UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Zobecněné lineární modely se smíšenými efekty s aplikací ve výzkumu HIV

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval(a): **Bc. Monika Launerová**
Studijní program: B1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2018

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Monika Launerová

Název práce: Zobecněné lineární modely se smíšenými efekty s aplikací ve výzkumu HIV

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2018

Abstrakt: Cílem práce je analýza dat týkajících se výskytu metylovaného cytosinu v dinukleotidech HIV viru. Budeme testovat hypotézu, zda se výskyt těchto vazeb zvyšuje s rostoucí délkou doby od nákazy. Prostředkem bude zobecněný lineární model obsahující jak pevné, tak i náhodné efekty.

Klíčová slova: AIDS, alternativní rozdělení, HIV, kategoriální proměnné, logistická regrese, metoda maximální věrohodnosti, normální rozdělení, pomocné proměnné, reálná aplikace, simulace Monte Carlo, software R, spojovací funkce, zobecněné lineární modely se smíšenými efekty

Počet stran: 52

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Monika Launerová

Title: Generalized Linear Mixed Models with application in HIV research

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2018

Abstract: The aim of this thesis is to analyse the data for the presence of the methylated cytosine in the dinucleotides of the HIV virus. We will test the hypothesis whether the presence of these elements increases with the increasing lenght of the infection. It will be used to generate the linear model containing both the fixed and random effects.

Key words: AIDS, Alternative Distribution, HIV, Categorical Data, Logistic Regression, Maximum Likelihood Estimation, Normal Distribution, Dummy Variables, Real Application, Simulation Monte Carlo, software R, Link Function, Generalized Linear Mixed Models

Number of pages: 52

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Úvod **7**

1 Kategoriální data **9**
1.1 Binomické rozdělení 10
1.2 Alternativní rozdělení 12
1.3 Multinomické rozdělení 12

2 Regresní analýza **14**
2.1 Vybrané typy lineárních regresních modelů 14
2.1.1 Klasické lineární regresní modely 15
2.1.2 Zobecněné lineární modely 16
2.1.3 Logistická regrese 19

3 Zobecněné lineární modely se smíšenými efekty **28**
3.1 Zobecněné lineární modely s jedním náhodným efektem 29
3.2 Logistické GLMM pro binární párová data 31
3.3 Logit-normální model pro binární data 33

4 Reálná aplikace - výzkum HIV **35**
4.1 Virus HIV 35
4.2 Data HIV 36
4.2.1 Pevné efekty (GLM) 39
4.2.2 Jednorozměrný smíšený efekt (GLMM) 40
4.2.3 Vícerozměrný smíšený efekt (GLMM) 44

Dodatek **46**
Další typy distribucí 46
Tabulky 47

Závěr **49**

Literatura **51**

Poděkování

Ráda bych poděkovala vedoucímu diplomové práce Mgr. Ondřeji Vencálkovi, Ph.D. za spolupráci a hlavně za čas, který mi věnoval při konzultacích. Také bych chtěla poděkovat rodině, přátelům a mému příteli za podporu během celého studia.

Úvod

Samotné téma diplomové práce - Zobecněné lineární modely se smíšenými efekty s aplikací ve výzkumu HIV - prozrazuje, o čem tato práce bude. V teoretické části bude představena potřebná problematika zobecněných lineárních modelů, která bude následně využita v části praktické. Praktická část se zabývá virem HIV a nemocí AIDS. Cílem práce je z dat o pacientech, kteří jsou nakaženi touto nemocí, zjistit, zda má délka doby od nákazy vliv na výskyt methylovaných vazeb.

Všude kolem nás se nachází mnoho bacilů a virů. Každý z nás měl někdy oslabenou imunitu. Občas je nám velmi zle, a i když je to třeba jen obyčejná chřipka, máme pocit, že téměř umíráme. Jak se potom musí asi cítit člověk, který se nakazí virem HIV? Tento virus napadá určitou skupinu bílých krvinek a snižuje jejich počet. To následně způsobuje selhání imunity a onemocnění AIDS. Člověk se pak stává velmi ohroženým a v okamžiku nákazy jiným virem opravdu může umřít. A to i na obyčejnou chřipku.

Toto onemocnění je převážně spojeno s rizikovým životním stylem. Nejčastější forma nákazy je nechráněný sex, z toho dvě třetiny případů jsou u homosexuálů nebo bisexuálů. Dále se nákaza vyskytuje u lidí, kteří užívali drogy pomocí použité jehly. V České republice se nejvíce nakažených objevuje v Praze, Středočeském a Jihomoravském kraji. K 28. 2. 2018 je v ČR nakaženo celkem 3048 mužů a 582 žen. Tyto a další statistiky lze nalézt v [13].

Tato práce pojednává o zobecněných lineárních modelech se smíšenými efekty a jejich následné aplikaci do výzkumu HIV. Budeme zjišťovat, zda se s délkou doby od nákazy zvyšuje výskyt methylovaného cytosinu v dinukleotidech viru HIV. Aby se nakažený subjekt mohl vyléčit, bylo by potřeba, aby v dinukleotidech viru tyto

methylované cytosiny neměl. Tedy čím jich má méně, tím lépe.

Na začátku budou představena kategoriální data a jejich nejznámější rozdělení. Následně bude nahlédnuto na regresní analýzu a vybrané typy lineárních regresních modelů. Dále již navazuje kapitola o zobecněných lineárních modelech se smíšenými efekty. Samotné praktické části je pak věnována poslední kapitola. Celá práce vychází především z knihy Categorical data analysis od Alana Agrestiho (viz [1]), ale bylo čerpáno i z dalších titulů zmíněných v sekci Literatura (str. 51).

Kapitola 1

Kategoriální data

První kapitolu věnujeme kategoriálním proměnným a jejich nejznámějším teoretickým rozdělením. Čerpáno bylo z [1], [2], [6], [8] a [12].

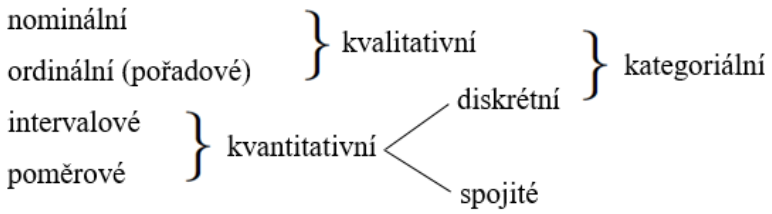
Proměnná, kterou není možné kvantifikovat, ale můžeme ji řadit do tříd, se nazývá kategoriální proměnná. Lze ji dělit podle několika kritérií. První možností dělení je proměnná vysvětlovaná a vysvětlující. Vysvětlující proměnné (nezávislé), které máme obvykle k dispozici, slouží k tomu, abychom si udělali nějakou představu o hodnotách vysvětlované proměnné (závislé).

Podle uspořádání kategorií rozlišujeme proměnné nominální, ordinální, intervalové a poměrové. Nominální proměnné nemůžeme přirozeně uspořádat, můžeme pouze říct, zda nabývají hodnoty stejné nebo různé, například jakou kdo studuje školu, fakultu a obor. Pro práci s nimi lze použít jen ty metody, které považují všechny kategorie za rovnocenné. Ordinální (pořadové) proměnné již uspořádat lze. Jako příklad můžeme uvést nejvyšší ukončené vzdělání dané osoby. Intervalové (rozdílové) proměnné jsou takové, pro jejichž dvě hodnoty můžeme vypočítat, o kolik je jedna hodnota větší než druhá. Poměrové (podílové) proměnné jsou takové, pro jejichž dvě hodnoty můžeme vypočítat, kolikrát je jedna hodnota větší než druhá.

Další možné dělení je na proměnné kvalitativní a kvantitativní. Kvalitativní proměnné jsou ty, jejichž hodnoty jsou popsány pouze slovně. Kvantitativní jsou číselně vyjádřené proměnné, které můžeme dále dělit na diskrétní (nabývají pouze

celočíselných hodnot) a spojité (nabývají libovolných hodnot z určitého intervalu). Všechny nominální proměnné jsou kvalitativní, všechny intervalové a poměrové proměnné jsou kvantitativní. U ordinálních proměnných toto rozdělení není zcela jednoznačné, přikloníme se však k proměnným kvalitativním. Nominální, ordinální a kvantitativní diskrétní proměnné můžeme souhrnně označit jako kategoriální.

Obrázek 1.1 shrnuje uvedené dělení.



Obrázek 1.1: Schéma kategoriálních proměnných

Podle jiného hlediska můžeme kategoriální data dělit na dichotomická a vícekategoriální. Dichotomická (alternativní) data nabývají pouze dvou kategorií, například muž a žena či kuřák a nekuřák. Vícekategoriální data nabývají více než dvou kategorií, například různé způsoby cesty do práce - autem, autobusem, na kole či pěšky.

1.1. Binomické rozdělení

Nechť n je přirozené číslo a $\pi \in (0, 1)$. Nechť veličina X nabývá pouze hodnot $0, 1, ..., n$, a to s pravděpodobnostmi

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \qquad k = 0, 1, ..., n.$$

Pak říkáme, že X má binomické rozdělení, a píšeme $X \sim \text{Bi}(n, \pi)$.

Pro střední hodnotu binomického rozdělení platí

$$E(X) = \sum_{k=0}^{n} kP(X=k) = \sum_{k=1}^{n} k\binom{n}{k}\pi^k(1-\pi)^{n-k} = \sum_{k=1}^{n} k\frac{n!}{(n-k)!k!}\pi^k(1-\pi)^{n-k} =$$

$$= \sum_{k=1}^{n} \frac{n!}{(n-k)!(k-1)!}\pi^k(1-\pi)^{n-k} =$$

$$= n\pi \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!}\pi^{k-1}(1-\pi)^{n-k} = n\pi \sum_{k=1}^{n} \binom{n-1}{k-1}\pi^{k-1}(1-\pi)^{n-k} =$$

$$= n\pi \sum_{j=0}^{n-1} \binom{n-1}{j}\pi^j(1-\pi)^{n-1-j} = n\pi(\pi+1-\pi)^{n-1} = n\pi.$$

Výpočet rozptylu využívá vztah

$$E(X^2) = E(X(X-1)) + E(X),$$

pro který platí

$$E(X(X-1)) = \sum_{k=0}^{n} k(k-1)\binom{n}{k}\pi^k(1-\pi)^{n-k} = \sum_{k-2}^{n} \frac{n!}{(k-2)!(n-k)!}\pi^k(1-\pi)^{n-k} =$$

$$= n(n-1)\pi^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!}\pi^{k-2}(1-\pi)^{n-k} =$$

$$= n(n-1)\pi^2 \sum_{j=0}^{n-2} \binom{n-2}{j}\pi^j(1-\pi)^{n-j-2} =$$

$$= n(n-1)\pi^2(\pi+(1-\pi))^{n-2} = n(n-1)\pi^2.$$

Potom

$$var(X) = E(X^2) - (E(X))^2 = (n(n-1)\pi^2 + n\pi) - (n\pi)^2 =$$

$$= (n\pi)^2 - n\pi^2 + n\pi - (n\pi)^2 = n\pi(1-\pi).$$

1.2. Alternativní rozdělení

Alternativní rozdělení je speciálním případem rozdělení binomického pro $n = 1$. Nechť $\pi \in (0, 1)$. Nechť veličina X nabývá pouze dvou hodnot $x_1 = 0, x_2 = 1$, a to s pravděpodobnostmi

$$\pi_1 = P(X = 0) = 1 - \pi, \qquad \pi_2 = P(X = 1) = \pi.$$

Pak říkáme, že X má alternativní (nula-jedničkové) rozdělení a píšeme $X \sim \text{Alt}(\pi)$.
Střední hodnota pro alternativní rozdělení se dá vyjádřit jako

$$E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$$

a rozptyl

$$var(X) = E(X^2) - (E(X))^2 = (0^2(1 - \pi) + 1^2\pi) - \pi^2 = \pi(1 - \pi).$$

1.3. Multinomické rozdělení

Multinomické rozdělení je rozšířením binomického rozdělení, kdy má sledovaná proměnná více než dvě kategorie. Nechť

$$Y_{ij} = \begin{cases} 1 & \text{pokud výsledek i-tého pokusu je z j-té kategorie} \\ 0 & \text{jinak} \end{cases}$$

Potom i-tý pokus můžeme psát jako vektor

$$Y_i = (Y_{i1}, ..., Y_{ic}), \qquad \sum_j Y_{ij} = 1,$$

kde c je celkový počet kategorií. Poznamenejme, že nám stačí znát $(c - 1)$ členů vektoru, neboť je pevně daný součet $\sum_j Y_{ij} = 1$.
Celkový počet pozorování v j-té kategorii je roven

$$N_j = \sum_i Y_{ij}.$$

Potom má vektor $N = (N_1, ..., N_c)$ multinomické rozdělení.

