je již snadné si domyslet

$$\log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5} \right) = \alpha_1 + \beta_1(\text{pohlavi=muz}) + \beta_2(\text{misto=vesnice}) + \beta_3(\text{bezpecnostni pas=ano}).$$

Logaritmickou věrohodnost pro multinomickou regresi s ordinální odezvou je vyjádřena vzorcem (2.11)

Maximum této logaritmické věrohodnosti budeme hledat graficky v závislosti na koeficientech β_1 a β_2 , kde koeficient β_1 vyjadřuje, jak fakt, že člověk je muž, zvyšuje pravděpodobnost lehkého zranění, a koeficient β_2 vyjadřuje, jak fakt, že se nehoda stala na vesnici, zvyšuje pravděpodobnost lehkého zranění. Za zbylé parametry do logarimické věrohodnosti dosadíme maximálně věrohodné odhady, které jsme už pro tento případ spočítali a jsou uvedeny v tabulce 2.4. K vykreslení grafů použijeme stejně jako v předešlých dvou případech funkci `persp3d`. Opět uvedeme nejprve dva grafy 2.7 a 2.8, z prvního lze vyčíst přibližnou hodnotu



Obrázek 2.7: Hodnota logaritmické věrohodnosti pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na parametru β_1 a β_2 z pohledu parametru β_1

parametru β_1 , pro niž nabývá logaritmická funkce maxima, a z druhého hodnotu parametru β_2 .

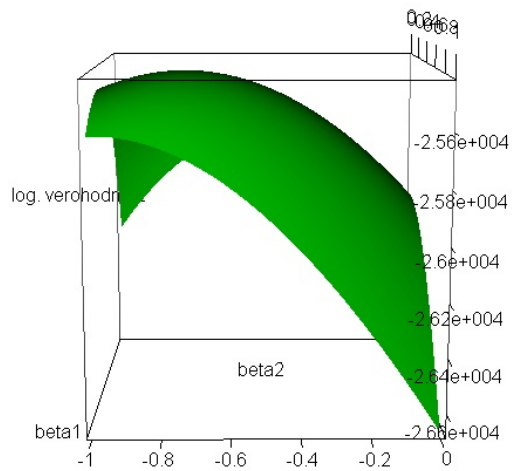
Z grafů 2.7 a 2.8 je tedy vidět, že logaritmická věrohodnost nabývá svého maxima přibližně pro hodnoty $\beta_1 = 0,5$ a $\beta_2 = -0,8$, což odpovídá maximálně věrohodným odhadům, které jsou uvedeny v tabulce 2.4 v podkapitole 2.4, tj. $\hat{\beta}_1 = 0,545$ a $\hat{\beta}_2 = -0,773$.

Když se pokusíme zobrazit logaritmickou věrohodnostní funkci pro větší interval hodnot (obrázek 2.9), tak zjistíme, že pro určité hodnoty se nevykreslí, neboť ve výpočtu dojde k situaci, že je třeba logaritmovat hodnotu příliš blízkou nule.

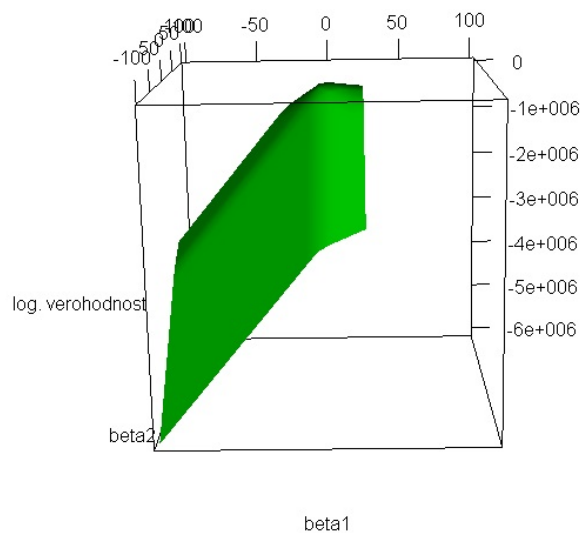
Ve všech třech případech jsme byli schopni relativně přesně nalézt odhady dvou parametrů graficky. Pokud bychom chtěli ovšem uvažovat o tomto grafickém způsobu hledání odhadů jako o alternativě k iterativní metodě, tak nastane problém, neboť pokud jsou součástí modelů další parametry, je třeba jako v našem případě nalézt odhady těchto dalších parametrů iterativní metodou, která nám však taky určí odhady hledaných dvou parametrů, naše grafické hledání je pak zbytečné. Dalo by se použít pouze na případy, kde v modelu figurují jen konstantní člen a jeden další parametr, tyto případy však nejsou příliš časté.

2.11 Naivní hledání odhadů parametrů β

V minulé podkapitole jsme hledali maximum logaritmické věrohodnosti graficky, cílem této podkapitoly bude opět hledání maxima logaritmické věrohodnosti avšak dosazením vybraných hodnot. Motivací je vyzkoušet nalezení odhadů tímto



Obrázek 2.8: Hodnota logaritmicke věrohodnosti pro model multinomicke regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na parametru β_1 a β_2 z pohledu parametru β_2



Obrázek 2.9: Širší pohled oproti obrázku 2.7, resp. 2.8 na logaritmicke věrohodnostní funkci pro data týkající se závažnosti zranění při automobilových nehodách

jednoduchým způsobem, jenž se ukáže, že vede k dobrým výsledkům. Na závěr této podkapitoly budeme také diskutovat použitelnost tohoto přístupu.

Opět budeme zkoumat tři v této práci popsané modely - model logistické regrese, model multinomické regrese s nominální odezvou a model multinomické regrese s ordinální odezvou.

Pro model logistické regrese použijeme data z předchozí podkapitoly týkající se rakoviny prostaty. Rovnice vyjadřující tento model je rovnice (2.11). Budeme studovat hodnotu logaritmické věrohodnosti v závislosti na hodnotách parametrů β_5 a β_7 . Do vytvořené funkce logaritmické věrohodnosti pro tato data budeme dosazovat hodnoty z intervalu $[-2, 2]$ a za zbylé parametry maximálně věrohodné odhady, které jsou uvedeny v předchozí podkapitole v tabulce 2.11. Hodnoty logaritmické věrohodnostní funkce jsou uvedeny v tabulce 2.12.

Z tabulky 2.12 je patrné, že jsme dosazovali vybrané hodnoty parametrů β_5 a β_7 ve třech etapách. V každé etapě jsme vybrali takové hodnoty studovaných parametrů, při nichž byla logaritmická věrohodnostní funkce největší, dále jsme zvolili okolí tohoto bodu, z něhož jsme pak dosazovali hodnoty v další etapě. Z tabulky 2.12 je také zřejmé, že se odhad parametrů v žádné ze tří etap neměnil a výsledný odhad tak je $\beta_5 = 0$ a $\beta_7 = 1$. Odhady získané iterativní metodou přitom jsou $\hat{\beta}_5 = 0,030$ a $\hat{\beta}_7 = 0,963$. Vidíme tedy, že při pouhých třech kolech dosazování vybraných hodnot jsme dospěli k relativně přesným odhadům.

Dalším modelem, u něhož budeme hledat odhad parametrů, je model multinomické regrese s nominální odezvou. Tento model budeme zkoumat na datech z tabulky 2.1 z podkapitoly 2.2. Jedná se o data zkoumající závislost mezi příslušností k politické straně a pohlavím a rasou. Rovnice představující model multinomické regrese s nominální odezvou pro tato data jsou rovnice (2.5) a (2.6). Logaritmická věrohodnostní funkce je pak dána vzorcem (2.4). Budeme hledat odhady pro parametry β_{11} a β_{12} tak, že budeme zkoumat, pro které hodnoty těchto parametrů nabývá logaritmická věrohodnostní funkce největší hodnoty. Hodnoty logaritmické věrohodnostní funkce v závislosti na vybraných hodnotách parametrů β_{11} a β_{12} jsou uvedeny v tabulce 2.13. Za zbylé parametry jsme dosadili maximálně věrohodné odhady, které jsou uvedeny v tabulce 2.2 v podkapitole 2.2.

Z tabulky 2.13 vyplývá, že jsme použili stejnou strategii hledání největší hodnoty logaritmické věrohodnostní funkce jako v předchozím případě. Narozdíl od předchozího případu se zde odhad parametrů měnil. Jak je z tabulky 2.13 vidět, výsledný odhad parametrů β_{11} a β_{21} je $\beta_{11} = 0,25$ a $\beta_{21} = -0,25$. Maximálně věrohodné odhady získané pomocí Newton-Raphsonovy metody jsou $\hat{\beta}_{11} = 0,220$ a $\hat{\beta}_{21} = -0,353$. Je tedy zřejmé, že ačkoliv odhad parametru β_{11} je poměrně přesný, tak abychom získali přesný odhad parametru β_{21} , bylo by třeba udělat ještě jedno kolo dosazování vybraných hodnot.

Prozkoumejme nyní třetí model z pohledu hledání odhadů pomocí dosazování vybraných hodnot. Tímto modelem je model multinomické regrese s ordinální odezvou. Tento model budeme zkoumat na datech z tabulky 2.3 z podkapitoly 2.4. Jedná se o data zabývající se závažností zranění při automobilových nehodách v závislosti na pohlaví, místě nehody a použití bezpečnostního pásu. Rovnice pro model multinomické regrese s ordinální odezvou pro daná data je reprezentována vzorcem (2.10) a zbylé rovnice je snadné si z této rovnice již odvodit. Logaritmická věrohodnostní funkce pro tento model je dána vzorcem (2.11). Budeme se snažit odhadnout parametry β_1 a β_2 . Za ostatní parametry dosadíme do

logaritmické věrohodnostní funkce maximální věrohodné odhady získané Newton-Raphsonovou metodou, které jsou uvedeny v tabulce 2.4 v podkapitole 2.4.

Z tabulky 2.14 je patrné, že v tomto případě stejně jako v předchozím se odhad parametrů β_1 a β_2 vyvíjel a že jsme opět využili stejnou strategii hledání odhadů jako v předchozích dvou případech. Odhady, které jsme touto strategií získali, jsou $\beta_1 = 0,5$ a $\beta_2 = -0,75$. Maximálně věrohodné odhady získané Newton-Raphsonovou metodou jsou $\hat{\beta}_1 = 0,545$ a $\hat{\beta}_2 = -0,773$. Odtud tedy vidíme, že hledání odhadů dosazením vybraných hodnot nám v tomto případě dalo poměrně přesné odhady.

Závěrem této kapitoly uveďme malé srovnání odhadů parametrů dle metody jejich získání. Přehled těchto odhadů je uveden v tabulce 2.15.

Z tabulky 2.15 je vidět, že hledání odhadů graficky i dosazováním hodnot dávají poměrně přesné odhady. Vyjímkou byl jen parametr β_{21} v případě dosazování vybraných hodnot a tato nepřesnost, jak už jsme psali, by se dala vylepšit, pokud bychom provedli ještě jedno kolo dosazování.

Zamysleme se však nad použitelností těchto způsobů hledání odhadů vzhledem k počtu odhadovaných parametrů. V případě grafického způsobu hledání odhadů jsme nutně omezeni na maximálně dva parametry, neboť více jak tři dimenze nelze graficky zobrazit. Naproti tomu v případě dosazování vybraných hodnot nejsme teoreticky omezováni, ale prakticky je třeba si uvědomit, že počet dosazení roste exponenciálně s počtem parametrů, což tvoří značnou komplikaci.

Je také nutné poznamenat, že hledání odhadů pomocí dosazování hodnot by nefungovalo, kdyby měla logaritmická věrohodnostní funkce divoký průběh, tj. kdyby měla i nějaká lokální minima a maxima.

2.12 Generování dat pro multinomickou regresi s nominální a ordinální odezvou

Úkolem této podkapitoly je generovat náhodná data, což prakticky znamená generovat odezvu při známých hodnotách regresorů, pro účely simulačních studií. Budeme pracovat se statistickým softwarem R.

Podívejme se nejprve na generování dat pro multinomickou regresi s nominální odezvou. Budeme uvažovat model se třemi kategoriemi a čtyřmi vysvětlujícími proměnnými, to znamená, že platí $Z \sim \text{Multinom}(n, \pi_1, \pi_2, \pi_3)$. Třetí kategorie je referenční. Počet pozorování stanovíme na 1000. Tento model reprezentuje následující rovnice, která vychází z rovnice (2.1)

$$\log \left(\frac{\pi_j(\mathbf{x}_i)}{\pi_3(\mathbf{x}_i)} \right) = \alpha_j + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \beta_{j3}x_{i3} + \beta_{j4}x_{i4}, \quad j = 1, \dots, 2, n = 1, \dots, 1000,$$

kde hodnoty proměnných x_{i1}, x_{i2}, x_{i3} a x_{i4} by bylo možné libovolně zadat, avšak mi si je vygenerujeme. Od proměnné x_{i1} budeme chtít, aby nabývala přirozených čísel od 1 do 20, k čemuž použijeme v R funkci `sample`. Proměnné x_{i2} a x_{i4} zvolíme jako binární, kde 0 a 1 budou nabývat se stejnou pravděpodobností, k jejich vygenerování opět využijeme funkci `sample`. Proměnná x_{i3} bude nabývat reálných hodnot mezi 10 a 20 a použijeme k tomu funkci `runif`.

Dále je třeba zvolit parametry, ty byly zadány tak, jak uvádí tabulka 2.16.

Nyní přejdeme k popisu generování vysvětlované proměnné Z_i , tato proměnná nabývá hodnot 1, 2 nebo 3 s pravděpodobnostmi $\pi_1(\mathbf{x}_i)$, $\pi_2(\mathbf{x}_i)$ a $\pi_3(\mathbf{x}_i)$. První dvě pravděpodobnosti jsou dány vzorcem (2.2). Poslední pravděpodobnost je dána vzorcem (2.3). Proměnná Z_i má tedy multinomické rozdělení s uvedenými pravděpodobnostmi. K jejímu vygenerování použijeme funkci `rmultinomial` z balíku `multinomRob`. V této funkci nastavíme `n=1` a použijeme ji 1000 krát. Uveďme ještě, že pravděpodobnost $\pi_3(\mathbf{x}_i)$ je nanejvýš vhodné zadat jako $1 - \pi_1(\mathbf{x}_i) - \pi_2(\mathbf{x}_i)$, aby se nestalo, že vlivem zaokrouhlování se pravděpodobnosti nevysčítají na 1.

Model multinomické regrese s nominální odezvou pro námi vygenerovaná data si odhadneme. Využijeme k tomu funkci `multinom` z balíku `nnet`, kde bude třeba předem upravit vysvětlovanou proměnnou tak, aby za referenční hodnotu byla považována hodnota 3, k tomu nám poslouží funkce `relevel`. Výsledné odhady parametrů spolu s dalšími charakteristikami jsou uvedeny v tabulce 2.16.

Jak je z tabulky 2.16 vidět, odhady parametrů příliš neodpovídají námi zvoleným parametrům. Vypočítané p -hodnoty vycházejí z testu poměrem věrohodností, kde se testovala nulovost vždy dvou parametrů příslušejících k jedné vysvětlující proměnné. U parametrů náležejících k vysvětlujícím proměnným x_{i1} , x_{i2} a x_{i3} je p -hodnota rovna nule pochopitelná. S podivem ovšem je, že se zamítla nulovost parametrů β_{14} a β_{24} , neboť vysvětlovaná proměnná x_{i4} by neměla mít žádný význam.

Pokusme se nyní vygenerovat data pro multinomickou regresi s ordinální odezvou. Budeme uvažovat model se čtyřmi kategoriemi a třemi vysvětlujícími proměnnými. Počet pozorování stanovíme opět na 1000. Tento model reprezentují následující tři rovnice

$$\log \left(\frac{\pi_1(\mathbf{x}_i)}{\pi_2(\mathbf{x}_i) + \pi_3(\mathbf{x}_i) + \pi_4(\mathbf{x}_i)} \right) = \alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \quad (2.12)$$

$$\log \left(\frac{\pi_1(\mathbf{x}_i) + \pi_2(\mathbf{x}_i)}{\pi_3(\mathbf{x}_i) + \pi_4(\mathbf{x}_i)} \right) = \alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \quad (2.13)$$

$$\log \left(\frac{\pi_1(\mathbf{x}_i) + \pi_2(\mathbf{x}_i) + \pi_3(\mathbf{x}_i)}{\pi_4(\mathbf{x}_i)} \right) = \alpha_3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \quad (2.14)$$

kde hodnoty proměnných x_{i1} , x_{i2} a x_{i3} náhodně vygenerujeme. Proměnné x_{i1} a x_{i3} budou v našem případě binární a budou 0 a 1 nabývat se stejnou pravděpodobností. K jejich vygenerování použijeme funkci `sample`. Proměnná x_{i2} bude nabývat reálných čísel mezi 10 a 20. K tomu, abychom vygenerovali tuto proměnnou, použijeme funkci `runif`.

Hodnoty parametrů zvolíme tak, jak je uvedeno v tabulce 2.17.

Nyní je potřeba vygenerovat vysvětlovanou proměnnou Z_i . Tato proměnná má multinomické rozdělení a nabývá hodnot 1, 2, 3 a 4. Je potřeba však určit pravděpodobnosti, s jakými těchto hodnot nabývá. Vzorce pro jednotlivé pravděpodobnosti si pro tento případ odvodíme. U těchto pravděpodobností již nebudeme uvádět závislost na \mathbf{x}_i , ale je zřejmé, že tato závislost existuje.

Upravíme rovnice (2.12), (2.13) a (2.14) na následující tvary

$$\pi_1 = e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} (\pi_2 + \pi_3 + \pi_4), \quad (2.15)$$

$$\pi_1 + \pi_2 = e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} (\pi_3 + \pi_4), \quad (2.16)$$

$$\pi_1 + \pi_2 + \pi_3 = e^{\alpha_3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} \pi_4. \quad (2.17)$$

Dále je třeba si uvědomit, že platí vztah

$$1 = \pi_1 + \pi_2 + \pi_3 + \pi_4. \quad (2.18)$$

Z rovnice (2.15) a platnosti vztahu (2.18) odvodíme vzorec pro π_1

$$\pi_1 = e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} (1 - \pi_1),$$

$$\pi_1 (1 + e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}) = e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}},$$

$$\pi_1 = \frac{e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}. \quad (2.19)$$

Z rovnice (2.17) a platnosti vztahu (2.18) odvodíme vzorec pro π_4

$$1 - \pi_4 = e^{\alpha_3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} \pi_4,$$

$$\pi_4 = \frac{1}{1 + e^{\alpha_3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}. \quad (2.20)$$

Z rovnice (2.16) a s použitím vztahu (2.18) odvodíme vzorec pro π_2

$$\pi_1 + \pi_2 = e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} (1 - \pi_1 - \pi_2),$$

$$\pi_2 (1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}) = e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} - \pi_1 (1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}),$$

$$\pi_2 = \frac{e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} - \pi_1,$$

$$\pi_2 = \frac{e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} - \frac{e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}. \quad (2.21)$$

Rovněž z rovnice (2.16) a s použitím vztahu (2.18) odvodíme vzorec pro π_3

$$1 - \pi_3 - \pi_4 = e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} (\pi_3 + \pi_4),$$

$$1 - \pi_4 (1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}) = \pi_3 (1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}),$$

$$\pi_3 = \frac{1}{1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} - \pi_4,$$

$$\pi_3 = \frac{1}{1 + e^{\alpha_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} - \frac{1}{1 + e^{\alpha_3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}. \quad (2.22)$$

K vygenerování hodnoty proměnné Z_i použijeme stejně jako v případě modelu multinomické regrese s nominální odezvou funkci `rmultinomial`, kde zadáme jednotlivé pravděpodobnosti podle vzorců (2.19), (2.20), (2.21) a (2.22). Připomeňme však, že i zde je kvůli chybám zaokrouhlování vhodné zadat jednu pravděpodobnost jako rozdíl jedničky a zbylých pravděpodobností.

Důležité u uvedených vzorců je si všimnout, že musí platit $\pi_1 > 0$, $\pi_2 > 0$, $\pi_3 > 0$ a $\pi_4 > 0$. První a poslední nerovnost je bez problémů splněna pro všechny hodnoty parametrů. Z nerovnosti $\pi_2 > 0$ však plyne, že musí platit $a_1 < a_2$, a z nerovnosti $\pi_3 > 0$ plyne, že $a_2 < a_3$, tj. musí nutně platit

$$a_1 < a_2 < a_3.$$

Na platnost tohoto vztahu je nutné myslet při volbě hodnot parametrů.

Model multinomické regrese s ordinální odezvou pro námi vygenerovaná data si odhadneme. Použijeme k tomu funkci `vg1m` z balíku `VGAM`. Výsledné odhady a další charakteristiky jsou uvedeny v tabulce 2.17.

Z tabulky 2.17 je patrné, že odhady parametrů relativně dobře odpovídají námi zadaným hodnotám parametrů. Uvedená p -hodnota vychází z testů poměrem věrohodností, kde se testovala nulovost jednotlivých parametrů. Test na nulovost parametru β_2 vyšel dle předpokladu tak, že nulovost zamítáme. Naopak u parametru β_3 nulovost zamítnout nemůžeme, což není nijak zvláštní, neboť tento parametr by neměl mít žádný význam. Překvapivé ovšem je, že nulovost parametru β_1 nemůžeme zamítnout.

β_5	β_7	$L(\boldsymbol{\beta}, \mathbf{Y})$	β_5	β_7	$L(\boldsymbol{\beta}, \mathbf{Y})$	β_5	β_7	$L(\boldsymbol{\beta}, \mathbf{Y})$
-2	-2	-10118,84	-1	0	-4519,86	-0,5	0,5	-2247,95
-2	-1	-9064,84	-1	0,5	-3993,03	-0,5	0,75	-1986,72
-2	0	-8010,84	-1	1	-3469,54	-0,5	1	-1733,52
-2	1	-6957,64	-1	1,5	-2989,05	-0,5	1,25	-1507,98
-2	2	-5980,70	-1	2	-2664,34	-0,5	1,5	-1344,29
-1	-2	-6627,85	-0,5	0	-2774,38	-0,25	0,5	-1376,16
-1	-1	-5573,85	-0,5	0,5	-2247,95	-0,25	0,75	-1118,76
-1	0	-4519,86	-0,5	1	-1733,53	-0,25	1	-882,75
-1	1	-3469,54	-0,5	1,5	-1344,29	-0,25	1,25	-712,40
-1	2	-2664,34	-0,5	2	-1289,55	-0,25	1,5	-659,40
0	-2	-3136,86	0	0	-1029,24	0	0,5	-513,29
0	-1	-2082,86	0	0,5	-513,29	0	0,75	-297,36
0	0	-1029,24	0	1	-194,65	0	1	-194,65
0	1	-194,65	0	1,5	-467,23	0	1,25	-260,45
0	2	-1070,83	0	2	-1070,83	0	1,5	-467,23
1	-2	$-\infty$	0,5	0	-525,29	0,25	0,5	-304,45
1	-1	$-\infty$	0,5	0,5	-553,96	0,25	0,75	-318,46
1	0	$-\infty$	0,5	1	-922,04	0,25	1	-448,20
1	1	$-\infty$	0,5	1,5	-1522,23	0,25	1,25	-679,23
1	2	$-\infty$	0,5	2	$-\infty$	0,25	1,5	-972,53
2	-2	$-\infty$	1	0	$-\infty$	0,5	0,5	-553,96
2	-1	$-\infty$	1	0,5	$-\infty$	0,5	0,75	-696,29
2	0	$-\infty$	1	1	$-\infty$	0,5	1	-922,04
2	1	$-\infty$	1	1,5	$-\infty$	0,5	1,25	-1206,37
2	2	$-\infty$	1	2	$-\infty$	0,5	1,5	-1522,23

Tabulka 2.12: Hodnoty logaritmickej vĕrohodnosti pro model logistickej regrese pro data tŕykajŕící se pokroĕilosti stádia rakoviny prostaty v závislosti na parametrech β_5 a β_7

β_{11}	β_{21}	$L(\beta, Y)$	β_{11}	β_{21}	$L(\beta, Y)$	β_{11}	β_{21}	$L(\beta, Y)$
-2	-2	-1282,80	-1	-1	-1109,05	-0,5	-1	-1069,70
-2	-1	-1248,93	-1	-0,5	-1115,53	-0,5	-0,75	-1065,86
-2	0	-1312,23	-1	0	-1149,00	-0,5	-0,5	-1067,82
-2	1	-1491,47	-1	0,5	-1211,75	-0,5	-0,25	-1076,18
-2	2	-1764,64	-1	1	-1303,05	-0,5	0	-1091,42
-1	-2	-1157,86	-0,5	-1	-1069,70	-0,25	-1	-1059,81
-1	-1	-1109,05	-0,5	-0,5	-1067,82	-0,25	-0,75	-1053,67
-1	0	-1149,00	-0,5	0	-1091,42	-0,25	-0,5	-1053,04
-1	1	-1303,05	-0,5	0,5	-1143,66	-0,25	-0,25	-1058,56
-1	2	-1556,85	-0,5	1	-1224,77	-0,25	0	-1070,75
0	-2	-1129,86	0	-1	-1057,05	0	-1	-1057,05
0	-1	-1057,05	0	-0,5	-1045,03	0	-0,75	-1048,45
0	0	-1056,26	0	0	-1056,26	0	-0,5	-1045,03
0	1	-1162,35	0	0,5	-1094,84	0	-0,25	-1047,44
0	2	-1376,50	0	1	-1162,35	0	0	-1056,26
1	-2	-1218,88	0,5	-1	-1073,86	0,25	-1	-1061,70
1	-1	-1119,98	0,5	-0,5	-1050,72	0,25	-0,75	-1050,55
1	0	-1068,42	0,5	0	-1047,77	0,25	-0,5	-1044,18
1	1	-1103,06	0,5	0,5	-1069,93	0,25	-0,25	-1043,28
1	2	-1248,76	0,5	1	-1120,31	0,25	0	-1048,47
2	-2	1402,99	1	-1	-1119,98	0,5	-1	-1073,86
2	-1	-1285,67	1	-0,5	-1085,97	0,5	-0,75	-1060,11
2	0	-1192,71	1	0	-1068,42	0,5	-0,5	-1050,72
2	1	-1153,38	1	0,5	-1072,62	0,5	-0,25	-1046,39
2	2	-1206,47	1	1	-1103,06	0,5	0	-1047,77

Tabulka 2.13: Hodnoty logaritmické věrohodnosti pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na parametrech β_{11} a β_{21}

β_1	β_2	$L(\beta, Y)$	β_1	β_2	$L(\beta, Y)$	β_1	β_2	$L(\beta, Y)$
-2	-2	-53475,38	0	-2	-30328,44	0	-1	-26178,34
-2	-1	-42668,48	0	-1,5	-27612,58	0	-0,75	-25854,82
-2	0	-36881,45	0	-1	-26178,34	0	-0,5	-25745,86
-2	1	-35273,88	0	-0,5	-25745,86	0	-0,25	-25817,78
-2	2	-36344,67	0	0	-26040,02	0	0	-26040,02
-1	-2	-38423,03	0,5	-2	-28369,67	0,25	-1	-25747,91
-1	-1	-30943,22	0,5	-1,5	-26389,11	0,25	-0,75	-25549,14
-1	0	-28485,90	0,5	-1	-25534,91	0,25	-0,5	-25545,67
-1	1	-29197,33	0,5	-0,5	-25526,04	0,25	-0,25	-25705,47
-1	2	-31466,09	0,5	0	-26111,50	0,25	0	-26000,00
0	-2	-30328,44	1	-2	-27448,43	0,5	-1	-25534,91
0	-1	-26178,34	1	-1,5	-26047,91	0,5	-0,75	-25441,63
0	0	-26040,02	1	-1	-25617,32	0,5	-0,5	-25526,04
0	1	-27949,43	1	-0,5	-25899,75	0,5	-0,25	-25758,14
0	2	-30731,86	1	0	-26676,99	0,5	0	-26111,50
1	-2	-27448,43	1,5	-2	-27300,50	0,75	-1	-25502,21
1	-1	-25617,32	1,5	-1,5	-26323,60	0,75	-0,75	-25496,81
1	0	-26676,99	1,5	-1	-26184,31	0,75	-0,5	-25653,51
1	1	-29100,06	1,5	-0,5	-26658,51	0,75	-0,25	-25944,44
1	2	-32083,54	1,5	0	-27558,29	0,75	0	-26345,27
2	-2	-27697,36	2	-2	-27697,36	1	-1	-25617,32
2	-1	-27064,24	2	-1,5	-27011,75	1	-0,75	-25684,21
2	0	-28637,57	2	-1	-27064,24	1	-0,5	-25899,75
2	1	-31261,68	2	-0,5	-27660,99	1	-0,25	-26238,15
2	2	-39320,88	2	0	-28637,57	1	0	-26676,99

Tabulka 2.14: Hodnoty logaritmické věrohodnosti pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na parametrech β_1 a β_2

Parametr	Odhady		
	Newton-Raphson	Graficky	Vybrané hodnoty
β_5	0,030	0	0
β_7	0,963	1	1
β_{11}	0,220	0,22	0,25
β_{21}	-0,353	-0,35	-0,25
β_1	0,545	0,5	0,5
β_2	-0,773	-0,8	-0,75

Tabulka 2.15: Přehled odhadů parametrů pro zkoumané modely dle způsobu jejich získání

Parametr	Zvoleno	Odhad	e ^{odhad}	Směr. odchylka	p-hodnota
α_1	3,4	4,710	111,030	2,710	
α_2	2	3,010	20,280	2,759	
β_{11}	0,05	0,234	1,264	0,130	0
β_{12}	-0,3	-1,361	0,256	1,126	0
β_{13}	0,1	-0,035	0,966	0,163	0
β_{14}	0	0,480	1,617	0,930	0
β_{21}	0,4	0,570	1,768	0,133	0
β_{22}	0,2	-0,707	0,493	1,140	0
β_{23}	-0,2	-0,312	0,732	0,167	0
β_{24}	0	0,593	1,809	0,946	0

Tabulka 2.16: Odhady parametrů a další charakteristiky pro model multinomické regrese s nominální odezvou pro námi vygenerovaná data

Parametr	Zvoleno	Odhad	e ^{odhad}	Směr. odchylka	p-hodnota
α_1	-1	-0,796	0,451	0,407	
α_2	2	2,215	9,165	0,355	
α_3	5	5,187	178,929	0,394	
β_1	-0,3	-0,201	0,818	0,126	0,109
β_2	-0,2	-0,214	0,807	0,023	0
β_3	0	-0,071	0,932	0,125	0,573

Tabulka 2.17: Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro námi vygenerovaná data

3. Asymptotické testy

Tato kapitola se zabývá asymptotickými testy v kontextu logistické a multinomické regrese. Použijeme v ní obecné postupy pro konstrukci Waldova testu a testu poměrem věrohodností.

3.1 Test poměrem věrohodností pro logistickou regresi

Úkolem této podkapitoly je odvodit testovou statistiku testu poměrem věrohodností. Cílem nebude vyjádření testové statistiky jen na základě odezvy a regresorů, nýbrž ve vyjádření budou figurovat také maximálně věrohodné odhady parametrů, neboť tyto parametry získáváme iterativní metodou a jejich vyjádření jako funkce odezvy a regresorů není možné.