Označíme-li π_j pravděpodobnost, že $Y_{ij} = 1$, můžeme vyjádřit pravděpodobnostní funkci multinomického rozdělení s parametry $n, \pi_1, ..., \pi_c$ jako

$$P(N = n) = P(N_1 = n_1, ..., N_c = n_c) = \frac{n!}{n_1! \cdots n_c!} \pi_1^{n_1} \cdots \pi_c^{n_c},$$

kde $n = \sum_j n_j$.

Celkové počty pozorování v rámci jednotlivých kategorií mají binomické rozdělení s parametry n a π_j. Protože N_j je součet hodnot $Y_{ij},\ i = 1, \ldots, n$ ze všech n nezávislých pokusů, pro multinomické rozdělení platí

$$E(N_j) = E\Big(\sum_{i=1}^{n} Y_{ij}\Big) = \sum_{i=1}^{n} E(Y_{ij}) = \sum_{i=1}^{n} \Big(1 \cdot \pi_j + 0 \cdot (1 - \pi_j)\Big) = \sum_{i=1}^{n} \pi_j = n\pi_j,$$

$$var(N_j) = var\Big(\sum_{i=1}^{n} Y_{ij}\Big) = \sum_{i=1}^{n} var(Y_{ij}) = \sum_{i=1}^{n} \Big(E(Y_{ij})^2 - [E(Y_{ij})]^2\Big)$$

$$= \sum_{i=1}^{n} \Big([1^2 \cdot \pi_j + 0^2 \cdot (1 - \pi_j)] - [\pi_j]^2\Big) = \sum_{i=1}^{n} \Big(\pi_j - \pi_j^2\Big) = n\pi_j(1 - \pi_j).$$

Pro vyjádření kovariance nejprve odvodíme, že $E(Y_{ij}Y_{ik}) = 0$ pro $\forall j \neq k$, neboť pouze jedna z veličin Y_{ij} je v každém z pokusů rovna jedné, zbylé veličiny jsou nulové.

$$cov(N_j, N_k) = cov\Big(\sum_{i=1}^{n} Y_{ij}, \sum_{i=1}^{n} Y_{ik}\Big) = \sum_{i=1}^{n} cov(Y_{ij}, Y_{ik})$$

$$= \sum_{i=1}^{n} \Big(E(Y_{ij}, Y_{ik}) - E(Y_{ij})E(Y_{ik})\Big) = \sum_{i=1}^{n} \Big(0 - \pi_j\pi_k\Big) = -n\pi_j\pi_k.$$

Kapitola 2

Regresní analýza

Regresní analýza popisuje vztah mezi proměnnými. Jejím úkolem je vysvětlit závislost jedné proměnné na jiných proměnných. Budeme uvažovat jednu vysvětlovanou proměnnou Y, která závisí na proměnných vysvětlujících $x_1, x_2, \ldots, x_n$. Představíme si vybrané typy lineárních regresních modelů včetně logistické regrese. Čerpáno bylo z [4], [6], [8] a [10].

Nechť náhodná veličina Y má pro danou hodnotu $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ a parametry $\beta_1, \beta_2, \ldots, \beta_p$ podmíněnou střední hodnotu $E(Y|\mathbf{x}) = h(\mathbf{x}, \beta_1, \beta_2, \ldots, \beta_p)$. Funkce h proměnné $\mathbf{x}$ se nazývá regresní funkce a parametry $\beta_1, \beta_2, \ldots, \beta_p$ regresní koeficienty.

Podle tvaru regresní funkce rozlišujeme různé typy regresních modelů, modely lineární či nelineární v parametrech. My se zeměříme na modely lineární v parametrech.

2.1. Vybrané typy lineárních regresních modelů

V této kapitole uvedeme přehled vybraných regresních modelů lineárních v parametrech. Nejprve představíme klasický lineární model, poté ho rozšíříme na zobecněný lineární model s pevnými efekty a v další kapitole na model jak s pevnými, tak s náhodnými efekty.

2.1.1. Klasické lineární regresní modely

Nejjednodušším typem regresního modelu je klasický lineární model.

Obecný lineární model pro i-té pozorovnání má tvar

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i,$$

kde $i = 1, 2, ..., n$. Rovnice pro všechna pozorování můžeme uspořádat jako soustavu rovnic a tu pak zapsat pomocí matic

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde vektor pozorovnání $\mathbf{Y}$, matice regresorů $\mathbf{x}$, vektor koeficientů $\boldsymbol{\beta}$ a vektor náhodných chyb $\boldsymbol{\varepsilon}$ mají následující podobu

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n. \end{pmatrix}.$$

Pro tento model předpokládáme:

1. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$,
2. $var(\mathbf{Y}|\mathbf{x}) = var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, kde σ^2 je neznámá kladná konstanta,
3. $\mathbf{X}$ je pevně daná matice typu $n \times (p + 1)$,
4. matice $\mathbf{X}$ má plnou sloupcovou hodnost, tj. $h(\mathbf{X}) = p + 1$, přičemž $p < n$,
5. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Kvůli normálnímu rozdělení náhodných chyb se klasické lineární modely někdy nazývají také normální lineární regresní modely. Nevýhodou tohoto modelu je, že data, která vstupují do modelu, musí splňovat poměrně silné předpoklady:

1. náhodné chyby ε_i jsou navzájem nezávislé a normálně rozdělené,

2. všechny náhodné chyby mají stejný konstantní rozptyl.

U lineární regrese se uvažuje, že jak proměnná vysvětlovaná, tak vysvětlující představují spojité náhodné veličiny. V případě, kdy závisle proměnná Y je dichotomická náhodná veličina, tj. veličina, která může nabývat pouze dvou hodnot, již není možné použít lineární regresi k predikci její střední hodnoty.

2.1.2. Zobecněné lineární modely

V této kapitole ukážeme takové typy modelů, v nichž jsou některé z předpokladů klasického lineárního modelu porušeny.

Často není některý z předpokladů klasického lineárního modelu splněn. Potom se předpoklady modelu uvolňují a zkoumají se vlastnosti takto nově vytvořených typů modelů.

Nechť $Y \sim Alt(\pi)$. Je vidět, že použití modelu ve tvaru $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \boldsymbol{\varepsilon}$ není korektní, neboť na levé straně vystupuje diskrétní veličina, zatímco na pravé straně má chybová složka dle vlastností lineárního regresního modelu normální rozdělení. Pokud vyjádříme střední hodnotu

$$E(Y_i|\mathbf{x}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i\boldsymbol{\beta},$$

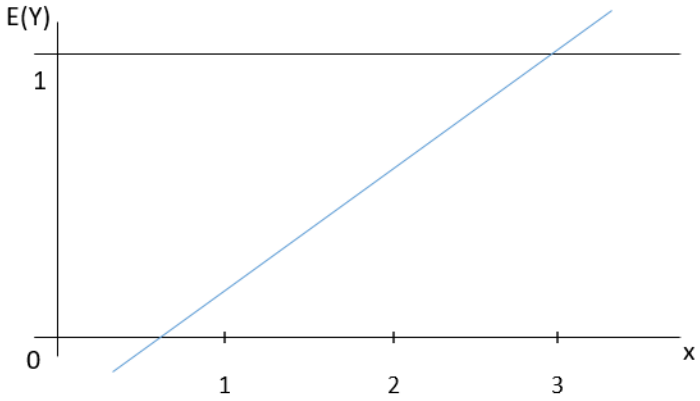
nebo úspornějším maticovým zápisem

$$E(\mathbf{Y}) = \mathbf{x}\boldsymbol{\beta},$$

vidíme, že střední hodnota náhodné veličiny je nulová.

Objevuje se tu ovšem problém. Lineární funkce se může nacházet v celém oboru reálných čísel, zatímco střední hodnota alternativního rozdělení je číslo z intervalu $(0, 1)$. Tuto situaci znázorňuje obrázek 2.1.

Pokusíme se tedy nalézt tzv. spojovací funkci, která interval $(0, 1)$ transformuje na celé $\mathbf{R}$.



Obrázek 2.1: Problém diskrétní hodnoty závisle proměnné

Spojovací funkce

Spojovací funkce (anglicky link-function) je funkce, která určuje vztah mezi vysvětlovanou proměnnou a regresní funkcí. Označíme ji g. Potom platí

$$g(E(\mathbf{Y}|\mathbf{x})) = \mathbf{x}\boldsymbol{\beta}.$$

Volba spojovací funkce není zcela libovolná, v tabulce 2.1 jsou znázorněny typické kombinace spojovacích funkcí a očekávaných distribucí vysvětlované proměnné. O Poissonově, gama a Gaussově rozdělení si čtenář může přečíst v dodatku, str.46.

Typ proměnné	„Typická“ spojovací funkce	Referenční distribuce
počty (frekvence)	log	Poissonova
pravděpodobnost (relativní frekvence)	logit nebo probit	binomická
rozměry, poměry	inverze nebo log	gama
vzácné typy měření	identita	Gaussova (normální)

Tabulka 2.1: Volba spojovací funkce.[9]

Uvažujeme, že $\mathbf{Y}$ má alternativní rozdělení, což je speciální případ binomického rozdělení, tedy můžeme předpokládat, že jako spojovací funkci použijeme

logit nebo probit. Tyto funkce jsou objasněny níže.

Víme, že pro střední hodnotu alternativního rozdělení platí $E(\mathbf{Y}) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$. Tedy můžeme psát

$$g(\pi) = \mathbf{x}\boldsymbol{\beta}.$$

Zde se nacházíme na intervalu $(0, 1)$, neboť pravděpodobnost $\pi \in (0, 1)$.

Šance

Šance (anglicky odds) vyjadřuje, kolikrát je (v dané kategorii) pravděpodobnější, že jev A nastane, než že nenastane. Matematicky ji můžeme zapsat jako

$$\text{šance}(A) = \frac{P(A)}{1 - P(A)},$$

kde P(A) značíme pravděpodobnost, že nastane jev A. Hodnoty šance leží v intervalu $(0, \infty)$.

Dále je třeba transformovat interval $(0, \infty)$ na interval $(-\infty, \infty)$. K tomu použijeme logaritmus, čímž zavedeme logitovou funkci:

$$\log\left(\frac{P(A)}{1 - P(A)}\right) = \text{logit}(P(A)).$$

Tím jsme se dostali na celé $\mathbf{R}$. Našli jsme tedy hledanou spojovací funkci:

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \text{logit}(\pi).$$

Poznámka 2.1.1 *Použili jsme logitovou funkci. Probitovou funkci a porovnání logitové a probitové funkce nalezneme například v [5].*

Označíme-li logit $(\pi) = \log(\frac{\pi}{1-\pi}) = k$ a upravíme

$$\log(\frac{\pi}{1 - \pi}) = k$$

$$\frac{\pi}{1-\pi} = e^k,$$

$$\pi = (1-\pi)e^k = e^k - \pi e^k,$$

$$\pi + \pi e^k = \pi(1+e^k) = e^k,$$

potom dostaneme

$$\pi = \frac{e^k}{1+e^k} \quad \Longrightarrow \quad \lim_{k\to\infty} \frac{e^k}{1+e^k} = 1$$

$$\Longrightarrow \quad \lim_{k\to-\infty} \frac{e^k}{1+e^k} = 0.$$

Pro $k = 1$ platí

$$\pi = \frac{e}{1+e}.$$

Pro $k = -1$

$$\pi = \frac{e^{-1}}{1+e^{-1}} = \frac{\frac{1}{e}}{1+\frac{1}{e}} = \frac{\frac{1}{e}}{\frac{1+e}{e}} = \frac{1}{1+e}.$$

Poznámka 2.1.2 *Vidíme, že platí* $\frac{e}{1+e} + \frac{1}{1+e} = 1$.