Jak bylo popsáno v kapitole 1.1, test poměrem věrohodností testuje nulovou hypotézu $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ pro zvolené $\boldsymbol{\theta}^0 \in \Theta$. V případě logistické regrese chceme testovat nulovost jen jednoho parametru. Nechť se jedná například o j -tý parametr, tedy budeme testovat hypotézu $\beta_j = 0$.

Vyjádříme nejprve logaritmickou věrohodnost s dosazenými maximálně věrohodnými odhady parametrů $\boldsymbol{\beta}$

$$L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) = \sum_{i=1}^n \left\{ y_i \ln \left[\frac{e^{\hat{\boldsymbol{\beta}}^T \mathbf{x}_i}}{1 + e^{\hat{\boldsymbol{\beta}}^T \mathbf{x}_i}} \right] + (1 - y_i) \ln \left[1 - \frac{e^{\hat{\boldsymbol{\beta}}^T \mathbf{x}_i}}{1 + e^{\hat{\boldsymbol{\beta}}^T \mathbf{x}_i}} \right] \right\}. \quad (3.1)$$

Nyní vyjádříme logaritmickou věrohodnost za platnosti nulové hypotézy, tedy dosadíme $\beta_j = 0$ a za ostatní parametry maximálně věrohodné odhady $\tilde{\boldsymbol{\beta}}$ spočítané pro podmodel

$$L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) = \sum_{i=1}^n \left\{ y_i \ln \left[\frac{\exp(\tilde{\beta}_1 x_i^{(1)} + \dots + \tilde{\beta}_{j-1} x_i^{(j-1)} + \tilde{\beta}_{j+1} x_i^{(j+1)} + \tilde{\beta}_p x_i^{(p)})}{1 + \exp(\tilde{\beta}_1 x_i^{(1)} + \dots + \tilde{\beta}_{j-1} x_i^{(j-1)} + \tilde{\beta}_{j+1} x_i^{(j+1)} + \tilde{\beta}_p x_i^{(p)})} \right] + (1 - y_i) \ln \left[1 - \frac{\exp(\tilde{\beta}_1 x_i^{(1)} + \dots + \tilde{\beta}_{j-1} x_i^{(j-1)} + \tilde{\beta}_{j+1} x_i^{(j+1)} + \tilde{\beta}_p x_i^{(p)})}{1 + \exp(\tilde{\beta}_1 x_i^{(1)} + \dots + \tilde{\beta}_{j-1} x_i^{(j-1)} + \tilde{\beta}_{j+1} x_i^{(j+1)} + \tilde{\beta}_p x_i^{(p)})} \right] \right\}.$$

Odtud již stačí dosadit do vzorce $LR = -2 \left\{ L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) - L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) \right\}$ a dostáváme testovou statistiku pro test poměrem věrohodností pro logistickou regresi. Tato testová statistika má za nulové hypotézy asymptoticky rozdělení χ_1^2 .

Správnost uvedené testové statistiky jsme ověřili na příkladu z podkapitoly 2.10 týkající se rakoviny prostaty, kdy jsme porovnali p -hodnoty příslušející k dané testové statistice při testování nulovosti jednotlivých parametrů s p -hodnotami, které jsme získali funkcí `lrtest` z balíku `lmtest`, a vyšly zcela totožné.

3.2 Test poměrem věrohodností pro multinomickou regresi s nominální odezvou

Tato podkapitola se zabývá opět testem poměrem věrohodností, ale tentokrát pro multinomickou regresi s nominální odezvou. Cílem bude rovněž vyjádřit testovou

statistiku pomocí odezvy, regresorů a maximálně věrohodných odhadů.

Vyjádříme výraz $-2 \left\{ L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) - L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) \right\}$ a získáme tím testovou statistiku testu poměrem věrohodností. Za vektor parametrů $\hat{\boldsymbol{\theta}}_n$ vezmeme maximálně věrohodné odhady parametrů $\hat{\beta}_{jm}, j = 1, \dots, J-1$. Budeme tak testovat m -té parametry u všech rovnic, když tedy nepočítáme konstantní členy. Dále položíme $\boldsymbol{\theta}^0 = \mathbf{0}$.

Nyní dosadíme do logaritmické věrohodnosti maximálně věrohodné odhady

$$L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) = \sum_{j=1}^{J-1} \left[\hat{\alpha}_j \left(\sum_{i=1}^n y_{ij} \right) + \sum_{k=1}^p \hat{\beta}_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{J-1} \exp \left(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \mathbf{x}_i \right) \right]. \quad (3.2)$$

Dále vyjádříme logaritmickou věrohodnost za platnosti nulové hypotézy, kde $\tilde{\alpha}_j, j = 1, \dots, J-1$ a $\tilde{\beta}_{jk}, j = 1, \dots, J-1, k = 1, \dots, p, k \neq m$ jsou maximálně věrohodné odhady v podmodelu

$$L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) = \sum_{j=1}^{J-1} \left[\tilde{\alpha}_j \left(\sum_{i=1}^n y_{ij} \right) + \sum_{k=1}^{m-1} \tilde{\beta}_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) + \sum_{k=m+1}^p \tilde{\beta}_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{J-1} \exp \left(\tilde{\alpha}_j + \sum_{k=1}^{m-1} \tilde{\beta}_{jk} x_{ik} + \sum_{k=m+1}^p \tilde{\beta}_{jk} x_{ik} \right) \right].$$

Výslednou testovou statistiku získáme dosazením do vzorce $LR = -2 \left\{ L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) - L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) \right\}$.

Tato testová statistika testu poměrem věrohodností pro multinomickou regresi s nominální odezvou má za nulové hypotézy asymptoticky rozdělení χ_{J-1}^2 , protože testujeme nulovost právě $J-1$ parametrů.

Správnost uvedené testové statistiky jsme ověřili na příkladu z podkapitoly 2.2 týkající se příslušnosti k politické straně, kdy jsme porovnali p -hodnoty příslušející k dané testové statistice při testování nulovosti jednotlivých parametrů s p -hodnotami, které jsme získali z testových statistik vypočtených jako rozdíl deviancí, a vyšly zcela totožné.

3.3 Test poměrem věrohodností pro multinomickou regresi s ordinální odezvou

K odvození testu poměrem věrohodností pro multinomickou regresi s ordinální odezvou opět využijeme obecný postup pro konstrukci tohoto testu. Protože budeme testovat nulovost pouze jednoho parametru, budeme uvažovat $\hat{\theta}_n = \hat{\beta}_l$ a $\theta_0 = 0$. Testovat tedy budeme l -tý parametr, když nezapočítáváme konstantní člen.

Vyjádříme logaritmicou věrohodnost, do které dosadíme maximálně věrohodné odhady

$$\begin{aligned} L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) &= \log \prod_{i=1}^n \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] = \sum_{i=1}^n \sum_{j=1}^J \log \pi_j(\mathbf{x}_i)^{y_{ij}} = \\ &= \sum_{i=1}^n \sum_{j=1}^J \log \left(\frac{\exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)} - \frac{\exp(\hat{\alpha}_{j-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\hat{\alpha}_{j-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)} \right)^{y_{ij}}. \end{aligned}$$

Dále vyjádříme logaritmicou věrohodnost za platnosti nulové hypotézy

$$\begin{aligned} L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) &= \sum_{i=1}^n \sum_{j=1}^J y_{ij} \log \left(\frac{\exp(\tilde{\alpha}_j + \sum_{k=1}^{l-1} \tilde{\beta}_k x_{ik} + \sum_{k=l+1}^p \tilde{\beta}_k x_{ik})}{1 + \exp(\tilde{\alpha}_j + \sum_{k=1}^{l-1} \tilde{\beta}_k x_{ik} + \sum_{k=l+1}^p \tilde{\beta}_k x_{ik})} - \right. \\ &\quad \left. - \frac{\exp(\tilde{\alpha}_{j-1} + \sum_{k=1}^{l-1} \tilde{\beta}_k x_{ik} + \sum_{k=l+1}^p \tilde{\beta}_k x_{ik})}{1 + \exp(\tilde{\alpha}_{j-1} + \sum_{k=1}^{l-1} \tilde{\beta}_k x_{ik} + \sum_{k=l+1}^p \tilde{\beta}_k x_{ik})} \right), \end{aligned}$$

kde $\tilde{\beta}_k$, $k = 1, \dots, p$, $k \neq l$ jsou maximálně věrohodné odhady spočítané v podmodelu.

Teď už jen stačí dosadit do již uvedeného vztahu $LR = -2 \left\{ L^{(n)}(\boldsymbol{\theta}^0, \mathbf{Y}) - L^{(n)}(\hat{\boldsymbol{\theta}}_n, \mathbf{Y}) \right\}$ a dostaneme testovou statistiku pro test poměrem věrohodností pro multinomickou regresi s ordinální odezvou. Tato testová statistika má za nulové hypotézy asymptoticky rozdělení χ_1^2 .

Správnost uvedené testové statistiky jsme ověřili na příkladu z podkapitoly 2.4 týkající se závažnosti zranění při automobilových nehodách, kdy jsme porovnali p -hodnoty příslušející k dané testové statistice při testování nulovosti jednotlivých parametrů s p -hodnotami, které jsme získali z testových statistik vypočítaných jako rozdíl deviancí, a vyšly zcela totožné.

3.4 Waldův test pro logistickou regresi

V této kapitole odvodíme Waldův test pro případ logistické regrese. Použijeme k tomu obecný postup konstrukce tohoto testu. Budeme testovat nulovost jednoho parametru, proto budeme uvažovat $\hat{\boldsymbol{\theta}}_n = \hat{\beta}_j$ a $\boldsymbol{\theta}^0 = 0$. Dále je třeba rozhodnout, co dosadíme za $F^{(n)}$. Zvolíme možnost $J^{(n)}(\hat{\boldsymbol{\theta}}_n)$. Než počítat jen jeden prvek Fisherovy informační matice, bude jednodušší v tomto případě spočítat celou Fisherovu informační matici. Požadované $J^{(n)}(\hat{\boldsymbol{\theta}}_n)$ bude pak j -tým prvkem na diagonále této matice. Fisherovu informační matici určíme podle následujícího vzorce

$$J^{(n)}(\boldsymbol{\beta}) = \mathbb{E} \frac{\partial L}{\partial \boldsymbol{\beta}} \left(\frac{\partial L}{\partial \boldsymbol{\beta}} \right)^T.$$

Tento vzorec nás vede k otázce, jak vypadá vektor derivací logaritmicke věrohodnostní funkce. Nejprve odvodíme jeho j -tý prvek

$$L(\beta_j) = \sum_{i=1}^n y_i \ln(\pi(\mathbf{x}_i)) + (1 - y_i) \ln(1 - \pi(\mathbf{x}_i)),$$

$$\begin{aligned}
1 - \pi(\mathbf{x}_i) &= \frac{1}{1 + e^{\beta^T \mathbf{x}_i}}, \\
\frac{\partial \pi(\mathbf{x}_i)}{\partial \beta_j} &= \frac{x_i^{(j)} e^{\beta^T \mathbf{x}_i} (1 + e^{\beta^T \mathbf{x}_i}) - e^{\beta^T \mathbf{x}_i} x_i^{(j)}}{(1 + e^{\beta^T \mathbf{x}_i})^2} = \\
&= \frac{x_i^{(j)} e^{\beta^T \mathbf{x}_i}}{(1 + e^{\beta^T \mathbf{x}_i})^2} = x_i^{(j)} \pi(\mathbf{x}_i) (1 - \pi(\mathbf{x}_i)).
\end{aligned}$$

Po přípravných výpočtech můžeme nyní přistoupit k samotnému odvození derivace logaritmické věrohodnostní funkce podle parametru β_j

$$\begin{aligned}
\frac{\partial L}{\partial \beta_j} &= \sum_{i=1}^n \left[y_i \frac{1}{\pi(\mathbf{x}_i)} x_i^{(j)} \pi(\mathbf{x}_i) (1 - \pi(\mathbf{x}_i)) - (1 - y_i) \frac{1}{1 - \pi(\mathbf{x}_i)} x_i^{(j)} \pi(\mathbf{x}_i) (1 - \pi(\mathbf{x}_i)) \right] = \\
&= \sum_{i=1}^n x_i^{(j)} [y_i (1 - \pi(\mathbf{x}_i)) - (1 - y_i) \pi(\mathbf{x}_i)] = \sum_{i=1}^n x_i^{(j)} (y_i - \pi(\mathbf{x}_i)).
\end{aligned}$$

Takže pro vektor derivací logaritmické věrohodnostní funkce platí

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n (y_i - \pi(\mathbf{x}_i)) \mathbf{x}_i.$$

Tento vztah lze také zapsat ve tvaru

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\pi}(\mathbf{X})),$$

kde matice \mathbf{X} má po řádcích vektory $(x_1^{(1)}, \dots, x_1^{(p)})$ až $(x_n^{(1)}, \dots, x_n^{(p)})$. Pro vektor \mathbf{Y} platí $\mathbf{Y} = (y_1, \dots, y_n)^T$ a pro vektor $\boldsymbol{\pi}(\mathbf{X})$ platí $\boldsymbol{\pi}(\mathbf{X}) = (\pi(\mathbf{x}_1), \dots, \pi(\mathbf{x}_n))^T$.

Nyní dosadíme do vzorce pro výpočet Fisherovy informační matice a využijeme toho, že platí $\mathbf{E}\mathbf{Y} = \boldsymbol{\pi}(\mathbf{X})$

$$\begin{aligned}
J^{(n)}(\boldsymbol{\beta}) &= \mathbf{E} \frac{\partial L}{\partial \boldsymbol{\beta}} \left(\frac{\partial L}{\partial \boldsymbol{\beta}} \right)^T = \mathbf{E} \mathbf{X}^T (\mathbf{Y} - \mathbf{E}\mathbf{Y}) (\mathbf{Y} - \mathbf{E}\mathbf{Y})^T \mathbf{X} = \\
&= \mathbf{X}^T \mathbf{E} (\mathbf{Y} - \mathbf{E}\mathbf{Y}) (\mathbf{Y} - \mathbf{E}\mathbf{Y})^T \mathbf{X}.
\end{aligned}$$

Matice $\mathbf{E}(\mathbf{Y} - \mathbf{E}\mathbf{Y})(\mathbf{Y} - \mathbf{E}\mathbf{Y})^T$ je matice rozptylů jednotlivých pozorování. Každé pozorování má alternativní rozdělení s pravděpodobností úspěchu $\pi(\mathbf{x}_i)$. Jednotlivá pozorování jsou vzájemně nezávislá, takže rozptylová matice je diagonální a vypadá následovně

$$\mathbf{R}(\boldsymbol{\beta}) = \text{diag} \{ \pi(\mathbf{x}_1)(1 - \pi(\mathbf{x}_1)), \dots, \pi(\mathbf{x}_n)(1 - \pi(\mathbf{x}_n)) \}.$$

Fisherovu matici tedy můžeme zapsat ve tvaru

$$J^{(n)}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{R}(\boldsymbol{\beta}) \mathbf{X} = \sum_{i=1}^n \pi(\mathbf{x}_i) (1 - \pi(\mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^T.$$

[5]

Požadovaný j -tý prvek na diagonále Fisherovy informační matice je našem případě tedy

$$J^{(n)}(\hat{\beta}_j) = \sum_{i=1}^n \hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i))(x_i^{(j)})^2,$$

kde $\hat{\pi}(\mathbf{x}_i) = \frac{e^{\hat{\beta}^T \mathbf{x}_i}}{1 + e^{\hat{\beta}^T \mathbf{x}_i}}$, neboť je třeba za parametry dosadit jejich maximálně věrohodné odhady.

Nyní můžeme dosadit do vzorce pro Waldův test a vychází nám

$$W = \hat{\beta}_j^2 \sum_{i=1}^n (x_i^{(j)})^2 \hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i)).$$

Tato testová statistika je však pro případ logistické regrese příliš zjednodušující a je třeba uvažovat ještě tzv. rušivé parametry. Mějme tedy obecně p parametrů $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$, kde $p \geq 2$. Nechť $1 \leq k \leq p$ a označme $\boldsymbol{\tau} = (\theta_1, \dots, \theta_k)^T$ a $\boldsymbol{\psi} = (\theta_{k+1}, \dots, \theta_p)^T$. Uvažujme nyní Fisherovu informační matici ve tvaru

$$J(\boldsymbol{\theta}) = \begin{pmatrix} J_{11}(\boldsymbol{\theta}) & J_{12}(\boldsymbol{\theta}) \\ J_{21}(\boldsymbol{\theta}) & J_{22}(\boldsymbol{\theta}) \end{pmatrix},$$

kde $J_{11}(\boldsymbol{\theta})$ je matice typu $k \times k$. Dále uvažujme matici danou vzorcem

$$J_{11.2}(\boldsymbol{\theta}) = J_{11}(\boldsymbol{\theta}) - J_{12}(\boldsymbol{\theta}) [J_{22}(\boldsymbol{\theta})]^{-1} J_{21}(\boldsymbol{\theta}) \quad (3.3)$$

Testovou statistiku Waldova testu při zahrnutí rušivých parametrů a testování nulové hypotézy $H_0^* : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ představuje následující vztah, jež je mírnou modifikací vzorce uvedeného v [4, str. 183]

$$W^* = (\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0)^T J_{11.2}(\boldsymbol{\theta})(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0).$$

Tato testová statistika má za platnosti nulové hypotézy asymptoticky rozdělení χ_k^2 .

Správnost uvedené testové statistiky jsme ověřili na příkladu z podkapitoly 2.10 týkající se rakoviny prostaty, kdy jsme porovnali p -hodnoty příslušející k dané testové statistice při testování nulovosti jednotlivých parametrů s p -hodnotami, které jsme získali zabudovanou funkcí `summary`, a vyšly zcela totožné.

3.5 Waldův test pro multinomickou regresi s nominální odezvou

Cílem této kapitoly je odvození Waldova testu pro multinomickou regresi s nominální odezvou. K tomuto odvození využijeme obecný postup konstrukce tohoto testu. Musíme opět zvolit, co budeme uvažovat jako $F^{(n)}$. Stejně jako v případě logistické regrese budeme chtít dosadit $J^{(n)}(\hat{\boldsymbol{\theta}}_n)$, tedy Fisherovu informační matici. Jednotlivé prvky Fisherovy informační matice vypočítáme pomocí tohoto vzorce, který je uveden v knize [5], jedná se o prvek na l -tém řádku a v h -tém sloupci Fisherovy informační matice

$$J_{lh}^{(n)}(\boldsymbol{\theta}) = -E \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_l \partial \theta_h}.$$

Logaritmická věrohodnostní funkce, ze které budeme vycházet, je následující

$$\sum_{j=1}^{J-1} \left[\alpha_j \left(\sum_{i=1}^n y_{ij} \right) + \sum_{k=1}^p \beta_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i) \right].$$

Nejdříve vypočítáme první parciální derivaci podle parametru β_{lm} , což je m -tý parametr v l -té rovnici

$$\frac{\partial L}{\partial \beta_{lm}} = \beta_{lm} \left(\sum_{i=1}^n x_{im} y_{il} \right) - \sum_{i=1}^n \frac{\exp(\alpha_l + \beta_l^T \mathbf{x}_i) x_{im}}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i)}.$$

Nyní odvodíme druhou parciální derivaci logaritmické věrohodnostní funkce

$$\frac{\partial^2 L}{\partial \beta_{lm} \partial \beta_{hm}} = - \sum_{i=1}^n x_{im}^2 \left\{ \frac{\exp(\alpha_l + \beta_l^T \mathbf{x}_i)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i)} - \frac{\exp(\alpha_l + \beta_l^T \mathbf{x}_i) \exp(\alpha_h + \beta_h^T \mathbf{x}_i)}{\left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i) \right]^2} \right\}$$

Teď je zapotřebí spočítat střední hodnotu této druhé parciální derivace a změnit znaménko na opačné. Tímto dostaneme prvek ve Fisherově informační matici. Střední hodnota je zřejmá. Nenachází se tam žádné y_{ij} , tudíž se jedná o střední hodnotu z konstanty a střední hodnota konstanty je sama konstanta. Prvek na l -tém řádku a v h -tém sloupci Fisherovy informační matice v případě, že $l \neq h$, je tedy

$$J_{lh}(\boldsymbol{\theta}) = - \sum_{i=1}^n x_{im}^2 \frac{\exp(\alpha_l + \beta_l^T \mathbf{x}_i) \exp(\alpha_h + \beta_h^T \mathbf{x}_i)}{\left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i) \right]^2}$$

Na diagonále Fisherovy informační matice je l -tý prvek dán vzorcem

$$J_{ll}(\boldsymbol{\theta}) = \sum_{i=1}^n x_{im}^2 \left\{ \frac{\exp(\alpha_l + \beta_l^T \mathbf{x}_i)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i)} - \frac{\exp(\alpha_l + \beta_l^T \mathbf{x}_i) \exp(\alpha_h + \beta_h^T \mathbf{x}_i)}{\left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x}_i) \right]^2} \right\} \quad (3.4)$$

Odtud již dosadíme do vztahu pro Waldův test

$$W = \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 \right)^T J^{(n)}(\hat{\boldsymbol{\theta}}_n) \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 \right), \quad (3.5)$$

kde budeme uvažovat, že $\hat{\boldsymbol{\theta}}_n$ je $\hat{\beta}_{jm}$, $j = 1, \dots, J-1$ a $\boldsymbol{\theta}^0 = \mathbf{0}$, protože testujeme nulovost parametrů.

Tato testová statistika má za nulové hypotézy asymptoticky rozdělení χ_{J-1}^2 , protože testujeme nulovost právě $J-1$ parametrů.

Uvedená testová statistika je pro náš obecný případ multinomické regrese s nominální odezvou příliš zjednodušující a stejně jako v případě logistické regrese

budeme muset ještě uvažovat vliv rušivých parametrů. Takto uvažovaná statistika je dána vzorcem

$$W^* = (\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0)^T J_{11.2}(\boldsymbol{\theta})(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0),$$

kde $J_{11.2}(\boldsymbol{\theta})$ představuje vzorec (3.3), který zahrnuje všechny parametry modelu a k jehož spočítání je třeba znát rozšířenou Fisherovu informační matici oproti Fisherově informační matici, s kterou jsme počítali v původní testové statistice. Prvky této matice jsou dány následujícím dvěma vzorci. V případě, že $l \neq h$, kde l a h představují rovnici, která obsahuje daný parametr, se jedná o prvky

$$J_{(J-1)(m-1)+l,(J-1)(r-1)+h}(\boldsymbol{\theta}) = - \sum_{i=1}^n x_{im}x_{ir} \frac{\exp(\boldsymbol{\beta}_l^T \mathbf{x}_i) \exp(\boldsymbol{\beta}_h^T \mathbf{x}_i)}{\left[1 + \sum_{j=1}^{J-1} \exp(\boldsymbol{\beta}_j^T \mathbf{x}_i)\right]^2},$$

kde m a r udávají pořadí parametrů v rovnici. Poznamenejme, že parametr α_l , resp. α_h uvažujeme nyní jako jeden z parametrů vektoru $\boldsymbol{\beta}_l$, resp. $\boldsymbol{\beta}_h$, který nyní obsahuje $p + 1$ parametrů.

Zbývající prvky Fisherovy informační matice jsou dány vzorcem

$$J_{(J-1)(m-1)+l,(J-1)(r-1)+l}(\boldsymbol{\theta}) = \sum_{i=1}^n x_{im}x_{ir} \left\{ \frac{\exp(\boldsymbol{\beta}_l^T \mathbf{x}_i)}{1 + \sum_{j=1}^{J-1} \exp(\boldsymbol{\beta}_j^T \mathbf{x}_i)} - \frac{\exp(\boldsymbol{\beta}_l^T \mathbf{x}_i) \exp(\boldsymbol{\beta}_h^T \mathbf{x}_i)}{\left[1 + \sum_{j=1}^{J-1} \exp(\boldsymbol{\beta}_j^T \mathbf{x}_i)\right]^2} \right\}$$

Správnost takto spočítané testové statistiky jsme ověřili na příkladu z podkapitoly 2.2 týkající se příslušnosti k politické straně, kdy jsme porovnali p -hodnoty příslušející k dané testové statistice při testování nulovosti jednotlivých parametrů s p -hodnotami, které jsme získali z testových statistik vypočtených funkcí `wald.test` z balíku `aod`, a vyšly zcela totožné.

3.6 Waldův test pro multinomickou regresi s ordinální odezvou

Waldův test pro multinomickou regresi s ordinální odezvou odvodíme z obecného postupu. Za $F^{(n)}$ zvolíme opět $J^{(n)}(\hat{\boldsymbol{\theta}}_n)$. Bude tedy třeba spočítat k -tý prvek na diagonále Fisherovy informační matice. Opět využijeme vzorce uvedeného v [5], který bude v našem případě

$$J_{kk}^{(n)}(\boldsymbol{\theta}) = -\mathbb{E} \frac{\partial^2 L(\boldsymbol{\theta})}{\partial^2 \theta_k}.$$

Spočítejme tedy druhou parciální derivaci logaritmické věrohodnostní funkce. Věrohodnostní funkce pro multinomickou regresi s ordinální odezvou je následující

$$\prod_{i=1}^n \left\{ \left(\frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right)^{y_{i1}} \left[\prod_{j=2}^{J-1} \left(\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} \right)^{y_{ij}} \right] \left(1 - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right)^{y_{iJ}} \right\}.$$

Z toho odvodíme logaritmickou věrohodnostní funkci, která je

$$\sum_{i=1}^n \left[y_{i1} \log \left(\frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) + \sum_{j=2}^{J-1} y_{ij} \log \left(\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) + y_{iJ} \log \left(1 - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \right].$$

Nejprve spočítáme první parciální derivaci logaritmické věrohodnostní funkce

$$\begin{aligned} \frac{\partial L}{\partial \beta_k} &= \sum_{i=1}^n \left\{ y_{i1} \frac{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \cdot \frac{x_{ik} \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)] - x_{ik} \exp[2(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} + \right. \\ &+ \sum_{j=2}^{J-1} y_{ij} \frac{1}{\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}} \cdot \\ &\cdot \left\{ \frac{x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} - \frac{x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} + \\ &\left. + y_{iJ} [1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] \cdot \left\{ - \frac{x_{ik} \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} \right\} = \end{aligned}$$

Nyní postupně upravíme první parciální derivaci

$$\begin{aligned} &= \sum_{i=1}^n \left\{ y_{i1} \frac{x_{ik}}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} + \right. \\ &+ \sum_{j=2}^{J-1} y_{ij} \frac{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)][1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]}{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \cdot \\ &\cdot \left\{ \frac{x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2 [1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} - \right. \\ &\left. - \frac{x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2 [1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} - \\ &\left. - y_{iJ} \frac{x_{ik} \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right\}. \end{aligned}$$

V součtu přes j zkrátíme v prvním zlomku čitatel s druhými mocninami ve jmenovatelích druhých dvou zlomků. Následně upravíme čitatele těchto druhých dvou zlomků. Vyjde nám

$$\begin{aligned} &x_{ik} \{ \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) + 2 \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \\ &+ \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\ &2 \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] \} = \\ &= x_{ik} \{ [\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] - \\ &- \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) [\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] \} = \\ &= x_{ik} [\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] \cdot \\ &\cdot [1 - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]. \end{aligned}$$

Zkrátíme první závorku se jmenovatelem prvního zlomku v součtu přes j a dostáváme tak první derivaci, která vypadá následovně

$$\begin{aligned} & \sum_{i=1}^n x_{ik} \left\{ y_{i1} \frac{1}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} + \right. \\ & + \sum_{j=2}^{J-1} y_{ij} \frac{1 - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)][1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]} + \\ & \left. + y_{iJ} \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right\}. \end{aligned}$$

Nyní přistoupíme k výpočtu druhé parciální derivace logaritmické věrohodnostní funkce. Zderivujeme-li první člen první derivace, dostaneme

$$-y_{i1} x_{ik}^2 \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^2}.$$

Derivace druhého členu bude derivací podílu. Když zderivujeme funkci v čitateli dostaneme

$$\begin{aligned} & - [\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \\ & x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] = \\ & = -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i). \end{aligned}$$

Zderivujeme-li funkci ve jmenovateli, máme

$$\begin{aligned} & [1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \\ & + x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] = \\ & = x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) + x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \\ & + 2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i). \end{aligned}$$

Protože se jedná o derivaci podílu, tak jmenovatel výsledného výrazu bude

$$[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2 [1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2.$$

Čítenel tohoto výrazu bude tvořen následující funkcí

$$\begin{aligned} & -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) \cdot \\ & \cdot [1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)][1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] - \\ & - [1 - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] [x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) + \\ & + x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + 2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]. \end{aligned}$$

Tuto funkci roznásobíme a dostaneme

$$\begin{aligned} & -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\ & -2x_{ik} \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\ & -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] - \\ & -2x_{ik} \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] - \end{aligned}$$

$$\begin{aligned}
& -x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\
& -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \\
& +x_{ik} \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \\
& +x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] + \\
& +2x_{ik} \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)].
\end{aligned}$$

Můžeme funkci zjednodušit

$$\begin{aligned}
& -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\
& -x_{ik} \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)] \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\
& -x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)] - \\
& -x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - x_{ik} \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \\
& -2x_{ik} \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i).
\end{aligned}$$

Nyní tuto funkci ještě upravíme

$$\begin{aligned}
& -x_{ik} \left\{ \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + 2 \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) + \exp[2(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]] + \right. \\
& + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + 2 \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) + \exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]] \left. \right\} = \\
& = -x_{ik} \left\{ \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2 + \right. \\
& + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2 \left. \right\}.
\end{aligned}$$

Druhý člen druhé parciální derivace logaritmické věrohodnostní funkce tedy je

$$-\sum_{j=2}^{J-1} y_{ij} x_{ik}^2 \left\{ \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} + \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\}.$$

Derivace posledního členu první parciální derivace logaritmické věrohodnostní funkce je

$$\begin{aligned}
& -y_{iJ} x_{ik} \left\{ \frac{x_{ik} \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) [1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} - \right. \\
& \left. - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) x_{ik} \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} = \\
& = -y_{iJ} x_{ik}^2 \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2}.
\end{aligned}$$

Uveďme nyní celkovou druhou parciální derivaci logaritmické věrohodnostní funkce

$$\begin{aligned}
& -\sum_{i=1}^n \left\{ y_{i1} x_{ik}^2 \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} + \right. \\
& + \sum_{j=2}^{J-1} y_{ij} x_{ik}^2 \left\{ \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} + \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} + \\
& \left. + y_{iJ} x_{ik}^2 \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\}.
\end{aligned}$$

K tomu, abychom získali k -tý prvek na diagonále Fisherovy informační matice, musíme počítat střední hodnotu této druhé parciální derivace a změnit znaménko na opačné. Střední hodnotu v tomto případě získáme tak, že dosadíme za příslušné y_{ij} pravděpodobnost příslušné kategorie a tím skončíme, neboť druhý člen, do něhož bychom dosadili nulu a přenásobili jej výrazem $1 - p$, kde p je příslušná pravděpodobnost, vypadne. Výsledný k -tý prvek na diagonále Fisherovy informační matice

$$\begin{aligned} & \sum_{i=1}^n x_{ik}^2 \left\{ \frac{\exp[2(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^3} + \right. \\ & + \sum_{j=2}^{J-1} \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right] \cdot \\ & \cdot \left\{ \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} + \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} + \\ & \left. + \left(1 - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\}. \end{aligned}$$

Odtud již lze dosadit do vzorce pro Waldův test, kde využijeme toho, že $\hat{\theta}_n = \hat{\beta}_k$ a $\theta^0 = 0$, a dostaneme

$$\begin{aligned} W &= \hat{\beta}_k^2 \sum_{i=1}^n x_{ik}^2 \left\{ \frac{\exp[2(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^3} + \right. \\ & + \sum_{j=2}^{J-1} \left[\frac{\exp(\alpha_j + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)} \right] \cdot \\ & \cdot \left\{ \frac{\exp(\alpha_j + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^2} + \frac{\exp(\alpha_{j-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{j-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^2} \right\} + \\ & \left. + \left(1 - \frac{\exp(\alpha_{J-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)} \right) \frac{\exp(\alpha_{J-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^2} \right\}. \quad (3.6) \end{aligned}$$

Výše uvedená testová statistika má za nulové hypotézy asymptoticky rozdělení χ_1^2 , protože testujeme nulovost právě jednoho parametru.