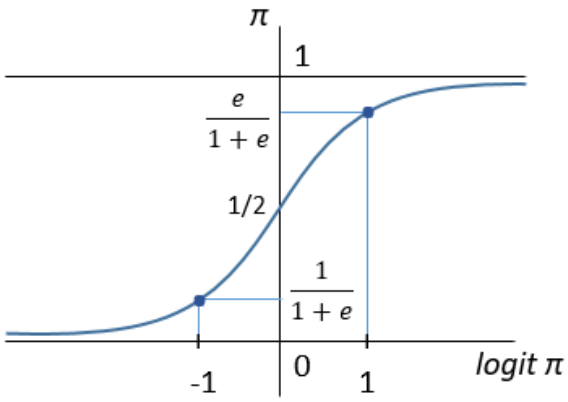
2.1.3. Logistická regrese

Logistická regrese představuje regresní vztah diskrétní povahy, speciálně vztah binárních dat se dvěma definovanými kategoriemi.
Logistický model má tvar

$$\text{logit}(\pi(\mathbf{x})) = \mathbf{x}\beta,$$

kde logit je logaritmus šance.

Graficky lze logistický model interpretovat jako sigmoidální křivku $\pi(\mathbf{x})$, která vyjadřuje odhad závislosti pravděpodobnosti výskytu sledovaného jevu v závislosti na $\mathbf{x}$ (viz obrázek 2.2).



Obrázek 2.2: Křivka logistického modelu

Interpretace parametrů

Nyní se podívame, jak lze parametry $\beta_0, \beta_1, ...\beta_p$ interpretovat.

Nejprve vezmeme v úvahu nejjednodušší případ, kdy máme jednu vysvětlující proměnnou x_1, která nabývá pouze dvou hodnot, $x_1 = \{0, 1\}$. Můžeme si to například představit tak, že buňka buď je (1) nebo není (0) vystavena působení nějaké látky. Potom má funkce logistické regrese tvar

$$\log\left(\frac{\pi(x_1)}{1-\pi(x_1)}\right) = \beta_0 + \beta_1 x_1.$$

Pokud na ni látka nepůsobí, $x_1 = 0$, pak dosazením do rovnice získáme

$$\log\left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_0.$$

Pro buňku, na kterou látka působí, získáme vztah

$$\log\left(\frac{\pi(1)}{1-\pi(1)}\right) = \beta_0 + \beta_1.$$

Máme tedy dvě rovnice o dvou neznámých. Z první rovnice pomocí exponenciální funkce získáme šanci pro $x_1 = 0$:

$$\log\left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_0 \qquad \Rightarrow \qquad e^{\beta_0} = \frac{\pi(0)}{1-\pi(0)}.$$

Odečtením rovnic

$$\log\left(\frac{\pi(1)}{1-\pi(1)}\right) - \log\left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_0 + \beta_1 - \beta_0$$

$$\log \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \log \frac{\text{šance}(1)}{\text{šance}(0)} = \beta_1 \qquad \Rightarrow \qquad e^{\beta_1} = \frac{\text{šance}(1)}{\text{šance}(0)}$$

získáváme poměr šancí (anglicky odds ratio). Poměr šancí udává, kolikrát větší šanci má výskyt daného jevu mezi případy s $x_1 = 1$ než pro ty s $x_1 = 0$.

Nyní můžeme vyslovit hypotézy. Pokud nás bude zajímat, zda x_1 má vliv na Y, testujeme $H_0 : \beta_1 = 0$. Platí, že $e^{\beta_1} = e^0 = 1$. Tedy

$$\frac{\text{šance}(1)}{\text{šance}(0)} = 1 \qquad \Rightarrow \qquad \text{šance}(1) = \text{šance}(0).$$

Při nulové hypotéze $H_0 : \beta_1 < 0$ platí $e^{\beta_1} < 1$, a tedy

$$\frac{\text{šance}(1)}{\text{šance}(0)} < 1 \qquad \Rightarrow \qquad \text{šance}(1) < \text{šance}(0).$$

Nyní úvahy rozšíříme na tři možné hodnoty vysvětlující proměnné, $x \in \{0, 1, 2\}$, kde 0 značí, že na buňku nepůsobí žádná látka, 1, že je vystavena působení nějaké látky A a 2, že na ni působí látka B.

Zde si již musíme definovat tzv. pomocné proměnné (angl. dummy variables).

$$\text{Pomocné proměnné} \begin{cases} x_1 = \begin{cases} 1 \text{ ...působí látka A} \\ 0 \text{ ...jinak} \end{cases} \\ x_2 = \begin{cases} 1 \text{ ...působí látka B} \\ 0 \text{ ...jinak} \end{cases} \end{cases} \tag{2.1}$$

Pro buňku, na kterou působí látka A, pak platí: $(x_1, x_2) = (1, 0)$
Pro buňku vystavenou působení látky B: $(x_1, x_2) = (0, 1)$
A pro neošetřenou buňku: $(x_1, x_2) = (0, 0)$

Jako referenční skupinu zvolíme skupinu s neošetřenými buňkami a s ní budeme porovnávat ostatní skupiny.

Logistická funkce má tedy tvar

$$\log \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

kde vektor $\mathbf{x} = (x_1, x_2)$.

Nejprve budeme uvažovat skupinu, kde působí látka A. Ta nám spolu s referenční skupinou poskytne dvě rovnice

$$\log\left(\frac{\pi(0,0)}{1-\pi(0,0)}\right) = \beta_0$$

$$\log\left(\frac{\pi(1,0)}{1-\pi(1,0)}\right) = \beta_0 + \beta_1.$$

Z první rovnice opět hned získáme vyjádřenou šanci pro skupinu neošetřenou

$$\log\left(\frac{\pi(0,0)}{1-\pi(0,0)}\right) = \beta_0 \qquad \Rightarrow \qquad e^{\beta_0} = \frac{\pi(0,0)}{1-\pi(0,0)}.$$

Odečtením rovnic

$$\log\left(\frac{\pi(1,0)}{1-\pi(1,0)}\right) - \log\left(\frac{\pi(0,0)}{1-\pi(0,0)}\right) = \beta_0 + \beta_1 - \beta_0$$

$$\log \frac{\frac{\pi(1,0)}{1-\pi(1,0)}}{\frac{\pi(0,0)}{1-\pi(0,0)}} = \log \frac{\text{šance}(1,0)}{\text{šance}(0,0)} = \beta_1 \qquad \Rightarrow \qquad e^{\beta_1} = \frac{\text{šance}(1,0)}{\text{šance}(0,0)}$$

získáváme poměr šancí skupiny ošetřené látkou A vzhledem k referenční skupině.

Obdobné výpočty provedeme pro skupinu, která je ošetřena látkou B. S referenční skupinou tvoří dvě rovnice

$$\log\left(\frac{\pi(0,0)}{1-\pi(0,0)}\right) = \beta_0$$

$$\log\left(\frac{\pi(0,1)}{1-\pi(0,1)}\right) = \beta_0 + \beta_2.$$

Odečtením rovnic

$$\log\left(\frac{\pi(0,1)}{1-\pi(0,1)}\right) - \log\left(\frac{\pi(0,0)}{1-\pi(0,0)}\right) = \beta_0 + \beta_2 - \beta_0$$

$$\log \frac{\frac{\pi(0,1)}{1-\pi(0,1)}}{\frac{\pi(0,0)}{1-\pi(0,0)}} = \log \frac{\text{šance}(0,1)}{\text{šance}(0,0)} = \beta_2 \qquad \Rightarrow \qquad e^{\beta_2} = \frac{\text{šance}(0,1)}{\text{šance}(0,0)}$$

získáváme poměr šancí skupiny, která je ošetřena látkou B vzhledem k referenční skupině.

Odhad regresního parametru

Pro odhad parametrů β_0 a β_1 můžeme využít odhady šance při různých hodnotách x. Nechť n_{ij} pro $x = i$ a $Y = j$ je zjištěná četnost, kde $i, j = 0, 1$. Dále nechť $n_{i.} = n_{i0} + n_{i1}$. Potom platí

$$\widehat{\text{šance}}(0) = \frac{\frac{n_{01}}{n_{0.}}}{\frac{n_{00}}{n_{0.}}} = \frac{n_{01}}{n_{00}},$$

$$\widehat{\text{šance}}(1) = \frac{\frac{n_{11}}{n_{1.}}}{\frac{n_{10}}{n_{1.}}} = \frac{n_{11}}{n_{10}}.$$

Odhady parametrů $\hat{\boldsymbol{\beta}}$ poté jsou

$$\hat{\beta}_0 = \log \frac{n_{01}}{n_{00}}, \quad \hat{\beta}_1 = \log \frac{\widehat{\text{šance}}(1)}{\widehat{\text{šance}}(0)} = \log \frac{n_{00}n_{11}}{n_{01}n_{10}}.$$

Pokud by i, j nabývalo více hodnot, můžeme regresní parametry odhadnout například pomocí metody maximální věrohodnosti. Uvážíme-li nejjednodušší případ, kdy je náhodný výběr z alternativního rozdělení s parametrem π, budeme maximalizovat rovnici:

$$l(\pi(\mathbf{x}_i)) = \prod_{i=1}^{n} \left(\pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{n-y_i}\right) \to \max.$$

Po zlogaritmování dostáváme tzv. logaritmickou věrohodnostní funkci

$$L(\pi(\mathbf{x}_i)) = \sum_{i=1}^{n} \left(y_i \log \pi(\mathbf{x}_i) + (n-y_i)\log(1-\pi(\mathbf{x}_i))\right) \to \max.$$

Tuto funkci zderivujeme podle π a položíme rovnu nule. Tím získáváme rovnici věrohodnosti

$$\frac{\partial L}{\partial \pi(\mathbf{x}_i)} = \frac{\sum_{i=1}^{n} y_i}{\pi(\mathbf{x}_i)} - \frac{n - \sum_{i=1}^{n} y_i}{1 - \pi(\mathbf{x}_i)} = 0.$$

Jejím řešením je

$$\hat{\pi}(\mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.$$

Tedy parametr $\hat{\pi} = \bar{y}$ je maximem logaritmické věrohodnostní funkce.

Obecné odvození lze nalézt například v [1] nebo [8].