Stejně jako v případě logistické regrese a multinomické regrese s nominální odezvou bude třeba vzít do úvahy ještě rušivé parametry. Za tímto účelem spočítáme rozšířenou Fisherovu informační matici, která bude odpovídat všem parametrům obsažených v modelu. Prvky této matice jsou dány několika vzorci, které si postupně uvedeme

$$\begin{aligned} J_{11}(\boldsymbol{\theta}) &= -E \frac{\partial^2 L}{\partial^2 \alpha_1} = \sum_{i=1}^n \left\{ \frac{\exp[2(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^3} + \right. \\ & + \left(\frac{\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \cdot \\ & \cdot (1 + \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)) \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i) \cdot \end{aligned}$$

$$\begin{aligned}
& \cdot \left[\frac{1}{(1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))(\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))} - \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)(\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i) - 2\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i) - 1)}{(1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))^2(\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right] \Bigg\}, \\
J_{12}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \alpha_1 \partial \alpha_2} = -\sum_{i=1}^n \left\{ \left(\frac{\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \cdot \right. \\
& \left. \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)}{(\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right\}.
\end{aligned}$$

Dále pro prvky Fisherovy informační matice platí

$$J_{13}(\boldsymbol{\theta}) = J_{14}(\boldsymbol{\theta}) = \dots = J_{1,J-1}(\boldsymbol{\theta}) = 0.$$

Dalšími prvky jsou pro $1 < j < J - 1$

$$\begin{aligned}
J_{jj}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \alpha_j^2} = \sum_{i=1}^n \left\{ \frac{2 \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))^2} - \right. \\
& \frac{\exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)](1 + 2 \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i))}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))^3(\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i))} + \\
& \left. + \frac{\exp[2(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)](1 + 2 \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i))}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))^3(\exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))} \right\}.
\end{aligned}$$

Pro $1 < j < J - 2$ dostáváme následující prvky Fisherovy informační matice

$$\begin{aligned}
J_{j,j+1}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \alpha_j \partial \alpha_{j+1}} = -\sum_{i=1}^n \left\{ \left(\frac{\exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i)} - \right. \right. \\
& \left. \left. - \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \cdot \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i) \cdot \right. \\
& \left. \cdot \left[\frac{1}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))(\exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))} - \right. \right. \\
& \left. \left. - \frac{1 + \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))(\exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right] \right\}
\end{aligned}$$

Pro další prvky platí

$$\begin{aligned}
J_{J-1,J-1}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \alpha_{J-1}^2} = \sum_{i=1}^n \left\{ \left(\frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} - \right. \right. \\
& \left. \left. - \frac{\exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \cdot \right. \\
& \left. \cdot \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)(1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)) \cdot \right. \\
& \left. \cdot \left[\frac{1}{(1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i))(\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i))} - \right. \right. \\
& \left. \left. - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)(1 + 2 \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i))}{(1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i))^2(\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right] \right\},
\end{aligned}$$

$$\begin{aligned}
J_{J-1, J-2}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \alpha_{J-1} \partial \alpha_{J-2}} = -\sum_{i=1}^n \left\{ \left(\frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} - \right. \right. \\
&\quad \left. \left. - \frac{\exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \cdot \right. \\
&\quad \left. \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right. \\
&\quad \cdot \left[\frac{1}{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)} + \right. \\
&\quad \left. \left. + \frac{1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) - \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right] \right\}.
\end{aligned}$$

Další prvky Fisherovy informační matice vystihuje vzorec

$$\begin{aligned}
J_{J-1+k, J-1+m}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \beta_k \partial \beta_m} = \sum_{i=1}^n x_{ik} x_{im} \left\{ \frac{\exp[2(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^3} + \right. \\
&\quad \left. + \sum_{j=2}^{J-1} \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right] \cdot \right. \\
&\quad \cdot \left\{ \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} + \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\} + \\
&\quad \left. + \left(1 - \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right) \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)]^2} \right\}.
\end{aligned}$$

Pro prvky, které byly získány derivací podle nějaké parametru α a nějakého parametru β , platí následující tři vztahy

$$\begin{aligned}
J_{1, J-1+k}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \beta_k \partial \alpha_1} = \sum_{i=1}^n x_{ik} \left\{ \frac{\exp[2(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)]^3} + \right. \\
&\quad \left. + \left[\frac{\exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i)} \right] \cdot \right. \\
&\quad \cdot \left[\frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i))} + \right. \\
&\quad \left. \left. + \frac{(1 - \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))}{(1 + \exp(\alpha_1 + \boldsymbol{\beta}^T \mathbf{x}_i))^2 (1 + \exp(\alpha_2 + \boldsymbol{\beta}^T \mathbf{x}_i))} \right] \right\},
\end{aligned}$$

$$\begin{aligned}
J_{j, J-1+k}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \beta_k \partial \alpha_j} = \sum_{i=1}^n x_{ik} \left\{ \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \right. \right. \\
&\quad \left. \left. - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right] \cdot \right. \\
&\quad \cdot \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i))} + \right.
\end{aligned}$$

$$\begin{aligned}
& + \frac{(1 - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)) \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \Big] + \\
& + \left[\frac{\exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} \right] \cdot \\
& \cdot \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i))} + \right. \\
& \left. + \frac{(1 - \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)) \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right] \Big\},
\end{aligned}$$

$$\begin{aligned}
J_{J-1, J-1+k}(\boldsymbol{\theta}) &= -\mathbb{E} \frac{\partial^2 L}{\partial \beta_k \partial \alpha_{J-1}} = \sum_{i=1}^n x_{ik} \left\{ \left[\frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} - \right. \right. \\
& \left. \left. - \frac{\exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right] \cdot \right. \\
& \cdot \left[\frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i))} + \right. \\
& \left. + \frac{(1 - \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)) \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i) \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_{J-2} + \boldsymbol{\beta}^T \mathbf{x}_i))(1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i))^2} \right] + \\
& \left. + \frac{\exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{(1 + \exp(\alpha_{J-1} + \boldsymbol{\beta}^T \mathbf{x}_i))^3} \right\}.
\end{aligned}$$

Dále ze spojitosti logaritmické věrohodnostní funkce platí

$$J_{a,b}(\boldsymbol{\theta}) = J_{b,a}(\boldsymbol{\theta}).$$

Zbylé prvky Fisherovy informační matice jsou rovny nule.

Testová statistika, která zahrnuje rušivé parametry, je dána vzorcem

$$W^* = (\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0)^T J_{11.2}(\boldsymbol{\theta}) (\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0),$$

kde $J_{11.2}(\boldsymbol{\theta})$ představuje vzorec (3.3), který zahrnuje všechny parametry modelu.

Správnost takto spočítané testové statistiky jsme ověřili na příkladu z podkapitoly 2.4 týkající se závažnosti zranění při automobilových nehodách, kdy jsme porovnali p -hodnotu příslušející k dané testové statistice při testování nulovosti parametru s p -hodnotou, kterou jsme získali z testové statistiky vypočtené funkcí `wald.test` z balíku `aod`, a vyšla totožná.

3.7 Ekvivalence testů pro dvě kategorie

Cílem této podkapitoly je dokázat, že pokud máme binární odezvu, tj. odezvu o dvou kategoriích, pak jsou ekvivalentní všechny uvedené testy poměru věrohodností a rovněž jsou ekvivalentní všechny uvedené Waldovy testy. Je zřejmé, že pokud $J = 2$, pak testové statistiky mají za nulové hypotézy všechny asymptoticky rozdělení χ_1^2 . Zbývá však dokázat, že se rovnají jejich statistiky.

Začneme testy poměrem věrohodností a zformulujeme si první větu.

Věta 14. *Statistika testu poměrem věrohodností v modelu logistické regrese je přesně rovna statistice testu poměrem věrohodností v modelu multinomické regrese s nominální odezvou v případě binární odezvy.*

Důkaz. Úkolem v tomto případě bude dokázat rovnost logaritmických věrohodností, neboť pak už je zřejmé, že se budou rovnat i testové statistiky. Vyjádřeme nejdříve logaritmickou věrohodnost pro multinomickou regresi s nominální odezvou v případě dvou kategorií, tj. $J = 2$ dosadíme do vzorce (3.2) a dostaneme

$$\begin{aligned} & \left[\hat{\alpha}_1 \left(\sum_{i=1}^n y_{i1} \right) + \sum_{k=1}^p \hat{\beta}_{1k} \left(\sum_{i=1}^n x_{ik} y_{i1} \right) \right] - \\ & - \sum_{i=1}^n \log \left[1 + \exp \left(\hat{\alpha}_1 + \hat{\beta}_1^T \mathbf{x}_i \right) \right]. \end{aligned} \quad (3.7)$$

Nyní se pokusíme upravit logaritmickou věrohodnost pro logistickou regresi (vzorec (3.1)) na tvar (3.7)

$$\begin{aligned} & \sum_{i=1}^n \left\{ y_i \hat{\beta}^T \mathbf{x}_i - y_i \ln(1 + e^{\hat{\beta}^T \mathbf{x}_i}) + (1 - y_i) \cdot \left[\ln 1 - \ln(1 + e^{\hat{\beta}^T \mathbf{x}_i}) \right] \right\} = \\ & = \sum_{i=1}^n \left\{ y_i \hat{\beta}^T \mathbf{x}_i - y_i \ln(1 + e^{\hat{\beta}^T \mathbf{x}_i}) - \ln(1 + e^{\hat{\beta}^T \mathbf{x}_i}) + y_i \ln(1 + e^{\hat{\beta}^T \mathbf{x}_i}) \right\} = \\ & = \sum_{i=1}^n \left[y_i \hat{\beta}^T \mathbf{x}_i - \ln(1 + e^{\hat{\beta}^T \mathbf{x}_i}) \right] \end{aligned} \quad (3.8)$$

Odtud vidíme, že vzorce (3.7) a (3.8) jsou až na formální odlišnosti ve značení totožné. \square

Dokažme teď rovnost výše uvedených vzorců pro logaritmickou věrohodnost s logaritmickou věrohodností pro multinomickou regresi s ordinální odezvou. Zformulujme nejprve větu.

Věta 15. *Statistika testu poměrem věrohodností v modelu logistické regrese je přesně rovna statistice testu poměrem věrohodností v modelu multinomické regrese s ordinální odezvou v případě binární odezvy.*

Důkaz. Vyjdeme z věrohodnosti, kterou reprezentuje vzorec (2.8) a využijeme toho, že $J = 2$

$$\begin{aligned} & \sum_{i=1}^n \left\{ y_{i1} \log \left(\frac{\exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)} \right) + \right. \\ & \left. + y_{i2} \log \left(1 - \frac{\exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)} \right) \right\} \end{aligned} \quad (3.9)$$

Vzorec (3.9) dále upravíme s využitím toho, že platí $y_{i2} = 1 - y_{i1}$

$$\begin{aligned} & \sum_{i=1}^n \left\{ y_{i1} (\alpha_1 + \hat{\beta}^T \mathbf{x}_i) - y_{i1} \log(1 + \exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)) - \right. \\ & \left. - \log(1 + \exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)) + y_{i1} \log(1 + \exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)) \right\} \\ & = \sum_{i=1}^n \left[y_{i1} (\alpha_1 + \hat{\beta}^T \mathbf{x}_i) - \log(1 + \exp(\alpha_1 + \hat{\beta}^T \mathbf{x}_i)) \right]. \end{aligned} \quad (3.10)$$

Odtud vidíme, že vzorce (3.8) a (3.10) jsou ekvivalentní, tudíž jsou ekvivalentní i testové statistiky testů poměru věrohodností v modelu logistické regrese a v modelu multinomické regrese s ordinální odezvou pro případ dvou kategorií. \square

Z platnosti dvou výše uvedených vět je zřejmé, že test poměrem věrohodností v případě dvou kategorií je ekvivalentní pro všechny tři studované modely.

Zaměříme se nyní na Waldovy testy. Budeme uvažovat verzi Waldova testu bez zahrnutí rušivých parametrů, neboť přítomnost inverzní matice u testu s rušivými parametry nám znemožňuje jejich porovnání. Uvedeme si a dokážeme dvě věty. Zformulujeme první z nich.

Věta 16. *Statistika Waldova testu v modelu logistické regrese je přesně rovna statistice Waldova testu v modelu multinomické regrese s nominální odezvou v případě binární odezvy.*

Důkaz. Nejprve si připomeňme vzorec pro Waldův test pro logistickou regresi při testování nulovosti m -tého parametru

$$W = \hat{\beta}_m^2 \sum_{i=1}^n (x_i^{(m)})^2 \hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i)). \quad (3.11)$$

Pokusíme se dokázat, že vzorec (3.11) je ekvivalentní testové statistice Waldova testu pro multinomickou regresi s nominální odezvou. Dosadíme do vzorce (3.5), kde matice $J^{(n)}(\hat{\theta}_n)$ bude pro $J = 2$ obsahovat jediný prvek daný vzorcem (3.4), kde $l = 1$, a upravíme jej

$$\begin{aligned} & \sum_{i=1}^n \hat{\beta}_{1m}^2 x_{im}^2 \left\{ \frac{\exp(\alpha_1 + \hat{\beta}_1^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \hat{\beta}_1^T \mathbf{x}_i)} - \frac{\exp[2(\alpha_1 + \hat{\beta}_1^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \hat{\beta}_1^T \mathbf{x}_i)]^2} \right\} = \\ & = \sum_{i=1}^n \hat{\beta}_{1m}^2 x_{im}^2 [\hat{\pi}(\mathbf{x}_i) - \hat{\pi}^2(\mathbf{x}_i)] = \\ & = \sum_{i=1}^n \hat{\beta}_{1m}^2 x_{im}^2 \hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i)), \end{aligned} \quad (3.12)$$

kde $\hat{\pi}(\mathbf{x}_i) = \frac{\exp(\alpha_1 + \hat{\beta}_1^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \hat{\beta}_1^T \mathbf{x}_i)}$.

Vidíme, že testová statistika Waldova testu pro logistickou regresi reprezentovaná vzorcem (3.11) je ekvivalentní testové statistice Waldova testu pro multinomickou regresi s nominální odezvou, kterou představuje vzorec (3.12). \square

Nyní dokážeme, že těmto testovým statistikám odpovídá i testová statistika Waldova testu pro multinomickou regresi s ordinální odezvou.