Pokud se využije toho, že platí

$$\log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \sum_{j=0}^{p} \beta_j x_{ij}, \tag{2.2}$$

$$1 - \pi(\mathbf{x}_i) = \pi(\mathbf{x}_i) e^{-\sum_{j=0}^{p} \beta_j x_{ij}} = \frac{e^{\sum_{j=0}^{p} \beta_j x_{ij}}}{1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}}} e^{-\sum_{j=0}^{p} \beta_j x_{ij}} = \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}}\right)^{-1},$$

můžeme věrohodnostní funkci upravit

$$\begin{aligned}
l(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(\pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{n - y_i} \right) \\
&= \prod_{i=1}^{n} \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)^{y_i} \left(1 - \pi(\mathbf{x}_i) \right)^{n} \\
&= \prod_{i=1}^{n} e^{\log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) y_i} \prod_{i=1}^{n} \left(1 - \pi(\mathbf{x}_i) \right)^{n} \\
&= e^{\sum_{i=1}^{n} y_i \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)} \prod_{i=1}^{n} \left(1 - \pi(\mathbf{x}_i) \right)^{n} \\
&= e^{\sum_{i=1}^{n} y_i \sum_{j=0}^{p} \beta_j x_{ij}} \prod_{i=1}^{n} \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}} \right)^{-n} \\
&= e^{\sum_{j=0}^{p} (\sum_{i=1}^{n} y_i x_{ij}) \beta_j} \prod_{i=1}^{n} \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}} \right)^{-n}.
\end{aligned}$$

Opět ji zlogaritmujeme

$$L(\boldsymbol{\beta}) = \sum_{j=0}^{p} \left(\sum_{i=1}^{n} y_i x_{ij} \right) \beta_j - \sum_{i=1}^{n} n \log \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}} \right)$$

a parciální derivace logaritmu položíme rovny nule

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{n} n x_{ij} \frac{e^{\sum_{k=0}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{p} \beta_k x_{ik}}} = 0, j = 0, ..., p.$$

S využitím toho, že výraz $\frac{e^{\sum_{k=0}^{p} \beta_k x_{ik}}}{1+e^{\sum_{k=0}^{p} \beta_k x_{ik}}}$ je roven odhadu $\hat{\pi}_i$, získáme opět normální rovnice a z nich odhad parametru $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n \hat{\pi}_i x_{ij} = 0, \qquad j = 0, 1, ..., p.$$

Pokud využijeme rovnost $\hat{\mu}_i = n\hat{\pi}_i$, můžeme rovnice zapsat jako

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \hat{\mu}_i x_{ij} = 0, \qquad j = 0, 1, ..., p.$$

Intervaly spolehlivosti a testování parametru β_j

Nyní určíme jednotlivé odhady parametru β_j. Krajní body získáme ze vztahu

$$\hat{\beta}_j \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\hat{\boldsymbol{\beta}})_{jj}},$$

kde $u_{1-\frac{\alpha}{2}}$ je $(1 - \frac{\alpha}{2})$-kvantil normálního rozdělení $\mathrm{N}(0, 1)$ a $\widehat{var}(\hat{\boldsymbol{\beta}})_{jj}$ je odhad varianční matice. Tedy přesnosti jednotlivých odhadů $\hat{\beta}_j$ určíme jako odmocninu z diagonálních prvků varianční matice. Tuto matici získáme jako inverzi z informační matice I s prvky

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g \partial \boldsymbol{\beta}_h} \right) \quad \text{pro} \quad g, h = 0, 1, ..., p.$$

Výše jsme již odvodili, že

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g} = \sum_{i=1}^{n} y_i x_{ig} - \sum_{i=1}^{n} x_{ig} \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}}.$$

Tedy nám nyní stačí provést druhou derivaci podle $\boldsymbol{\beta}_h$ opět s využitím platnosti vztahů 2.1.3.

$$\frac{\partial L^2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g \boldsymbol{\beta}_h} = -\sum_{i=1}^{n} x_{ig} x_{ih} n \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}) - e^{\sum_{k=1}^{p} \beta_k x_{ik}} e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}})^2}$$

$$= -\sum_{i=1}^{n} n x_{ig} x_{ih} \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}} \frac{1}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}} = -\sum_{i=1}^{n} x_{ig} x_{ih} n \pi_i (1 - \pi_i).$$

Vzhledem k tomu, že druhá derivace nezávisí na y_i, platí rovnost střední hodnoty informační matice a pozorované hodnoty

$$I = \mathbf{X}^T \mathrm{diag}[n\pi_i(1 - \pi_i)]\mathbf{X} = \mathbf{X}^T W \mathbf{X}.$$

Vidíme, že informační matice je rovna záporně vzatému hessiánu. Lze ji tedy získat již v průběhu iteračního procesu při hledání odhadů parametru $\boldsymbol{\beta}$. Odhad varianční matice je roven

$$\widehat{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \widehat{W} \mathbf{X})^{-1}.$$

Výhodou je, že matice W je pozitivně definitní. Pokud má matice plánu $\mathbf{X}$ plnou sloupcovou hodnost, je pozitivně definitní i informační matice I a tedy i hessián H. Ten pak lze snadno invertovat a použít v iteračním procesu.

Také lze takto testovat hypotézu $H_0 : \beta_j = 0$ pomocí Waldovy testovací statistiky

$$z = \frac{\hat{\beta}_j}{\sqrt{\widehat{var}(\hat{\boldsymbol{\beta}})_{jj}}}.$$

Druhá mocnina této testovací statistiky má asymptoticky χ^2 rozdělení o 1 stupni volnosti.

Hodnocení kvality modelu

Jednou z možností jak ověřit, zda model dobře popisuje reálná data, je použití Akaikeho informačního kritéria (AIC). Toto kritérium zohledňuje počet parametrů p v modelu a dává přednost modelům jednodušším před složitějšími. Je

dáno vztahem

$$AIC = -2[(\hat{\mu}, y) - p].$$

Pokud se dle tohoto kritéria srovnává více modelů, budeme za nejlepší považovat ten, který má nejnižší hodnotu AIC.

Pro další možnosti jak ohodnotit kvalitu modelu odkazuji například na [6].

Kapitola 3

Zobecněné lineární modely se smíšenými efekty

Doposud jsme uvažovali pouze modely s pevnými efekty, které předpokládaly, že pozorování jsou vzájemně nezávislá. Tyto modely nejsou vhodné pro analýzu některých dat, například dat ve shlucích, kde pozorování v rámci jednoho shluku jsou podobnější než pozorování ze shluků odlišných. Pozorování v rámci jednoho shluku jsou tedy obvykle pozitivně korelované.

V této kapitole budeme uvažovat model jak s pevnými, tak s náhodnými efekty. Náhodné efekty zohledňují existenci shluků a nabývají stejných hodnot v rámci jednoho shluku a odlišných hodnot pro shluky rozdílné. Tato kapitola čerpá hlavně z [1], ale také částečně z [7] a [10].

Zobecněné lineární modely se smíšenými efekty (GLMM, z angl. Generalized Linear Mixed Models) se používají pro popis dat, která nemusí pocházet z normálního rozdělení, ale z libovolného rozdělení exponenciálního typu. Data v sobě navíc zahrnují více zdrojů variability. V modelu se projeví jak pevné, tak náhodné efekty. Pevné efekty popisují vliv faktorů, které působí na všechny kategorie. Příkladem může být působení nějaké nemoci na všechny pacienty v České republice. Náhodné efekty zachycují vlivy, které působí na jednotlivé kategorie. Jako kategorie můžeme uvažovat například způsob léčby v jednotlivých nemocnicích, které danou nemoc léčí.

Nechť y_{it} je hodnota pozorování t v i-tém shluku, $t = 1, \ldots, T_i$ (v jednotlivých shlucích se počty pozorování mohou lišit). Nechť rozdělení jednotlivých složek y_{it} závisí na pevných sloupcových vektorech $\mathbf{x}_{it}$ prostřednictvím vektoru parametrů $\boldsymbol{\beta}$ a na sloupcových vektorech $\mathbf{z}_{it}$ skrze vektor náhodných efektů $\mathbf{u}_i$. Dále nechť $\mu_{it} = E(Y_{it}|\mathbf{u}_i)$. Pak rovnice zobecněného lineárního modelu se smíšenými efekty má tvar

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i, \tag{3.1}$$

kde $g(\cdot)$ je spojovací funkce a $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

Pro rozptyl platí

$$var(Y_{it}|\mathbf{u}_i) = \phi_{it}\nu(\mu_{it}),$$

kde $\nu(\cdot)$ je funkce rozptylu, která popisuje závislost rozptylu na průměru. Často se volí $\phi_{it} = 1$ nebo $\phi_{it} = \frac{\phi}{\omega_{it}}$, kde ω_{it} je známá váha (např. počet dichotomických pokusů) a ϕ je neznámý parametr rozptylu.

3.1. Zobecněné lineární modely s jedním náhodným efektem

Uvažujeme-li $z_{it} = 1$, pak dostaneme model s jedním náhodným efektem

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + u^*_i\sigma,$$

kde jsme u_i nahradili $u^*_i\sigma$, $u^*_i \sim N(0, 1)$. Tato rovnice má tvar obyčejného zobecněného lineárního modelu s neznámými hodnotami u^*_i.

Náhodné efekty někdy reprezentují náhodné chyby měření vysvětlujících proměnných. Pokud bychom regresor x_{it} nahradili $x^*_{it} + \epsilon_i$, kde x^*_{it} je správná hodnota a ϵ_i chyba měření, potom můžeme ϵ_i považovat za náhodný efekt.

Regresní parametr $\boldsymbol{\beta}$ lze odhadnout pomocí metody maximální věrohodnosti, jak jsme ukázali v kapitole 2.1.3. Nyní k alternativnímu rozdělení přidáme distribuční funkci rozdělení náhodného efektu.

Budeme uvažovat hustotu normálního rozdělení ϕ pro dané σ^2.

$$l(\pi(\mathbf{x}_i)) = \prod_{i=1}^{n} \left(\pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{n-y_i} \phi_{\sigma^2}(u_i) \right) \to \max .$$

Zlogaritmováním dostáváme logaritmickou věrohodnostní funkci

$$L(\pi(\mathbf{x}_i)) = \sum_{i=1}^{n} \left(y_i \log \pi(\mathbf{x}_i) + (n - y_i) \log(1 - \pi(\mathbf{x}_i)) + \log(\phi_{\sigma^2}(u_i)) \right) \to \max .$$

Dle [1] platí

$$\log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma,$$

$$\begin{aligned} 1 - \pi(\mathbf{x}_i) = \pi(\mathbf{x}_i) e^{-\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma} &= \frac{e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma}}{1 + e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma}} e^{-\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma} \\ &= \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma} \right)^{-1}. \end{aligned}$$

Pak funkci můžeme zapsat jako

$$\begin{aligned} l(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(\pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{n-y_i} \phi_{\sigma^2}(u_i) \right) = \\ &= \prod_{i=1}^{n} \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)^{y_i} \left(1 - \pi(\mathbf{x}_i) \right)^{n} \phi_{\sigma^2}(u_i) \\ &= \prod_{i=1}^{n} e^{\log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)^{y_i}} \prod_{i=1}^{n} \left(1 - \pi(\mathbf{x}_i) \right)^{n} \phi_{\sigma^2}(u_i) \\ &= e^{\sum_{i=1}^{n} y_i \log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)} \prod_{i=1}^{n} \left(1 - \pi(\mathbf{x}_i) \right)^{n} \phi_{\sigma^2}(u_i) \\ &= e^{\sum_{i=1}^{n} y_i \sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma} \prod_{i=1}^{n} \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma} \right)^{-n} \phi_{\sigma^2}(u_i) \\ &= e^{\sum_{j=0}^{p} (\sum_{i=1}^{n} y_i x_{ij}) \beta_j + \sum_{i=1}^{n} u_i y_i \sigma} \prod_{i=1}^{n} \left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i \sigma} \right)^{-n} \phi_{\sigma^2}(u_i). \end{aligned}$$

Opět ji zlogaritmujeme

$$L(\boldsymbol{\beta}) = \sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i x_{ij}\right)\beta_j + \sum_{i=1}^{n} u_i y_i \sigma - \sum_{i=1}^{n} n \log\left(1+e^{\sum_{j=0}^{p}\beta_j x_{ij}+u_i\sigma}\right) + \log\left(\phi_{\sigma^2}(u_i)\right)$$

a parciální derivace logaritmu položíme rovny nule

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n x_{ij} \frac{e^{\sum_{k=0}^{p}\beta_k x_{ik}+u_i\sigma}}{1+e^{\sum_{k=0}^{p}\beta_k x_{ik}+u_i\sigma}} = 0, j = 0, ..., p.$$

S využitím toho, že výraz $\frac{e^{\sum_{k=0}^{p}\beta_k x_{ik}+u_i\sigma}}{1+e^{\sum_{k=0}^{p}\beta_k x_{ik}+u_i\sigma}}$ je roven odhadu $\hat{\pi}_i$, získáme opět normální rovnice a z nich odhad parametru $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n\hat{\pi}_i x_{ij} = 0, \qquad j = 0, 1, ..., p.$$

Pokud využijeme rovnost $\hat{\mu}_i = n\hat{\pi}_i$, můžeme rovnice zapsat

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \hat{\mu}_i x_{ij} = 0, \qquad j = 0, 1, ..., p.$$

3.2. Logistické GLMM pro binární párová data

Binární párová data jsou taková data s binomických rozdělením pravděpodobnosti, která mají v jednom shluku vždy jen dvě pozorování. Pro pozorování t v i-tém shluku je v případě úspěchu $y_{it} = 1$ a v případě selhání $y_{it} = 0$, $t = 1, 2$. Pro zobecněný linerání smíšený model platí

$$\text{logit}(P(Y_{it} = 1)) = \alpha_i + \beta x_t,$$

kde $x_1 = 0$ a $x_2 = 1$.

Příkladem těchto dat mohou být výsledky pacientů, ve kterých každý shluk reprezentuje pacienta a pro každého pacienta máme dvě hodnoty lékařských testů, před a po léčbě (například má nebo nemá v těle streptokoka). Hodnoty α_i jsou pro každého pacienta odlišné a charakterizují, jak je pacient náchylný k dané nemoci,

jsou to tedy náhodné efekty. Výraz βx_t zastupuje pevný efekt, který popisuje rozdíl mezi skupinami $x_t = 0$ a $x_t = 1$. Pak můžeme testovat hypotézu, zda je $\beta = 0$, tedy zda je u lékařských testů znatelný rozdíl před a po léčbě (např. zda se výskyt streptokoka u pacientů po léčbě znatelně snížil či nikoliv).

Ekvivalentně můžeme psát

$$\text{logit}(P(Y_{i1} = 0|u_i)) = \alpha + u_i, \tag{3.2}$$

$$\text{logit}(P(Y_{i2} = 1|u_i)) = \alpha + \beta + u_i, \tag{3.3}$$

kde $u_i = \alpha_i - \alpha$ pro nějakou konstantnu α. Proměnná u_i reprezentuje náhodný efekt pro i-tý shluk a pro všechna i jsou u_i vzájemně nezávislá a mají normální rozdělení s nulovou střední hodnotou a stejným neznámým rozptylem. Při pevném u_i jsou y_{i1} a y_{i2} nezávislé.

Model popsaný rovnicemi 3.2 a 3.3 je speciální případ modelu 3.1, kde $\mu_{it} = P(Y_{it} = 1|u_i)$, $g(\cdot)$ je logit, $\boldsymbol{\beta}' = (\alpha, \beta)$, $\mathbf{x}'_{i1} = (1, 0)$ a $\mathbf{x}'_{i2} = (1, 1)$ pro všechna i a $z_{it} = 1$ pro všechna i a t. Zobecněný lineární model se smíšenými efekty v této podobě se nazývá model s náhodným absolutním členem, který místo obvyklého pevného absolutního členu α používá absolutní člen náhodný $\alpha + u_i$. Nechť $Y_1 = \sum_i y_{i1}$ a $Y_2 = \sum_i y_{i2}$ jsou marginální distribuce, kde Y_1 má binomické rozdělní s parametrem $E\left(\frac{e^{\alpha+U}}{1+e^{\alpha+U}}\right)$, Y_2 má binomické rozdělení s parametrem $E\left(\frac{e^{\alpha+\beta+U}}{1+e^{\alpha+\beta+U}}\right)$, $U \sim N(0, \sigma^2)$. Pokud je rozptyl σ^2 velký, tzn. že hustota normálního rozdělení je rozprostřena mezi mnoha hodnotami, pak u_i mohou nabývat i velkých pozitivních nebo velkých negativních hodnot. Shluk s vysokými kladnými hodnotami u_i má relativně velké pravděpodobnosti $P(Y_{it} = 1|u_i)$ pro $t = 1$ i pro $t = 2$, tzn. že s velkou pravděpodobností bude platit $(y_{i1}, y_{i2}) = (1, 1)$. Shluk s vysokými zápornými hodnotami u_i má relativně malé pravděpodobnosti $P(Y_{it} = 1|u_i)$ pro $t = 1, 2$, tzn. že s velkou pravděpodobností bude $(y_{i1}, y_{i2}) = (0, 0)$. To platí pro model popsaný rovnicí 3.2 i 3.3, ve které sice β výsledné hodnoty trochu změní, ale za předpokladu vysokých hodnot σ^2 budou platit stejné závěry, tj. pravděpodobně bude $(y_{i1}, y_{i2}) = (1, 1)$ pro shluk s vysokými kladnými hodnotami u_i a $(y_{i1}, y_{i2}) = (0, 0)$ s velmi nízkými hodnotami u_i. Pro tyto modely jsou Y_1 a

Y_2 nezávislé, pouze pokud $\sigma = 0$.

Tento model je vzácným případem toho, že pevné efekty mají v modelu s náhodnými efekty pevně danou formu odhadu maximální věrohodnosti, a to:

$$\hat{\beta} = \log\left(\frac{\hat{\mu}_{21}}{\hat{\mu}_{12}}\right).$$

Pokud pro výběrovou hodnotu logaritmu poměru šancí platí

$$\log(n_{11}n_{22}/n_{12}n_{21}) \geq 0,$$

potom $\hat{\mu}_{ab} = n_{ab}$ pro $a = 1, 2$, $b = 1, 2$ a $\hat{\beta} = \log(n_{21}/n_{12})$.

3.3. Logit-normální model pro binární data

Nyní budou představeny modely s náhodnými efekty pro binární data, které se označují jako logit-normální modely. Pro jednorozměrný náhodný efekt má model podobu:

$$\text{logit}(P(Y_{it} = 1|u_i)) = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i, \tag{3.4}$$

kde u_i jsou nezávislé náhodné veličiny, $u_i \sim N(0, \sigma^2)$. Toto je speciální případ zobecněných lineárních smíšených modelů (3.1), ve kterém je spojovací funkce logitová a náhodné efekty tvoří náhodný absolutní člen. Obecněji pak může být spojovací funkcí v modelu 3.4 inverzní funkce k libovolné distribuční funkci. Pro tyto modely při pevném u_i jsou Y_{is} a Y_{it} nezávislé. Korelace těchto veličin je nezáporná, jak dále ukážeme.

Nechť distribuční funkce Φ je inverzní funkcí ke spojovací funkci. Potom pro $s \neq t$ platí

$$cov(Y_{is}, Y_{it}) = E[cov(Y_{is}, Y_{it}|u_i)] + cov[E(Y_{is}|u_i), E(Y_{it}|u_i)] =$$

$$= 0 + cov[\Phi(\mathbf{x}'_{is}\boldsymbol{\beta} + u_i), \Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)].$$

Obě distribuční funkce Φ jsou v u_i rostoucí, a proto je korelace nezáporná.

Poznámka 3.3.1 *Pokud $\sigma = 0$, logit-normální model 3.4 se zjednodušuje na obyčejnou logistickou regresi, která uvažuje všechna pozorování nezávislá.*

Speciálním případem těchto modelů je model s náhodným absolutním členem ve tvaru:

$$logit(P(Y_{it} = 1|u_i)) = u_i + \beta_t, \tag{3.5}$$

kde u_i jsou vzájemně nezávislé, $u_i \sim N(0, \sigma^2)$. Je možné do modelu přidat intercept α za přepokladu omezení na některé β_t, např. pokud $\beta_T = 0$. Tento model je tedy jakýmsi rozšířením modelu 3.3. Zde ale předpokládáme, že v každém shluku jsou více než dvě pozorování ($T > 2$). Jeho první využití bylo v psychometrii. Pravděpodobnost $P(Y_{it} = 1|u_i)$, že i-tý pacient správně odpoví na otázku t, závisí na celkové schopnosti pacienta, která je charakterizována proměnnou u_i, a také na snadnosti otázky charakterizované proměnnou β_t. Takovéto modely se nazývají item-response modely. Jejich logitovou verzí je pak Rashův model.

Kapitola 4

Reálná aplikace - výzkum HIV

Nyní aplikujeme výše uvedenou teorii na reálnou situaci týkající se pacientů nakažených virem HIV. Nejprve představíme onen virus. Veškeré informace o této problematice byly čerpány z [3] a [11]. Následně již budeme aplikovat jednotlivé zobecněné lineární modely na reálná data. Veškeré výpočty byly prováděny v softwaru R a jsou přiloženy na CD v souboru „script_hiv.R“ a „vystupy_hiv.RData“.

4.1. Virus HIV

HIV je virus, který v organismu napadá určitou skupinu bílých krvinek, ve kterých se dále množí. Následně bílé krvinky zabíjí a snižuje tím v těle napadeného člověka jejich počet. Jejich pokles vede k selhání imunity a rozvíjí se v onemocnění AIDS. Organismus se stává náchylný k řadě dalších infekčních a nádorových onemocnění. Člověk pak může umřít třeba i na chřipku.

DNA jsou biologické makromolekuly ve tvaru dvojité šroubovice. Dvě vlákna se kolem sebe ovíjejí a jsou spojena bázemi - adeninem, thyminem, cytosinem a guaninem. Různými kombinacemi těchto bází vzniká gen. Pokud se na cytosinovou bázi DNA naváže methylová skupina, dochází k vypnutí genu, tzv. inaktivaci. Potom se na danou část genu nemůže nic vázat, a tedy nelze spustit produkci genu. Pokud je inaktivován úsek genu, kam se začlenil HIV virus, stává se virus neaktivní a přechází do spícího stavu.

Samotné methylaci se zabránit nedá, protože k ní dochází během prvních týdnů po infekci, kdy pacienti o nákaze většinou ještě ani nevědí. Vědci AV ČR hledali způsob, jak gen znovu aktivovat, aby virostatika mohla virus zničit. Našli vhodnou kombinaci látek a nyní ji testují.

Naším úkolem je z reálných dat o pacientech, kteří jsou různou dobu nakaženi HIV virem, zjistit, zda se s rostoucí délkou doby od nákazy zvyšuje výskyt těchto neaktivních vazeb.

4.2. Data HIV

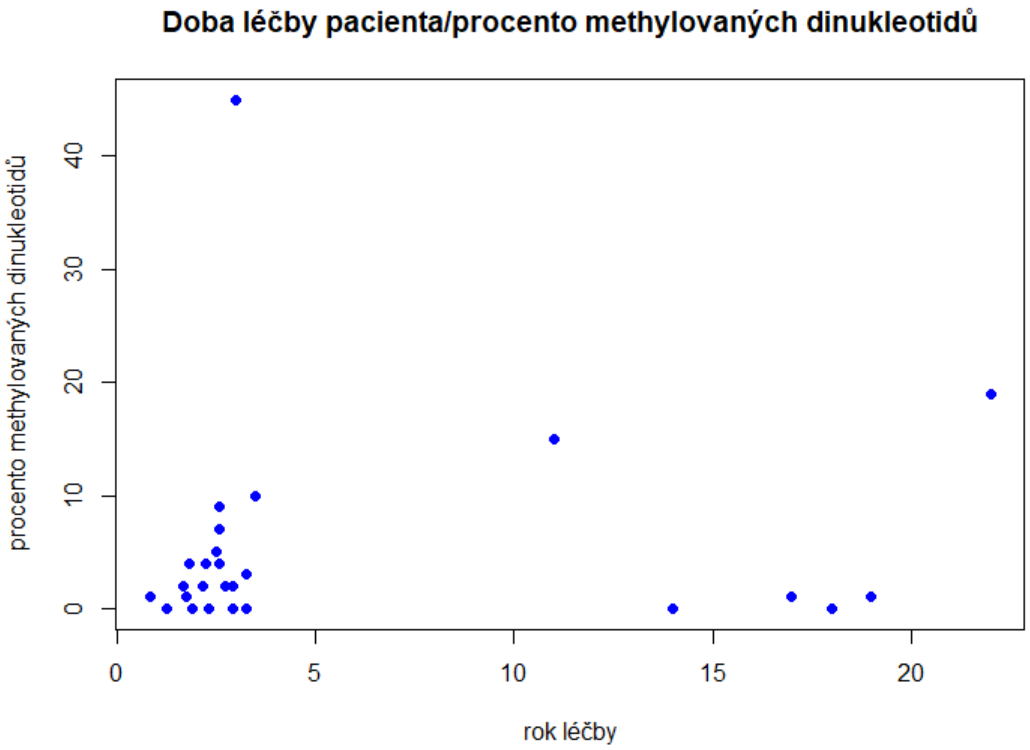
Námi analyzovaná data zahrnují informace o 26 pacientech, kteří jsou různou dobu nakaženi virem HIV. U každého pacienta známe délku trvání léčby. Dále ke každému pacientovi máme různý počet náhodně vybraných buněk a ke každé buňce hodnoty pro devět dinukleotidů. Pokud je na dané pozici jednička, znamená to, že se tam methylace vyskytuje. A naopak, pokud je tam nula, methylace se nevysktuje. Pro každého pacienta tedy máme různě velkou matici, která se skládá z jedniček a nul, viz obrázek 4.1.