Věta 17. *Statistika Waldova testu v modelu logistické regrese je přesně rovna statistice Waldova testu v modelu multinomické regrese s ordinální odezvou v případě binární odezvy.*

Důkaz. Testová statistika modelu multinomické regrese s ordinální odezvou je vyjádřena vzorcem (3.6), do něhož dosadíme $J = 2$ a dostaneme

$$\begin{aligned}
& \hat{\beta}_m^2 \sum_{i=1}^n x_{im}^2 \left\{ \frac{\exp[2(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]}{[1 + \exp(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^3} + \right. \\
& \left. + \left(1 - \frac{\exp(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)} \right) \frac{\exp(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{[1 + \exp(\alpha_1 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^2} \right\} = \\
& = \hat{\beta}_m^2 \sum_{i=1}^n x_{im}^2 [\hat{\pi}^2(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i)) + (1 - \hat{\pi}(\mathbf{x}_i))^2 \hat{\pi}(\mathbf{x}_i)] = \\
& = \hat{\beta}_m^2 \sum_{i=1}^n x_{im}^2 [\hat{\pi}^2(\mathbf{x}_i) - \hat{\pi}^3(\mathbf{x}_i) + \hat{\pi}(\mathbf{x}_i) - 2\hat{\pi}^2(\mathbf{x}_i) + \hat{\pi}^3(\mathbf{x}_i)] = \\
& = \hat{\beta}_m^2 \sum_{i=1}^n x_{im}^2 [\hat{\pi}(\mathbf{x}_i) - \hat{\pi}^2(\mathbf{x}_i)] = \\
& = \hat{\beta}_m^2 \sum_{i=1}^n x_{im}^2 \hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i)). \tag{3.13}
\end{aligned}$$

Odtud plyne, že uvedené testové statistiky Waldova testu testující nulovost parametru, které jsou vyjádřeny vzorci (3.11) a (3.13), jsou ekvivalentní v případě, že odezva nabývá dvou kategorií. \square

Z platnosti posledních dvou uváděných vět je patrné, že Waldův test je v případě binární odezvy ekvivalentní pro všechny tři studované modely.

4. Kvalita vína

Cílem této kapitoly je ilustrace teoretických poznatků pro hledání modelů. Konkrétně budeme chtít najít vhodný model předpovídající kvalitu vína. Vhodným modelem se pro tento případ jeví model multinomické regrese s ordinální odezvou. My však zkusíme aplikovat i model multinomické regrese s nominální odezvou, abychom viděli, jak se budou lišit. Tyto dva typy modelů použijeme na bílé víno. Model multinomické regrese s ordinální odezvou, pak aplikujeme ještě na data pro červené víno, abychom zjistili, zda ty proměnné, které jsou významné pro kvalitu bílého vína, jsou významné i pro kvalitu červeného vína. Data, která budeme používat byla získána ze stránky [8] a byla použita také v [9]. Pro data o bílém víně máme 4898 pozorování a pro data o červeném víně 1599 pozorování.

4.1 Model multinomické regrese s nominální odezvou pro bílé víno

V této podkapitole budeme kvalitu bílého vína modelovat pomocí multinomické regrese s nominální odezvou, přestože víme, že odezva je ordinální.

Výchozí model, který budeme uvažovat bude aditivní a bude obsahovat všech jedenáct vysvětlujících proměnných. Tento model reprezentuje následující rovnice

$$\begin{aligned} \log\left(\frac{\pi_j}{\pi_J}\right) = & \alpha_j + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \beta_{j3}x_{i3} + \beta_{j4}x_{i4} + \beta_{j5}x_{i5} + \\ & + \beta_{j6}x_{i6} + \beta_{j7}x_{i7} + \beta_{j8}x_{i8} + \beta_{j9}x_{i9} + \beta_{j10}x_{i10} + \\ & + \beta_{j11}x_{i11}, \quad j = 1, \dots, 6, \quad i = 1, \dots, 4898, \end{aligned} \quad (4.1)$$

kde proměnná

- x_{i1} reprezentuje stálou kyselost vína
- x_{i2} udává nestálou kyselost vína
- x_{i3} určuje množství kyseliny citrónové
- x_{i4} udává obsah zbytkového cukru
- x_{i5} představuje množství chloridů
- x_{i6} reprezentuje obsah volného oxidu siřičitého
- x_{i7} udává množství celkového oxidu siřičitého
- x_{i8} určuje hustotu vína
- x_{i9} znamená pH vína
- x_{i10} udává množství síranů
- x_{i11} reprezentuje procentuální podíl alkoholu ve víně

Vysvětlovaná proměnná pak určuje kvalitu vína a nabývá přirozených čísel od 3 do 9.

Model odhadneme pomocí funkce `multinom` z balíku `nnet` statistického softwaru R. Nebudou nás však ani tolik zajímat výsledné odhady, ale hodnota deviance, která vyšla v tomto případě 10660,85.

Nyní je třeba zvolit podmodel tak, že vyřadíme jednu proměnnou. Máme však jedenáct možností. Rozhodneme se podle deviance takto upraveného modelu. Vybereme ten model, který bude mít nejmenší devianci. V tabulce 4.1 uvádíme deviance těchto modelů.

Bez proměnné	Deviance	Bez proměnné	Deviance
x_{i1}	10703,16	x_{i7}	10671,17
x_{i2}	11003,66	x_{i8}	10672,23
x_{i3}	10666,22	x_{i9}	10696,81
x_{i4}	10726,89	x_{i10}	10696,20
x_{i5}	10685,89	x_{i11}	10868,11
x_{i6}	10724,45		

Tabulka 4.1: Deviance jednotlivých modelů při vyřazení jedné vysvětlující proměnné

Z tabulky 4.1 je patrné, že nejmenší devianci má model neobsahující proměnnou x_{i3} .

Nyní budeme testovat nulovou hypotézu, že platí podmodel, který neobsahuje proměnnou x_{i3} , proti alternativě, že platí původní model. Testová statistika je rozdílem deviancí těchto modelů, tedy se rovná 5,37. Počet stupňů volnosti je 6. Výsledná p -hodnota je 0,497, což znamená, že nemůžeme zamítnout podmodel neobsahující jako vysvětlující proměnnou množství kyseliny citrónové.

Dalším krokem bude zvolit druhý podmodel neobsahující x_{i3} a ještě jednu proměnnou. Z deseti nabízených možností opět vybereme ten model, co bude mít nejmenší devianci. Přehled těchto deviancí je v tabulce 4.2.

Bez proměnných	Deviance	Bez proměnných	Deviance
x_{i3}, x_{i1}	10702,11	x_{i3}, x_{i7}	10677,52
x_{i3}, x_{i2}	11016,98	x_{i3}, x_{i8}	10678,51
x_{i3}, x_{i4}	10733,00	x_{i3}, x_{i9}	10700,65
x_{i3}, x_{i5}	10691,50	x_{i3}, x_{i10}	10694,51
x_{i3}, x_{i6}	10726,31	x_{i3}, x_{i11}	10827,17

Tabulka 4.2: Deviance jednotlivých modelů při vyřazení dvou vysvětlujících proměnných

Z tabulky 4.2 je vidět, že nejmenší devianci má model neobsahující proměnné x_{i3} a x_{i7} . Otestujme tedy nulovou hypotézu, že platí model neobsahující tyto dvě proměnné, oproti alternativě, že platí model nezahrnující jen proměnnou x_{i3} . Testová statistika je rovna 11,3 a při 6 stupních volnosti dává p -hodnotu 0,080,

takže na hladině 5% nemůžeme zamítnout platnost podmodelu neobsahujícího proměnné představující množství kyseliny citrónové a množství celkového oxidu siřičitého.

Zvolíme nyní v pořadí již třetí podmodel. Tento podmodel nebude obsahovat proměnné x_{i3} , x_{i7} a ještě třetí proměnnou. Máme devět možností pro volbu takového modelu a přehled deviancí těchto modelů je v tabulce 4.3.

Bez proměnných	Deviance	Bez proměnných	Deviance
x_{i3}, x_{i7}, x_{i1}	10714,58	x_{i3}, x_{i7}, x_{i8}	10691,54
x_{i3}, x_{i7}, x_{i2}	11060,22	x_{i3}, x_{i7}, x_{i9}	10704,39
x_{i3}, x_{i7}, x_{i4}	10742,74	x_{i3}, x_{i7}, x_{i10}	10709,18
x_{i3}, x_{i7}, x_{i5}	10702,11	x_{i3}, x_{i7}, x_{i11}	10821,74
x_{i3}, x_{i7}, x_{i6}	10746,42		

Tabulka 4.3: Deviance jednotlivých modelů při vyřazení tří vysvětlujících proměnných

Z tabulky 4.3 je vidět, že nejmenší devianci má model neobsahující proměnné x_{i3} , x_{i7} a x_{i8} . Budeme-li testovat nulovou hypotézu, že platí tento podmodel, oproti alternativně že platí model neobsahující jen proměnné x_{i3} a x_{i7} , vyjde nám testová statistika 14,02. Při 6 stupních volnosti nám tato testová statistika dává p -hodnotu 0,029. Na 5% hladině zamítáme tedy tento podmodel.

Výsledkem našeho hledání vhodného modelu jako modelu multinomické regrese s nominální odezvou je model neobsahující jako vysvětlující proměnné množství kyseliny citrónové ve víně a množství celkového oxidu siřičitého. Odhady parametrů a další charakteristiky jsou pro tento model uvedeny v tabulce 4.5.

Parametr	Odhad	e ^{odhad}	Směr. odchylka
α_1	39,396	$1,286 \cdot 10^{17}$	2,458
α_2	26,983	$5,231 \cdot 10^{11}$	1,267
α_3	15,572	$5,792 \cdot 10^6$	0,774
α_4	28,176	$1,724 \cdot 10^{12}$	0,733
α_5	116,606	$4,377 \cdot 10^{50}$	0,842
α_6	69,307	$1,257 \cdot 10^{30}$	1,311
β_{11}	-0,824	0,439	0,250
β_{21}	-1,538	0,215	0,264
β_{31}	-1,830	0,160	0,253
β_{41}	-1,951	0,142	0,253
β_{51}	-1,822	0,162	0,255
β_{61}	-1,908	0,148	0,273
β_{12}	8,311	$4,072 \cdot 10^3$	1,388
β_{22}	10,120	$2,484 \cdot 10^4$	0,587
β_{32}	6,305	$5,474 \cdot 10^2$	0,428
β_{42}	0,380	1,463	0,429
β_{52}	-1,573	0,207	0,534
β_{62}	-1,083	0,339	0,852

Parametr	Odhad	e ^{odhad}	Směr. odchylka
β_{14}	-0,167	0,846	0,090
β_{24}	-0,231	0,794	0,078
β_{34}	-0,177	0,838	0,075
β_{44}	-0,113	0,893	0,075
β_{54}	-0,049	0,952	0,075
β_{64}	-0,020	0,980	0,076
β_{15}	30,234	$1,350 \cdot 10^{13}$	1,699
β_{25}	21,690	$2,629 \cdot 10^9$	2,743
β_{35}	22,153	$4,174 \cdot 10^9$	1,654
β_{45}	23,332	$1,358 \cdot 10^{10}$	1,640
β_{55}	3,989	53,994	3,256
β_{65}	10,905	$5,446 \cdot 10^4$	0,248
β_{16}	0,014	1,014	0,040
β_{26}	-0,068	0,935	0,040
β_{36}	-0,023	0,977	0,040
β_{46}	-0,018	0,982	0,040
β_{56}	-0,013	0,986	0,040
β_{66}	0,000	1,000	0,040
β_{18}	45,031	$3,603 \cdot 10^{19}$	2,428
β_{28}	78,987	$2,012 \cdot 10^{34}$	1,245
β_{38}	98,236	$4,608 \cdot 10^{42}$	0,763
β_{48}	77,796	$6,114 \cdot 10^{33}$	0,721
β_{58}	-22,398	$1,874 \cdot 10^{-10}$	0,832
β_{68}	19,048	$1,872 \cdot 10^8$	1,287
β_{19}	-13,048	$2,155 \cdot 10^{-6}$	2,247
β_{29}	-14,998	$3,064 \cdot 10^{-7}$	1,865
β_{39}	-16,023	$1,100 \cdot 10^{-7}$	1,784
β_{49}	-15,928	$1,210 \cdot 10^{-7}$	1,776
β_{59}	-14,590	$4,611 \cdot 10^{-7}$	1,784
β_{69}	-14,449	$5,305 \cdot 10^{-7}$	1,845
β_{110}	0,246	1,279	1,938
β_{210}	2,080	8,002	0,906
β_{310}	1,804	6,073	0,670
β_{410}	2,924	18,607	0,646
β_{510}	4,130	62,204	0,669
β_{610}	3,164	23,674	0,806
β_{111}	-3,169	0,042	0,520
β_{211}	-3,741	0,024	0,480
β_{311}	-3,787	0,023	0,474
β_{411}	-2,893	0,055	0,473
β_{511}	-2,423	0,089	0,473
β_{611}	-2,062	0,127	0,477

Tabulka 4.5: Odhady parametrů a další charakteristiky pro model multinomické regrese s nominální odezvou pro data týkající se kvality bílého vína

Interpretace odhadů uvedených v tabulce 4.5 bude popsána až při srovnání tohoto modelu s modelem multinomické regrese s ordinální odezvou.

4.2 Model multinomické regrese s ordinální odezvou pro bílé víno

Náplní této podkapitoly bude najít vhodný model multinomické regrese s ordinální odezvou pro data týkající se kvality bílého vína.

Stejně jako v případě modelu multinomické regrese s nominální odezvou bude výchozí uvažovaný model aditivní a bude obsahovat všech jedenáct vysvětlujících proměnných. Tento model reprezentuje 6 rovnic, z nichž první uvádíme a následujících 5 si lze snadno domyslet z této rovnice a z (2.7)

$$\log\left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6 + \pi_7}\right) = \alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \\ + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \\ + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11}, \quad (4.2)$$

kde význam proměnných x_{i1}, \dots, x_{i11} je stejný jako u výchozího modelu multinomické regrese s nominální odezvou a je uveden v předchozí podkapitole.

Model odhadneme pomocí funkce `vglm` z balíku `VGAM` statistického programu R. Strategie hledání vhodného modelu bude stejná jako v předchozím případě, a tak nás bude zajímat deviance uvažovaných modelů. U tohoto výchozího modelu je deviance 10900,89.

Nyní je třeba zvolit podmodel výchozího modelu. Tento podmodel vytvoříme tak, že vynecháme jednu vysvětlující proměnnou. Máme však 11 možností a rozhodneme se mezi nimi podle deviance. Vybereme model s nejnižší deviancí. Tyto deviance jsou uvedeny v tabulce 4.6.

Bez proměnné	Deviance	Bez proměnné	Deviance
x_{i1}	10916,09	x_{i7}	10901,74
x_{i2}	11169,70	x_{i8}	10965,10
x_{i3}	10901,15	x_{i9}	10951,89
x_{i4}	11015,44	x_{i10}	10949,83
x_{i5}	10901,08	x_{i11}	10933,38
x_{i6}	10928,91		

Tabulka 4.6: Deviance jednotlivých modelů při vyřazení jedné vysvětlující proměnné

Z tabulky 4.6 je vidět, že nejmenší devianci má model neobsahující vysvětlující proměnnou x_{i5} . Zvolíme ho proto jako náš první podmodel. A budeme testovat nulovou hypotézu, že platí podmodel, oproti alternativě, že platí výchozí model obsahující všechny vysvětlující proměnné. Testová statistika je rozdílem deviance podmodelu a deviance širšího modelu a rovná se 0,19. Počet stupňů volnosti je 1. Dostáváme tak p -hodnotu 0,663. Nemůžeme tedy zamítnout podmodel neobsahující jako vysvětlující proměnnou množství chloridů obsažených v bílém víně.

Budeme chtít dále zvolit druhý podmodel, který nebude obsahovat proměnnou x_{i5} a ještě jednu vysvětlující proměnnou. Máme 10 možností volby takového podmodelu. Přehled deviancí těchto modelů je uveden v tabulce 4.7.