12	dinukl1	dinukl2	dinukl3	dinukl4	dinukl5	dinukl6	dinukl7	dinukl8	dinukl9
buňka 1	0	0	0	0	0	0	0	0	0
buňka 2	0	0	0	0	0	0	0	0	0
buňka 3	0	0	0	0	0	0	0	0	0
buňka 4	0	0	0	0	0	0	0	0	0
buňka 5	0	0	0	0	0	0	0	0	0
buňka 6	0	0	0	0	0	0	0	0	0
buňka 7	0	0	0	0	0	0	1	0	0
buňka 8	0	0	0	0	0	1	0	0	0
buňka 9	0	0	0	0	0	1	0	0	0
buňka 10	0	0	0	0	0	1	0	0	0

Obrázek 4.1: Náhodně vybraná data o pacientovi

Každý pacient má tedy jiný počet methylovaných dinukleotidů. Maximální počet methylovaných dinukleotidů u našich pacientů je 53, ale jsou i pacienti, u kterých se nevyskytuje žádná methylace. Průměrný počet methylovaných dinukleotidů je 6,3. Pokud u pacientů nahlédneme na procentuální výskyt methylace,

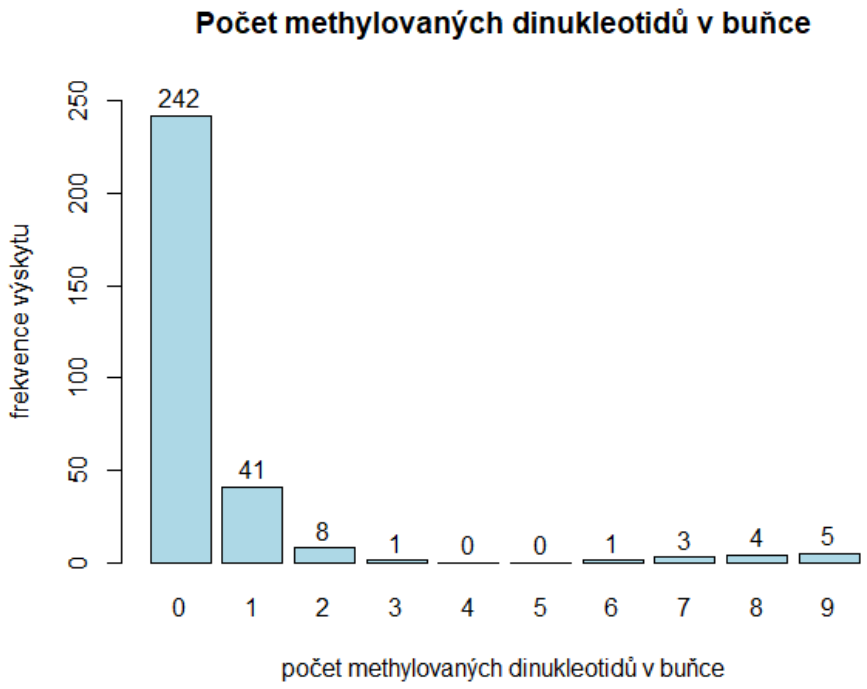
maximální hodnota je 45%, minimální 0% a průměrná 5,3%. Dále známe délku léčby pacientů. Nejkratší doba léčby u našeho vzorku pacientů je 10 měsíců, nejdelší pak 264 měsíců, tj. 22 let. Průměrná doba léčby je přibližně 69 měsíců, tedy necelých 6 let. Graf na obrázku 4.2 znázorňuje vztah mezi dobou léčení (pro lepší přehlednost je uvedena v letech) a procentem methylovaných dinukleotidů u jednotlvých pacientů. Vidíme, že nejvíce pacientů se léčí méně než 5 let a nemají



Obrázek 4.2: Procento methylovaných dinukleotidů u pacientů

příliš methylovaných dinukleotidů. Objevují se ale i výjimky jak v délce léčby, tak v počtu methylací.

U všech pacientů dohromady máme informace o 305 buňkách, které obsahují různý počet methylovaných dinukleotidů. Tyto počty znázorňuje graf na obrázku 4.3. Je patrné, že většina buněk žádné methylace neobsahuje a pokud ano, nejčastěji obsahuje pouze jeden methylovaný dinukleotid. Bohužel se u pacientů vyskytují i buňky, které mají methylovaných všech devět dinukleotidů.

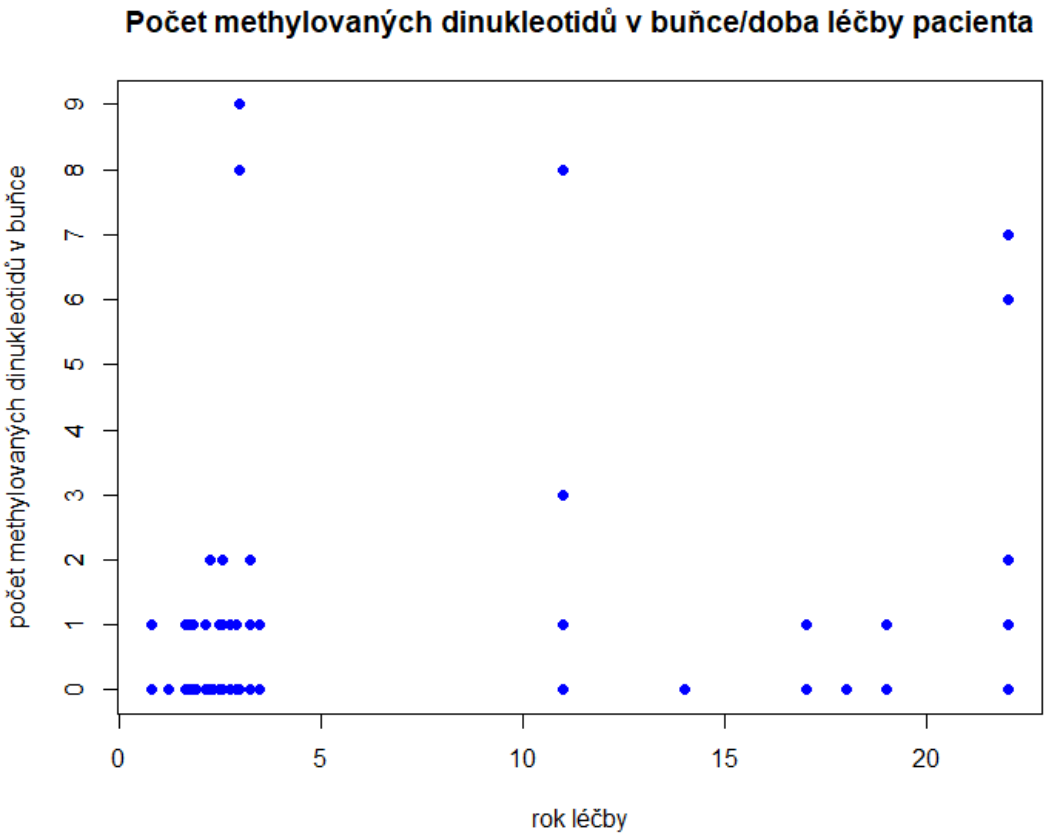


Obrázek 4.3: Počet methylovaných dinukleotidů v buňce

Vztah mezi délkou léčby pacienta a počtem methylovaných dinukleotidů v buňce znázorňuje graf na obrázku 4.4. Čekali bychom, že s rostoucí délkou léčby budou mít pacienti méně či více methylovaných dinukleotidů. Z grafu ale na první pohled žádná závislost vidět není. Zdá se tedy, že bude hodně záležet na náchylnosti pacienta nebo buněk k výskytu methylace. Tuto domněnku zkusíme ověřit matematicky.

Pokud bychom měli slovně popsat náš model, řekneme, že výskyt methylace závisí na čtyřech faktorech:

1. Pacientovi - náhodně vybereme pacienta (náhodný efekt),
2. Době léčby - každý pacient se nějakou pevně danou dobu léčí (pevný efekt),
3. Vybrané buňce - u daného pacienta náhodně vybereme některou z buněk (náhodný efekt),



Obrázek 4.4: Počet methylovaných dinukleotidů v závislosti na době léčení

4. Dinukleotidu - každá buňka pevně obsahuje devět dinukleotidů (pevný efekt).

Schématicky bychom model mohli popsat takto:

$$\text{Výskyt methylace} \sim \text{Doba léčby} + \text{Pacient} + \text{Buňka} + \text{Dinukleotid}.$$

4.2.1. Pevné efekty (GLM)

Nejprve se zkusíme úplně oprostit od náhodnosti. Nebudeme uvažovat vliv náhodných efektů - pacienta a buňky. Model se pak zúží do podoby

$$\text{Výskyt methylace} \sim \text{Doba léčby} + \text{Dinukleotid}.$$

Pro dinukleotid zavedeme pomocné proměnné (dummy variables) - identifikátory jednotlivých dinukleotidů - a budeme testovat podmodely obsahující nenulové efekty vždy jen u některých dinukleotidů. V „hlavním“ modelu (označíme ho 0) budeme uvažovat všechny dinukleotidy. A k tomuto modelu testujeme dva podmodely - nejprve zkusíme úplně vynechat proměnnou pro dinukleotid (model s označením 1) a poté vezmeme model s nejvíce odlišným dinukleotidem, což je po otestování dinukleotid číslo 5 (model s označením 2). Pokud otestujeme platnost podmodelu (1) modelu (0), dostaneme p-hodnotu 0,36. Tedy nulovou hypotézu na hladině významnosti $0,05$ nelze zamítnout a zdá se, že by se dalo pracovat jen s jednoduššším modelem (1). Po otestování platnosti podmodelu (2) modelu (0) je p-hodnota 0,73, tedy by bylo i možné místo modelu (0) pracovat s modelem (2). Otázkou je, zda se dá říci, že platí podmodel (1) modelu (2). Po otestování vyjde p-hodnota 0,036, tedy na hladině významnosti $\alpha = 0,05$ bychom hypotézu platnosti podmodelu zamítali. Ale je to poměrně blízko hranice významnosti, a tak raději tuto myšlenku ještě ověříme pomocí Akaikova informačního kritéria. Toto kritérium potvrzuje domněnku - model (2) je vhodnější než model (1). Tedy pro naše další výpočty budeme vycházet z modelu (2).

$$\text{Výskyt methylace} \sim \text{Doba léčby} + \text{Dinukleotid5}.$$

Pro všechny výpočty byl použit software R. Testy podmodelů byly provedeny pomocí příkazu anova(podmodel, model, test="Chisq"). Výpočet Akaikova informačního kritéria příkazem c(AIC(model1),AIC(model2),AIC(model0)). Toto řešení je uloženo na přiloženém CD pod názvem „vystupy_hiv.RData“.