Bez proměnných	Deviance	Bez proměnných	Deviance
x_{i5}, x_{i1}	10917,48	x_{i5}, x_{i7}	10901,92
x_{i5}, x_{i2}	11174,19	x_{i5}, x_{i8}	10968,89
x_{i5}, x_{i3}	10901,30	x_{i5}, x_{i9}	10954,82
x_{i5}, x_{i4}	11023,35	x_{i5}, x_{i10}	10950,45
x_{i5}, x_{i6}	10928,96	x_{i5}, x_{i11}	10933,42

Tabulka 4.7: Deviance jednotlivých modelů při vyřazení dvou vysvětlujících proměnných

Z tabulky 4.7 je patrné, že nejmenší devianci má model neobsahující proměnné x_{i5} a x_{i3} , proto tento model zvolíme jako druhý podmodel. Budeme nyní testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model, který neobsahuje proměnnou x_{i5} . Testová statistika je v tomto případě 0,22. Počet stupňů volnosti je 1. Z toho vychází p -hodnota, která je 0,639. Nemůžeme tedy zamítnout podmodel, ve kterém nevystupují jako vysvětlující proměnné množství kyseliny citrónové a množství chloridů obsažených ve víně.

Úkolem je nyní zvolit v pořadí již třetí podmodel. Tento podmodel nebude obsahovat proměnné x_{i5}, x_{i3} a ještě jednu vysvětlující proměnnou. Máme ovšem 9 možností, jak tento podmodel vybrat. Rozhodneme se opět na základě deviance, která je pro tyto modely uvedena v tabulce 4.8.

Bez proměnných	Deviance	Bez proměnných	Deviance
x_{i3}, x_{i5}, x_{i1}	10918,13	x_{i3}, x_{i5}, x_{i8}	10968,98
x_{i3}, x_{i5}, x_{i2}	11183,20	x_{i3}, x_{i5}, x_{i9}	10954,94
x_{i3}, x_{i5}, x_{i4}	11023,49	x_{i3}, x_{i5}, x_{i10}	10950,90
x_{i3}, x_{i5}, x_{i6}	10929,43	x_{i3}, x_{i5}, x_{i11}	10934,37
x_{i3}, x_{i5}, x_{i7}	10902,11		

Tabulka 4.8: Deviance jednotlivých modelů při vyřazení tří vysvětlujících proměnných

Z tabulky 4.8 je vidět, že nejmenší hodnoty deviance nabývá model neobsahující proměnné x_{i3}, x_{i5} a x_{i7} . Budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model, který nezahrnuje proměnné x_{i3} a x_{i5} . Testová statistika příslušného testu je 0,81 a počet stupňů volnosti je 1, z čehož dostáváme p -hodnotu 0,368. Odtud plyne, že nemůžeme zamítnout podmodel, který jako vysvětlující proměnné nezahrnuje množství kyseliny citrónové, množství chloridů a množství celkového oxidu siřičitého obsaženého ve víně.

Je tedy nutné zvolit čtvrtý podmodel. Tento podmodel nebude obsahovat proměnné x_{i3}, x_{i5}, x_{i7} a ještě jednu proměnnou. Máme 8 možností pro volbu takového podmodelu. Deviance uvažovaných modelů jsou v tabulce 4.9

Bez proměnných	Deviance	Bez proměnných	Deviance
$x_{i3}, x_{i5}, x_{i7}, x_{i1}$	10919,82	$x_{i3}, x_{i5}, x_{i7}, x_{i8}$	10977,90
$x_{i3}, x_{i5}, x_{i7}, x_{i2}$	11204,46	$x_{i3}, x_{i5}, x_{i7}, x_{i9}$	10956,93
$x_{i3}, x_{i5}, x_{i7}, x_{i4}$	11032,13	$x_{i3}, x_{i5}, x_{i7}, x_{i10}$	10951,28
$x_{i3}, x_{i5}, x_{i7}, x_{i6}$	10936,86	$x_{i3}, x_{i5}, x_{i7}, x_{i11}$	10934,38

Tabulka 4.9: Deviance jednotlivých modelů při vyřazení čtyř vysvětlujících proměnných

Z tabulky 4.9 vidíme, že nejmenší devianci má model, v kterém nevystupují proměnné x_{i3}, x_{i5}, x_{i7} a x_{i1} . Zvolíme jej tedy naším čtvrtým podmodelem a budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model, který neobsahuje proměnné x_{i3}, x_{i5} a x_{i7} . Testová statistika je 17,71. Počet stupňů volnosti je 1. Výsledná p -hodnota je $2,573 \cdot 10^{-5}$. Zamítneme tedy platnost modelu, který neobsahuje proměnné představující množství kyseliny citronové, množství chloridů, množství celkového oxidu siřičitého a stálou kyselost vína.

Jako výsledný model multinomické regrese s ordinální odezvou týkající se kvality bílého vína vyšel model, který má za vysvětlující proměnné stálou kyselost vína, nestálou kyselost vína, obsah zbytkového cukru, množství volného oxidu siřičitého, hustotu vína, pH vína, množství síranů a podíl alkoholu ve víně. Odhadnuté parametry a další charakteristiky tohoto modelu jsou uvedeny v tabulce 4.10.

Parametr	Odhad	e^{odhad}	Směr. odchylka
α_1	-467,754	$7,196 \cdot 10^{-204}$	54,510
α_2	-465,391	$7,639 \cdot 10^{-203}$	54,506
α_3	-462,352	$1,596 \cdot 10^{-201}$	54,501
α_4	-459,764	$2,122 \cdot 10^{-200}$	54,496
α_5	-457,512	$2,020 \cdot 10^{-199}$	54,495
α_6	-453,832	$8,006 \cdot 20^{-198}$	54,496
β_1	-0,244	0,784	0,057
β_2	5,073	159,638	0,293
β_4	-0,236	0,790	0,021
β_6	-0,011	0,989	0,002
β_8	478,463	$6,220 \cdot 10^{207}$	55,206
β_9	-2,095	0,123	0,279
β_{10}	-1,816	0,163	0,258
β_{11}	-0,423	0,655	0,072

Tabulka 4.10: Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se kvality bílého vína

4.3 Srovnání modelů a jejich interpretace

Úkolem této podkapitoly je interpretovat modely multinomické regrese s nominální a s ordinální odezvou, které nám vyšly jako nejlepší v předchozích dvou

podkapitolách. Dále bude třeba oba modely srovnat.

Zaměříme se nejprve na model multinomické regrese s nominální odezvou. Pokud vyšel odhad parametru u tohoto modelu záporný, znamená to, že zvýšení hodnoty proměnné příslušící tomuto parametru snižuje pravděpodobnost, že vysvětlovaná proměnná bude nabývat kategorie, ke které parametr náleží, oproti pravděpodobnosti, že bude nabývat referenční kategorie. V případě kladného odhadu parametru se tato pravděpodobnost zvýší. Exponenciála tohoto odhadu určuje kolikrát se daná pravděpodobnost zvýší, popř. sníží.

Z tabulky 4.5 vidíme, že zvýšení proměnné představující stálou kyselost vína snižuje jednotlivé pravděpodobnosti, že bude odezva patřit do prvních šesti kategorií, oproti pravděpodobnosti, že bude patřit do referenční kategorie. Jinými slovy to znamená, že vyšší stálá kyselost vína přispívá k tomu, že bude kvalita vína ohodnocena jako nejvyšší. Hodnoty odhadů parametrů náležející právě ke stálé kyselosti vína ukazují, že nebyla vzata do úvahy ordinalita odezvy, neboť zvýšení kyselosti o jednu jednotku snižuje pravděpodobnost, že bude víno hodnoceno jako nejméně kvalitní, oproti pravděpodobnosti, že bude nejkvalitnější, o polovinu. Avšak stejné zvýšení stálé kyselosti vína již sníží pravděpodobnost, že víno je hodnoceno jako druhé nejlepší, oproti pravděpodobnosti, že je nejkvalitnější, téměř sedmkrát.

Ze stejné tabulky je patrné, že vyšší hodnoty nestálé kyselosti vína budou u vína, které je hodnoceno stupněm 3, 4, 5, 6, ale také 9, který tvoří referenční kategorii.

Dále by se dalo usuzovat, že vyšší hodnoty zbytkového cukru a množství chloridů se budou nacházet u vína nižší kvality.

Můžeme také říct, že je pravděpodobnější, že vyšší hodnoty volného oxidu siřičitého se budou nacházet u kvalitněji hodnocených vína, ale také překvapivě u toho nejméně kvalitního.

Odhady parametrů náležející hustotě vína se nedají nijak souhrně interpretovat.

O pH lze říci, že bude pravděpodobně vysoké u nejlépe hodnocených vín.

Parametry týkající se množství síranů ve víně napovídají, že nižší hodnoty síranů lze očekávat u nejméně kvalitních vín a pak také u těch vín nejkvalitnějších.

K odhadům poslední skupiny parametrů týkajících se podílu alkoholu ve víně lze říci, že čím větší podíl alkoholu tím je pravděpodobnější, že víno bude kvalitnější.

Podívejme se nyní na model multinomické regrese s ordinální odezvou. Z tabulky 4.10 vyplývá, že vyšší hodnoty proměnných odpovídajících stálé kyselosti vína, obsahu zbytkového cukru a volného oxidu siřičitého, pH vína, množství síranů a procentuálnímu podílu alkoholu ve víně se budou objevovat pravděpodobněji u kvalitnějších vín. Zatímco vyšší hodnoty nestálé kyselosti a hustoty vína se budou objevovat pravděpodobněji u vín s nižší kvalitou.

Srovnáme výsledné odhady obou modelů mezi sebou. Lze říci, že o vlivu stálé i nestálé kyselosti vína se modely shodují. Ve významu zbytkového cukru na kvalitu vína se modely rozcházejí. Vliv přisuzovaný volnému oxidu siřičitému je téměř totožný. Ve významu podílu alkoholu ve víně se modely shodují. Vzhledem k rozporuplnosti odhadů multinomické regrese s nominální odezvou nelze ostatní parametry porovnat. Je třeba také zmínit, že proměnná odpovídající množství chloridů ve víně byla modelem multinomické regrese s ordinální odezvou označena

za nevýznamnou narozdíl od modelu s nominální odezvou.

Jako vhodnější model se jednoznačně jeví model multinomické regrese s ordinální odezvou. Je lepší zvolit tento model už z toho důvodu, že odezva zkoumaných dat je ordinální a tuto skutečnost nebral model multinomické regrese s nominální odezvou vůbec v potaz, jak bylo také vidět z rozporuplných výsledných odhadů parametrů.

Na závěr této podkapitoly je třeba okomentovat velmi vysokou hodnotu odhadu parametru β_8 , tento odhad je tak vysoký proto, že proměnná hustota, která přísluší tomuto parametru, má téměř nulovou variabilitu, konkrétně je $8,946 \cdot 10^{-6}$.

4.4 Hledání vhodného modelu pro data o červeném víně

Úkolem této kapitoly bude nalézt vhodný model pro data týkající se kvality červeného vína. Tímto modelem bude model multinomické regrese s ordinální odezvou. Cílem pak bude srovnat, zda stejné proměnné, které ovlivňují kvalitu bílého vína, ovlivňují také kvalitu červeného vína.

Začneme volbou výchozího modelu, kterým bude aditivní model s 11 proměnnými. Tento model reprezentuje 5 rovnic, neboť odezva, která představuje kvalitu vína, nabývá přirozených čísel od 3 do 8. První rovnice je následující a zbylé čtyři si lze z ní a z (2.7) domyslet

$$\log \left(\frac{\pi_1}{\pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6} \right) = \alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \\ + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \\ + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11},$$

kde význam proměnných x_{i1}, \dots, x_{i11} je stejný jako u výchozího modelu multinomické regrese s nominální odezvou pro kvalitu bílého vína a je uveden v podkapitole 4.1.

Model odhadneme pomocí funkce `vglm` z balíku `VGAM` statistického softwaru R. Nezájímají nás odhady parametrů, ale hodnota deviance, která je 3074,767.

Nyní je třeba zvolit podmodel tak, že vynecháme jednu proměnnou. Z 11 možností vybereme ten model, který má nejmenší devianci. Přehled těchto deviancí je v tabulce 4.11.

Bez proměnné	Deviance	Bez proměnné	Deviance
x_{i1}	3077,189	x_{i7}	3097,674
x_{i2}	3148,543	x_{i8}	3076,014
x_{i3}	3077,789	x_{i9}	3076,764
x_{i4}	3078,097	x_{i10}	3140,833
x_{i5}	3089,454	x_{i11}	3173,304
x_{i6}	3078,811		

Tabulka 4.11: Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení jedné vysvětlující proměnné

Z tabulky 4.11 vidíme, že model s nejmenší deviancí je model, který neobsahuje vysvětlující proměnnou x_{i8} . Zvolíme jej tedy jako podmodel a budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí výchozí model. Testová statistika tohoto testu je 1,247. Počet stupňů volnosti je 1. Příslušná p -hodnota je 0,264. Nemůžeme tedy zamítnout platnost podmodelu, který neobsahuje proměnnou odpovídající hustotě vína.

Zvolíme druhý podmodel. Tento podmodel nebude obsahovat proměnnou x_{i8} a ještě jednu proměnnou. Pro volbu takového podmodelu máme 10 možností a opět se rozhodneme podle deviancí, které jsou pro tyto modely uvedeny v tabulce 4.12.

Bez proměnných	Deviance	Bez proměnných	Deviance
x_{i8}, x_{i1}	3077,222	x_{i8}, x_{i6}	3080,612
x_{i8}, x_{i2}	3153,327	x_{i8}, x_{i7}	3100,065
x_{i8}, x_{i3}	3079,040	x_{i8}, x_{i9}	3082,301
x_{i8}, x_{i4}	3078,098	x_{i8}, x_{i10}	3141,513
x_{i8}, x_{i5}	3091,622	x_{i8}, x_{i11}	3330,391

Tabulka 4.12: Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení dvou vysvětlujících proměnných

Z tabulky 4.12 je patrné, že nejmenší devianci má model neobsahující proměnné x_{i8} a x_{i1} , a proto jej zvolíme jako podmodel. Budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model neobsahující vysvětlující proměnnou x_{i8} . Testová statistika příslušící tomuto testu je 1,208. Při jednom stupni volnosti dostáváme p -hodnotu 0,272. Nemůžeme zamítnout platnost podmodelu, který neobsahuje proměnné odpovídající stále kyselosti vína a hustotě vína.

Zvolíme v pořadí již třetí podmodel, do něhož nezařadíme proměnné x_{i8}, x_{i1} a ještě jednu další proměnnou. Na výběr této třetí proměnné máme 9 možností a rozhodneme se podle deviancí, které jsou uvedeny v tabulce 4.13.

Bez proměnných	Deviance	Bez proměnných	Deviance
x_{i8}, x_{i1}, x_{i2}	3154,422	x_{i8}, x_{i1}, x_{i7}	3106,126
x_{i8}, x_{i1}, x_{i3}	3079,069	x_{i8}, x_{i1}, x_{i9}	3090,608
x_{i8}, x_{i1}, x_{i4}	3079,522	x_{i8}, x_{i1}, x_{i10}	3144,777
x_{i8}, x_{i1}, x_{i5}	3096,543	x_{i8}, x_{i1}, x_{i11}	3331,425
x_{i8}, x_{i1}, x_{i6}	3082,099		

Tabulka 4.13: Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení tří vysvětlujících proměnných

Tabulka 4.13 ukazuje, že nejmenší devianci má model, který neuvažuje vysvětlující proměnné x_{i8}, x_{i1} a x_{i3} , proto jej zvolíme jako náš podmodel. Budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model neobsahující jen proměnné x_{i8} a x_{i1} . Testová statistika tohoto testu je 1,847. Počet

stupňů volnosti je 1, z čehož dostáváme p -hodnotu 0,174. Nemůžeme proto zamítnout platnost podmodelu, který neobsahuje vysvětlující proměnné představující stálou kyselost vína, hustotu vína a množství kyseliny citrónové.

Vzhledem k výsledku provedeného testu je třeba zvolit další podmodel. Tento podmodel nebude obsahovat 4 vysvětlující proměnné, kterými budou x_{i8} , x_{i1} a x_{i3} a ještě jedna zvolená proměnná. Pro volbu této čtvrté proměnné máme 8 možností a rozhodneme se podle deviancí, které jsou uvedeny v tabulce 4.14.

Bez proměnných	Deviance	Bez proměnných	Deviance
$x_{i8}, x_{i1}, x_{i3}, x_{i2}$	3167,339	$x_{i8}, x_{i1}, x_{i3}, x_{i7}$	3111,004
$x_{i8}, x_{i1}, x_{i3}, x_{i4}$	3080,848	$x_{i8}, x_{i1}, x_{i3}, x_{i9}$	3090,697
$x_{i8}, x_{i1}, x_{i3}, x_{i5}$	3100,544	$x_{i8}, x_{i1}, x_{i3}, x_{i10}$	3145,255
$x_{i8}, x_{i1}, x_{i3}, x_{i6}$	3085,148	$x_{i8}, x_{i1}, x_{i3}, x_{i11}$	3333,166

Tabulka 4.14: Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení čtyř vysvětlujících proměnných

V tabulce 4.14 vidíme, že nejmenší devianci má model, který nezahrnuje mezi vysvětlující proměnné x_{i8} , x_{i1} , x_{i3} a x_{i4} , zvolíme jej proto jako náš podmodel. Budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model, který neuvažuje proměnné x_{i8} , x_{i1} a x_{i3} . Příslušná testová statistika je 1,779. Počet stupňů volnosti je 1 a p -hodnota vychází 0,182. Nezamítáme proto platnost podmodelu, který neobsahuje vysvětlující proměnné představující stálou kyselost vína, hustotu vína, množství kyseliny citrónové a obsah zbytkového cukru.

Nyní je třeba zvolit další podmodel. Budeme vybírat z modelů, které neobsahují 5 z 11 námi uvažovaných vysvětlujících proměnných. Nebudeme do modelů zahrnovat proměnné x_{i8} , x_{i1} , x_{i3} a x_{i4} a ještě jednu zvolenou proměnnou. Pro volbu této proměnné máme 7 možností a vybereme ten podmodel, který bude mít nejmenší devianci. Deviance jednotlivých modelů jsou uvedeny v tabulce 4.15.

Bez proměnných	Deviance	Bez proměnných	Deviance
$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i2}$	3168,781	$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i9}$	3093,405
$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i5}$	3101,626	$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i10}$	3146,122
$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i6}$	3087,677	$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i11}$	3339,773
$x_{i8}, x_{i1}, x_{i3}, x_{i4}, x_{i7}$	3111,595		

Tabulka 4.15: Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení pěti vysvětlujících proměnných

Z tabulky 4.15 je patrné, že nejmenší devianci má model neobsahující vysvětlující proměnné x_{i8} , x_{i1} , x_{i3} , x_{i4} a x_{i6} . Budeme jej tedy považovat za náš podmodel a budeme testovat nulovou hypotézu, že platí tento podmodel, oproti alternativě, že platí model, který nazahrnuje mezi své vysvětlující proměnné x_{i8} , x_{i1} , x_{i3} a x_{i4} . Testová statistika tohoto testu je 6,829. Počet stupňů volnosti je 1 a příslušná p -hodnota je 0,009. Na hladině 5% tedy zamítáme platnost podmodelu.

Výsledným modelem, který se zabývá kvalitou červeného vína, je model, který obsahuje jako své vysvětlující proměnné nestálou kyselost vína, množství chloridů, obsah volného oxidu siřičitého, množství celkového oxidu siřičitého, pH vína, množství síranů a procentuální podíl alkoholu ve víně. Tabulka 4.16 udává výsledné odhady parametrů tohoto modelu.

Parametr	Odhad	e ^{odhad}	Směr. odchylka
α_1	-1,638	0,194	1,279
α_2	0,279	1,322	1,246
α_3	3,981	53,544	1,244
α_4	6,830	925,247	1,254
α_5	9,834	1865,710	1,283
β_2	3,070	21,538	0,324
β_5	5,778	323,244	1,257
β_6	-0,017	0,983	0,007
β_7	0,012	1,012	0,002
β_9	1,312	3,713	0,364
β_{10}	-2,784	0,062	0,342
β_{11}	-0,884	0,413	0,056

Tabulka 4.16: Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se kvality červeného vína

Z tabulky 4.16 je vidět, že vyšší hodnoty nestálé kyselosti vína, množství chloridů, množství celkového oxidu siřičitého a pH vína zvyšují pravděpodobnost, že víno bude nekvalitní, oproti pravděpodobnosti, že bude patřit mezi kvalitnější vína. Opačný efekt má obsah volného oxidu siřičitého, množství síranů a procentuální podíl alkoholu ve víně.

Parametr	e ^{odhad}	
	Bílé víno	Červené víno
α_1	$7,196 \cdot 10^{-204}$	0,194
α_2	$7,639 \cdot 10^{-203}$	1,322
α_3	$1,596 \cdot 10^{-201}$	53,544
α_4	$2,122 \cdot 10^{-200}$	925,247
α_5	$2,020 \cdot 10^{-199}$	1865,710
α_6	$8,006 \cdot 20^{-198}$	-
β_1	0,784	-
β_2	159,638	21,538
β_3	-	-
β_4	0,790	-
β_5	-	323,244
β_6	0,989	0,983
β_7	-	1,012
β_8	$6,220 \cdot 10^{207}$	-
β_9	0,123	3,713
β_{10}	0,163	0,062
β_{11}	0,655	0,413

Tabulka 4.17: Srovnání modelů multinomické regrese s ordinální odezvou pro bílé a červené víno

Tabulka 4.17 představuje srovnání modelů multinomické regrese s ordinální odezvou pro bílé a červené víno. Z této tabulky je patrné, že pro dva zmíněné modely jsou významné odlišné parametry. Také je třeba zmínit, že vyšší hodnoty proměnné reprezentující pH vína mají odlišný vliv na kvalitu bílého vína než na kvalitu červeného vína. Ve vlivu zbylých společných vysvětlujících proměnných se oba modely shodují.

Na závěr uvedme srovnání se studií popsanou v článku [9]. Tato studie ze stejných dat, které jsme použili my, zjišťovala důležitost jednotlivých vysvětlujících proměnných. Studie pracovala s třemi modely. Prvním byla multinomická regrese, druhým neuronové sítě a třetím metoda, která se anglicky nazývá support vector machines. Poslední model se jevil jako nejlepší. Když srovnáme důležitost jednotlivých proměnných u tohoto modelu a u našich modelů, tak multinomický regresní model pro ordinální data o bílém víně vyřadil jako nedůležité proměnné, které byly na 4., 5. a 10. místě v důležitosti výsledného modelu citovaného článku. Pro červené víno náš model vyřadil proměnné, které byly na 7., 8., 10. a 11. místě v důležitosti. Je tedy patrné, že pro model s červeným vínem jsme dostali relativně stejné výsledky.

Závěr

Tato diplomová práce se zabývala regresními modely s kategoriální odezvou, jež jsou jednou z partií, kterou se zabývá analýza kategoriálních dat.

Práce je rešerší přes důležitou problematiku, která se nevyučuje v základních kurzech na Matematicko-fyzikální fakultě.

Bylo vyzkoušeno teoretické odvození Waldova testu a testu poměrem věrohodností pro model logistické regrese, multinomické regrese s nominální odezvou a multinomické regrese s ordinální odezvou. Odvozené výsledky byly následně porovnány na konkrétních příkladech s příslušnými testy, které jsou implementovány ve statistickém softwaru R.

Popisované postupy byly ilustrovány na několika příkladech. Jednalo se jednak o malé příklady s důkladnějšími studiemi, tak také o jeden velký příklad, ve kterém figurovalo více regresorů.

Práce ovšem nevyčerpala celou problematiku, nazabývala se například testem lineárního trendu [11].

Seznam použité literatury

- [1] AGRESTI, Alan: *Categorical Data Analysis*. 2. vydání. Hoboken: John Wiley & Sons, Inc., 2002. ISBN 0-471-36093-7.
- [2] HOSMER, David W., LEMESHOW, Stanley: *Applied logistic regression*. 2. vydání. Hoboken: John Wiley & Sons, Inc., 2000. ISBN 0-471-35632-8.
- [3] AGRESTI, Alan: *An Introduction to Categorical Data Analysis*. 2. vydání. Hoboken: John Wiley & Sons, Inc., 2007. ISBN 978-0-471-22618-5.
- [4] ANDĚL, Jiří: *Základy matematické statistiky*. 1. vydání. Praha: MATFYZPRESS, 2007. ISBN 80-86732-40-1.
- [5] ZVÁRA, Karel: *Regrese*. 1. vydání. Praha: MATFYZPRESS, 2008. ISBN 978-80-7378-041-8.
- [6] <http://people.sc.fsu.edu/~jburkardt/datasets/regression/x20.txt>
- [7] <http://www.umass.edu/statdata/statdata/stat-logistic.html>
- [8] <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>
- [9] CORTEZ, P., CERDEIRA, A., ALMEIDA, F., MATOS, T., REIS, J.: *Modeling wine preferences by data mining from physicochemical properties*. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
- [10] PEKÁR, S., BRABEC, M.: *Moderní analýza biologických dat - Zobecněné lineární modely v prostředí R*. 1. vydání. Praha: Scientia, 2009. ISBN 978-80-86960-44-9.
- [11] KALINA, J.: *Some tests for evaluation of contingency tables (for biomedical applications)*. Journal of applied mathematics, statistics and informatics 7 (1), 37-50.

Seznam tabulek

1.1	Data o problému s O-ringem v závislosti na teplotě	7
1.2	Odhady parametrů a další charakteristiky pro model logistické regrese pro data týkající se problému s O-ringem v závislosti na teplotě	7
2.1	Data týkající se příslušnosti k politické straně v závislosti na pohlaví a rase	11
2.2	Odhady parametrů a další charakteristiky pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na pohlaví a rase	11
2.3	Data týkající se závažnosti zranění při automobilových nehodách v závislosti na pohlaví, místu nehody a použití bezpečnostního pásu	13
2.4	Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na pohlaví, místu nehody a použití bezpečnostního pásu	14
2.5	Procentuální podíl závažnosti zranění pro data týkající se zranění při automobilových nehodách v závislosti na pohlaví, místu nehody a použití bezpečnostního pásu	14
2.6	Přehled koeficientů, které jsou zahrnuty v modelu $Y \sim P * M * B$ v závislosti na hodnotách jednotlivých vysvětlujících proměnných	17
2.7	Odhady parametrů a další charakteristiky pro model lineární regrese pro data týkající se míry úmrtnosti na cirhózu jater	19
2.8	Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se míry úmrtnosti na cirhózu jater	20
2.9	Odhady parametrů pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na počtu iterací	21
2.10	Odhady parametrů pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na počtu iterací	22
2.11	Odhady parametrů a další charakteristiky pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty	23
2.12	Hodnoty logaritmické věrohodnosti pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty v závislosti na parametrech β_5 a β_7	35
2.13	Hodnoty logaritmické věrohodnosti pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na parametrech β_{11} a β_{21}	36
2.14	Hodnoty logaritmické věrohodnosti pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na parametrech β_1 a β_2	37
2.15	Přehled odhadů parametrů pro zkoumané modely dle způsobu jejich získání	37
2.16	Odhady parametrů a další charakteristiky pro model multinomické regrese s nominální odezvou pro námi vygenerovaná data	38

2.17	Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro námi vygenerovaná data	38
4.1	Deviance jednotlivých modelů při vyřazení jedné vysvětlující proměnné	57
4.2	Deviance jednotlivých modelů při vyřazení dvou vysvětlujících proměnných	57
4.3	Deviance jednotlivých modelů při vyřazení tří vysvětlujících proměnných	58
4.5	Odhady parametrů a další charakteristiky pro model multinomické regrese s nominální odezvou pro data týkající se kvality bílého vína	59
4.6	Deviance jednotlivých modelů při vyřazení jedné vysvětlující proměnné	60
4.7	Deviance jednotlivých modelů při vyřazení dvou vysvětlujících proměnných	61
4.8	Deviance jednotlivých modelů při vyřazení tří vysvětlujících proměnných	61
4.9	Deviance jednotlivých modelů při vyřazení čtyř vysvětlujících proměnných	62
4.10	Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se kvality bílého vína	62
4.11	Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení jedné vysvětlující proměnné	64
4.12	Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení dvou vysvětlujících proměnných	65
4.13	Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení tří vysvětlujících proměnných	65
4.14	Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení čtyř vysvětlujících proměnných	66
4.15	Deviance jednotlivých modelů zabývajících se kvalitou červeného vína při vyřazení pěti vysvětlujících proměnných	66
4.16	Odhady parametrů a další charakteristiky pro model multinomické regrese s ordinální odezvou pro data týkající se kvality červeného vína	67
4.17	Srovnání modelů multinomické regrese s ordinální odezvou pro bílé a červené víno	68

Seznam obrázků

1.1	Hodnota pravděpodobnosti $\pi(x)$ v závislosti na teplotě při letu . . .	8
2.1	Hodnota logaritmické věrohodnosti pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty v závislosti na parametru β_5 a β_7 z pohledu parametru β_5	24
2.2	Hodnota logaritmické věrohodnosti pro model logistické regrese pro data týkající se pokročilosti stádia rakoviny prostaty v závislosti na parametru β_5 a β_7 z pohledu parametru β_7	24
2.3	Širší pohled oproti obrázku 2.1, resp. 2.2 na logaritmickou věrohodnostní funkci pro data týkající se rakoviny prostaty	25
2.4	Hodnota logaritmické věrohodnosti pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na parametru β_{11} a β_{21} z pohledu parametru β_{11}	26
2.5	Hodnota logaritmické věrohodnosti pro model multinomické regrese s nominální odezvou pro data týkající se příslušnosti k politické straně v závislosti na parametru β_{11} a β_{21} z pohledu parametru β_{21}	26
2.6	Širší pohled oproti obrázku 2.4, resp. 2.5 na logaritmickou věrohodnostní funkci pro data týkající se příslušnosti k politické straně	27
2.7	Hodnota logaritmické věrohodnosti pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na parametru β_1 a β_2 z pohledu parametru β_1	28
2.8	Hodnota logaritmické věrohodnosti pro model multinomické regrese s ordinální odezvou pro data týkající se závažnosti zranění při automobilových nehodách v závislosti na parametru β_1 a β_2 z pohledu parametru β_2	29
2.9	Širší pohled oproti obrázku 2.7, resp. 2.8 na logaritmickou věrohodnostní funkci pro data týkající se závažnosti zranění při automobilových nehodách	29

Přílohy

Zdrojový kód k řešenému příkladu z logistické regrese z kapitoly 1.4:

```
problem=c(0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1,
           0, 0, 0, 0, 0, 0, 1, 0, 1)
teplota=c(66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78,
           67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58)
fit = glm(problem~teplota, family=binomial)
summary(fit)
anova(fit, test="Chisq")
```

Zdrojový kód k příkladu s multinomickou nominální odezvou z kapitoly 2.2:

```
require(nnet)
politika <- read.table("prvni.dat", header=TRUE)
politika$strana2 <- relevel(politika$strana, ref = "nezavisly")
test <- multinom(strana2 ~ pohlavi + rasa, data = politika)
summary(test)
```

Zdrojový kód k příkladu s kumulativním logitovým modelem s originální odezvou z kapitoly 2.4:

```
require(VGAM)
zraneni <- read.table("druhy.dat", header=TRUE)
fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ pohlavi+misto+bezpecnostnipas,
            family=cumulative(parallel=TRUE), data=zraneni)
summary(fit)
```