4.2.2. Jednorozměrný smíšený efekt (GLMM)

V této části k pevným efektům přiřadíme jeden náhodný efekt, a to efekt pacienta. Můžeme si představit, že každý pacient je jinak náchylný k výskytu methylovaných vazeb. Někteří jsou náchylní více, někteří méně a průměr je zhruba uprostřed. Tedy to výstižně popisuje hustotu normálního rozdělení. Náš model

bude mít pro i-tého pacienta dle 3.4 tvar

$$\text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_1 \text{Doba léčby} + \beta_2 \text{Dinukleotid5} + u(\text{Pacient}),$$

kde u je náhodná veličina, pro kterou platí $u \sim N(0, \sigma^2)$ a $\mathbf{x}$ je vektor obsahující informace o pacientovi - jaký je to pacient, co o něm víme (délka léčby) a co u něj měříme (jaký dinukleotid), tedy $\mathbf{x} =$ (pacient, délka léčby, dinukleotid). Snažíme se odhadnout parametry tohoto modelu, a to β_0, β_1, β_2, pro každého pacienta parametr náhodného efektu, tedy 26 parametrů u a také rozptyl σ^2. Tyto odhady jsou prováděny metodou maximální věrohodnosti. Pro tuto metodu je takové množství parametrů příliš, neboť existuje velké množství možných kombinací. Budeme tedy muset použít simulaci Monte Carlo. Otázkou je, jaký počet simulací zvolit. Bohužel jsme omezeni kapacitou počítače a už při 10^3 simulací se generování provádí přibližně hodinu. Usoudili jsme tedy, že 10^3 je dostatečné množství, pokud tuto simulaci spustíme vícekrát. My jsme ji spustili 30krát. Výstupy byly rozdílné. Například výsledek první simulace byl takovýto:

$$\hat{\beta}_0 = -3,621 \ \ (0,181),$$

$$\hat{\beta}_1 = 0,002 \ \ (0,001),$$

$$\hat{\beta}_2 = -0,786 \ \ (0,374),$$

kde v závorkách jsou uvedeny hodnoty směrodatných odchylek.

Pro u jsme získali rozptyl:

$$u \sim N(0; 1,685)$$

a pro každého pacienta jinou hodnotu náhodného efektu, tedy získáváme celkem 26 hodnot:

0.1766579 −0.9259571 −0.1961500 −0.4752471 −0.1944834 −0.8865888

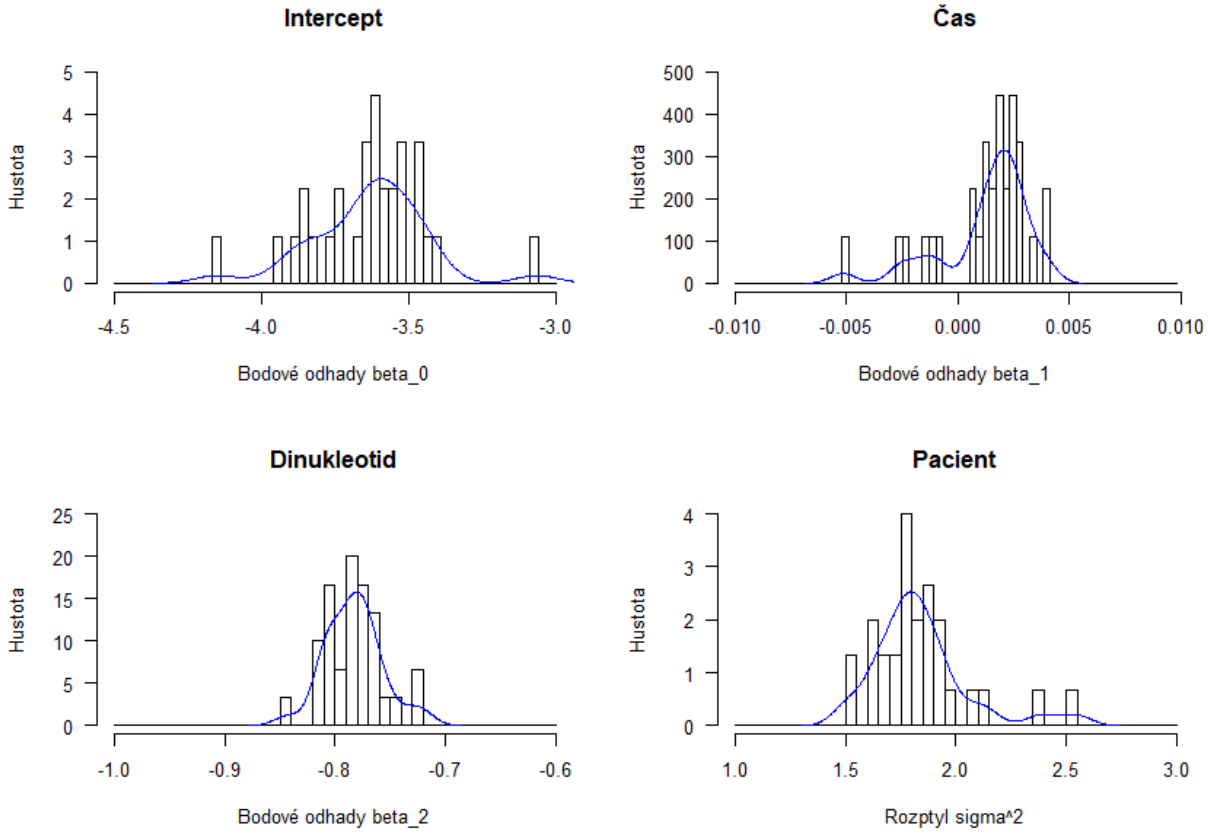
−0.2430839 0.2208474 − 0.9563604 0.9299452 0.6175557 0.3213349

$-0.1944834 \quad -0.7382298 \quad -0.3077203 \quad 3.1810302 \quad -0.7376712 \quad -0.1013790$

$1.0051489 \quad 1.5483468 \quad -1.1778042 \quad -0.6887438 \quad -1.0904121 \quad -0.5513574$

$1.6136429 \quad 0.2588970.$

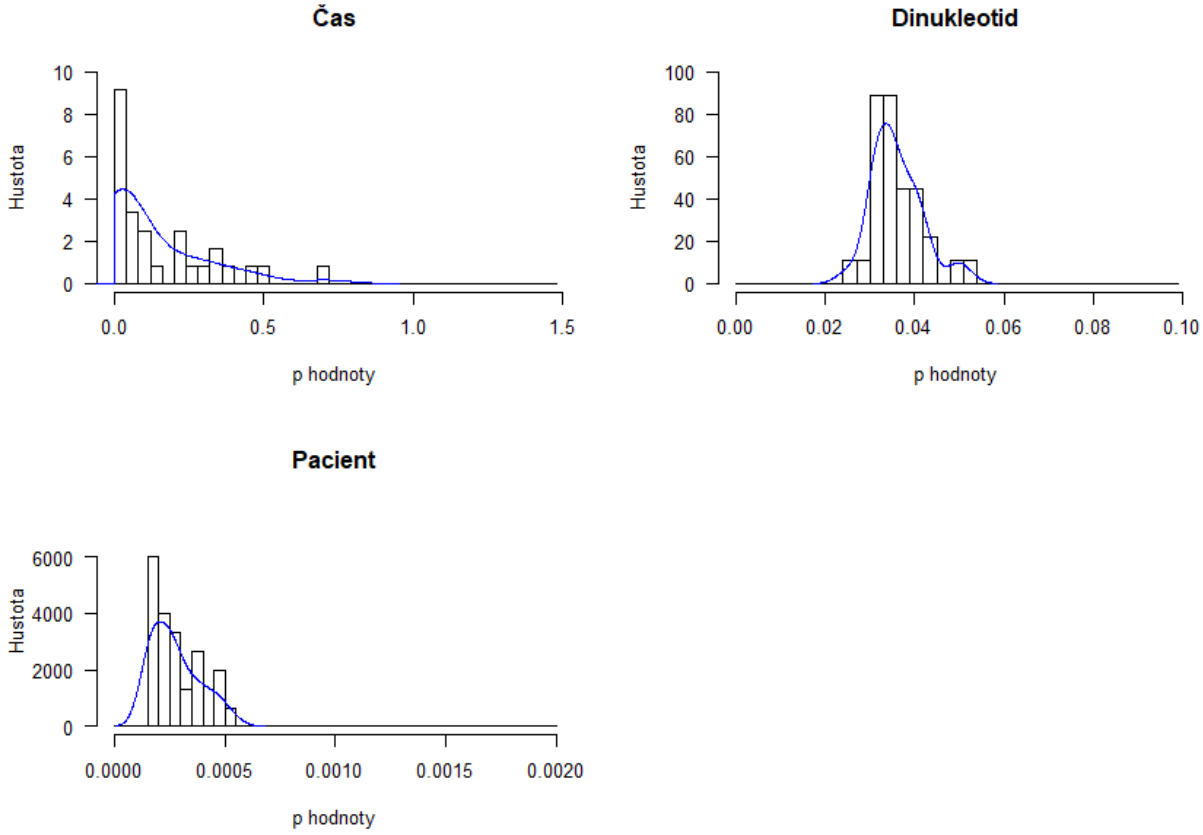
Přičemž efekt pacienta je významný ve všech 30 opakováních, ale doba léčby byla v některých iteracích významná a v některých nikoliv. Tabulky 4.7 a 4.8 obsahují p-hodnoty a bodové odhady všech 30 iterací[1]. Pro parametry β_0, β_1, β_2 a rozptyl σ^2 jsme hodnoty bodových odhadů a p-hodnot zakreslili do histogramů (viz obrázky 4.5 a 4.6). Vidíme, že bodové odhady získané při první simulaci



Obrázek 4.5: Histogramy bodových odhadů

přibližně odpovídají hodnotám bodových odhadů, které mají největší hustotu zastoupení. P-hodnota pro absolutní člen vyšla ve všech 30 opakováních menší

[1]Označením „čas“ je myšlena doba léčby.



Obrázek 4.6: Histogramy p-hodnot

než $0,0001$, tedy jsme její histogram ani nevykreslovali. P-hodnoty pro dinukleotid jsou všechny na hranici významnosti. U efektu pacienta se prokázala ve všech opakováních významnost. Problematičtější je ale doba léčby, která nabývá poměrně proměnlivých p-hodnot.

Rozhodli jsme se tedy zvýšit přesnost měření. Pro 10^4 byla doba léčby nevýznamná (p-hodnota rovna $0,435$). Obdobně pak i pro 10^5 simulací, které trvaly přibližně 72 hodin, je doba léčby nevýznamným efektem ($p = 0,69$). Získané výsledky jsou takovéto:

$$\hat{\beta}_0 = -3,908 \ \ (0,141),$$

$$\hat{\beta}_1 = 0,0004 \ \ (0,001),$$

$$\hat{\beta}_2 = -0,795 \ \ (0,374),$$

$$u \sim N(0; 2,435),$$

kde v závorkách u koeficientů $\hat{\boldsymbol{\beta}}$ jsou uvedeny hodnoty směrodatných odchylek.

Tedy na základě tohoto modelu bychom usoudili, že doba léčby vliv na výskyt methylace nemá.

V softwaru R má příkaz tvar:

$$\begin{aligned} \text{m1m} = \text{glmm}(&y \sim \text{cas} + \text{I}(\text{dinukleotid} == 5), \\ &\text{random} = \text{list}(\sim 0 + \text{as.factor}(\text{pacient})), \\ &\text{varcomps.names} = \text{c}(\text{"pacient"}), \text{data} = \text{data\_hiv}, \\ &\text{family.glmm} = \text{bernoulli.glmm}, m = 10\text{\^{}}5). \end{aligned}$$

Hodnoty pevných efektů získáme příkazem m1m$beta.pql a hodnoty náhodných efektů přikazem m1m$u.pql.

Na přiloženém CD lze nalézt postup a řešení pod názvy „script_hiv.R“ a „vystupy_hiv.RData“.

4.2.3. Vícerozměrný smíšený efekt (GLMM)

Nyní již do modelu přidáme i náhodný efekt buňky. Zobecněný lineární model se smíšenými efekty poté bude mít tvar

$$\text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_1 \text{Doba léčby} + \beta_2 \text{dinukleotid5} + u_1(\text{pacient}) + u_2(\text{bunka})$$

kde $u_1 \sim N(0, \sigma_1)$ a $u_2 \sim N(0, \sigma_2)$.

Nejprve jsme opět zkusili několikrát vygenerovat hodnoty pro 10^3 simulací. Tentokrát jsme získali jednoznačné výsledky. Významnost se potvrdila u doby léčby s p-hodnotou menší než $0,0001$ a u pacienta s p-hodnotou $0,0002$. Taktéž efekt buňky je pro hladinu významnosti $\alpha = 0,05$ významný (p-hodnota menší než $0,0001$). Efekt dinukleotidu je významný s p-hodnotou $0,007$. Pro jistotu jsme ještě provedli výpočet s 10^4 simulacemi, který již trval přibližně 5 hodin a předchozí výsledky nám potvrdil. Pouze mírně klesla významnost efektu dinukleotidu ($p = 0,01$). Pokud nahlédneme na výsledné koeficienty $\hat{\boldsymbol{\beta}}$:

$$\hat{\beta}_0 = -3,815 \ \ (0,151),$$

$$\hat{\beta}_1 = -0,002 \ \ (0,001),$$

$$\hat{\beta}_2 = -1,094 \ \ (0,441),$$

kde v závorkách jsou uvedeny hodnoty směrodatných odchylek, vidíme, že pokud se zvýší doba léčby o 40 let (480 měsíců), argument $\hat{\beta}_1$Doba léčby se sníží přibližně o jednotku. Tedy vliv doby léčby na výskyt methylace je o dost nižší než vliv ostatních pevných efektů. Ale vliv nejspíše má, tedy pro pacienty tu existuje naděje, že pokud se budou léčit dlouhou dobu, mohou se těchto methylací zbavit, a následně i viru HIV. Je třeba poukázat, že s rostoucí dobou léčby tento model ukazuje pokles výskytu methylovaných vazeb, zatímco výsledky předchozích modelů ukazovaly jeho nárůst.

Dále jsme získali rozptyly pro náhodné efekty, tedy víme, že platí:

$$u_1^{pacient} \sim N(0; 0,963),$$

$$u_2^{bunka} \sim N(0; 2,489).$$

Náhodnost pacienta i buňky se jeví jako důležité faktory pro výskyt methylace. Můžeme tedy říci, že mezi jednotlivými pacienty je velká variabilita, stejně tak i mezi jednotlivými buňkami.

V softwaru R se použije příkaz:

m2m=glmm(y ∼ cas+I(dinukleotid==5),

random=list(∼ 0+as.factor(pacient), 0+as.factor(bunka)),

varcomps.names=c("P","B"), data=data_hiv,

family.glmm = bernoulli.glmm, m = 10^4).

Hodnoty pevných a náhodných efektů se získají stejnými příkazy jako u jednorozměrného smíšeného efektu.

Na přiloženém CD lze opět nalézt postup a řešení pod názvy „script_hiv.R“ a „vystupy_hiv.RData“.

Dodatek

Další typy distribucí

V této sekci stručně popíšeme některá další rozdělení pravděpodobností náhodné veličiny. Pro detailnější popis či další možné distribuce čtenáře odkážeme na [2] a [8], odkud bylo i čerpáno.

Poissonovo rozdělení

Nechť X nabývá pouze hodnot $0, 1, \ldots$ s pravděpodobnostmi

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, \ldots,$$

kde $\lambda > 0$ je dané číslo. Pak říkáme, že X má Poissonovo rozdělení s parametrem λ, a píšeme $X \sim \mathrm{Po}(\lambda)$.

Gama rozdělení

Nechť $a > 0, p > 0$. Gama rozdělení $\mathrm{Ga}(a, p)$ má hustotu

$$f(x) = \frac{a^p}{\Gamma(p)} e^{-ax} x^{p-1}, \qquad x > 0.$$

Normální (Gaussovo) rozdělení

Nechť $\mu \in \mathbf{R}$ a $\sigma > 0$ jsou dané konstanty (parametry). Normální rozdělení je určeno hustotou

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

a označuje se symbolem $\mathrm{N}(\mu, \sigma^2)$.

Tabulky

Na závěr přidáváme tabulky s hodnotami, které jsme získali při výpočtech s více opakováními. Tyto tabulky jsou také přiloženy na CD v souborech „tab_bo.xlsx“ a „tab_p.xlsx“.

iterace	intercept	cas	dinukl	pacient_rozptyl
1	-3.621368	0.0018330	-0.7863370	1.6850
2	-3.622184	0.0015480	-0.7688350	1.8098
3	-3.937393	0.0038400	-0.7687780	1.6267
4	-3.590353	0.0013740	-0.7716550	1.6776
5	-3.529237	-0.0026751	-0.7629747	1.7585
6	-3.873891	0.0021317	-0.7712299	2.5448
7	-3.063675	-0.0051581	-0.7203305	1.8689
8	-3.739772	0.0023690	-0.7874130	1.8909
9	-3.630659	0.0012360	-0.8119900	1.9853
10	-3.624782	0.0019690	-0.8127520	1.7989
11	-3.434537	0.0018740	-0.7296710	1.9318
12	-3.639715	-0.0023160	-0.8074280	1.7711
13	-3.829393	0.0025229	-0.8022225	2.1418
14	-4.152664	0.0039001	-0.7788541	1.7053
15	-3.853925	0.0025400	-0.8111950	1.7847
16	-3.618614	0.0019080	-0.7790220	1.6225
17	-3.655256	-0.0015619	-0.8044050	1.9087
18	-3.756478	0.0032184	-0.7560649	1.9371
19	-3.845826	0.0023840	-0.8432260	1.7753
20	-3.478140	0.0027723	-0.7485933	1.8608
21	-3.476128	-0.0007963	-0.7872011	1.5062
22	-3.665340	0.0022010	-0.7811970	1.6471
23	-3.552643	0.0013970	-0.7803550	1.5136
24	-3.404267	0.0007375	-0.8061455	1.8787
25	-3.592658	0.0028560	-0.8057360	1.7880
26	-3.550120	0.0007729	-0.7663835	1.8032
27	-3.514759	0.0008621	-0.7943523	2.3846
28	-3.460582	-0.0011455	-0.7919251	2.0839
29	-3.730635	0.0026640	-0.7773820	1.8056
30	-3.520878	0.0016640	-0.7837970	1.7181

Obrázek 4.7: Tabulka bodových odhadů

iterace	intercept	cas	dinukl	pacient_rozptyl
1	2e-16	1.192e-01	0.03560	0.000308
2	2e-16	1.570e-01	0.04070	0.000370
3	2e-16	1.850e-03	0.03880	0.000210
4	2e-16	3.362e-01	0.03880	0.000459
5	2e-16	4.920e-03	0.04095	0.000159
6	2e-16	2.560e-02	0.03950	0.000164
7	2e-16	6.480e-08	0.05110	0.000156
8	2e-16	6.380e-02	0.03460	0.000323
9	2e-16	2.550e-01	0.03000	0.000358
10	2e-16	5.770e-02	0.03090	0.000247
11	2e-16	7.890e-02	0.04880	0.000531
12	2e-16	1.340e-02	0.03190	0.000163
13	2e-16	9.530e-03	0.03027	0.000172
14	2e-16	4.060e-05	0.03370	0.000162
15	2e-16	1.710e-02	0.03070	0.000231
16	2e-16	2.939e-01	0.03680	0.000454
17	2e-16	1.049e-01	0.03140	0.000158
18	2e-16	1.190e-03	0.04275	0.000267
19	2e-16	6.320e-02	0.02490	0.000260
20	2e-16	3.790e-03	0.04210	0.000205
21	2e-16	4.632e-01	0.03440	0.000239
22	2e-16	2.055e-01	0.03570	0.000383
23	2e-16	3.244e-01	0.03620	0.000292
24	2e-16	7.105e-01	0.03320	0.000237
25	2e-16	6.410e-03	0.03272	0.000278
26	2e-16	4.809e-01	0.04030	0.000252
27	2e-16	3.868e-01	0.03140	0.000176
28	2e-16	2.211e-01	0.03240	0.000158
29	2e-16	8.370e-02	0.03560	0.000471
30	2e-16	2.272e-01	0.03450	0.000397

Obrázek 4.8: Tabulka p-hodnot

Závěr

V diplomové práci jsme si nejprve představili kategoriální data a jejich nejznámější rozdělení. Druhá kapitola se věnovala regresní analýze, postupovali jsme od klasických lineárních regresních modelů přes zobecněné, až k zobecněným lineárním modelům se smíšenými efekty, kterým je věnována kapitola třetí. Čtvrtá kapitola se zabývala virem HIV. Byla zde popsána reálná data o pacientech, kteří jsou různou dobu tímto virem nakaženi. Následně jsme již na tato data aplikovali teorii zobecněných lineárních modelů.

Za použití softwaru R jsme nejprve otestovali model, ve kterém jsme neuvažovali vliv náhodných efektů. Po otestování platností jednodušších podmodelů jsme došli k závěru, že bude v následujících úvahách postačující uvažovat model, který obsahuje dobu léčby a dinukleotid 5. Tyto výpočty byly potvrzeny pomocí Akaikova informačního kritéria.

Další část reálné aplikace se věnuje zobecněným lineárním modelům s jedním náhodným efektem. Tedy jsme k pevným efektům přiřadili jeden náhodný efekt, efekt pacienta. Bodové odhady parametrů a odhad rozptylu pro náhodný efekt jsme získali pomocí metody maximální věrohodnosti prostřednictvím simulace Monte Carlo. Výpočty byly poměrně časově náročné a již pro 10^3 iterací trvaly přibližně hodinu. Spustili jsme simulaci vícekrát a na základě výsledků zjistili, že efekt pacienta je významný a efekt dinukleotidu je na hranici významnosti. Problematičtější však byla doba léčby, u které jsme získali poměrně rozdílné p-hodnoty. Zvýšili jsme tedy přesnost. Pro 10^4 i 10^5 iterací byla doba léčby nevýznamným efektem, tedy bychom na základě tohoto modelu usoudili, že doba léčby výskyt methylovaných vazeb neovlivňuje.

V závěru práce jsme se věnovali zobecněným lineárním modelům s více náhodnými efekty. Do modelu kromě náhodného efektu pacienta přibyl i náhodný efekt buňky. Opět jsme nejprve několikrát vygenerovali hodnoty pro 10^3 iterací, tentokrát s jednoznačnými výsledky. Tyto výsledky byly potvrzeny i simulací s 10^4 iteracemi, která trvala přibližně 5 hodin. Na základně tohoto modelu jsme zjistili, že jsou nejen efekty pacienta a buňky významné, ale významná je i doba léčby. Její koeficient má záporné znaménko, tedy podle tohoto modelu bychom řekli, že se s rostoucí dobou léčby snižuje výskyt methylovaných vazeb.

Tato práce pro mě byla velkým přínosem, neboť jsem si prohloubila znalosti regresní analýzy a rozšířila je o zobecněné lineární modely se smíšenými efekty. Následně jsem tyto modely použila při reálné aplikaci, a tedy se naučila, jak je využít v praxi. Také jsem díky této práci rozšířila své schopnosti při práci v softwaru R, se kterém jsem doposud pracovala jen velmi zřídka a okrajově.

Literatura

[1] AGRESTI, Alan. *Categorical data analysis.* 2nd ed. Hoboken: John Wiley & Sons, c2002. Wiley series in probability and statistics. ISBN 0-471-36093-7.

[2] ANDĚL, Jiří. *Základy matematické statistiky.* Vyd. 1. Praha: Matfyzpress, 2005. ISBN 80-86732-40-1.

[3] BORECKÁ, Kateřina. Jak na spící HIV. *Česká Televize* [online]. 2011 [cit. 2017-08-30]. Dostupné z: http://www.ceskatelevize.cz/porady/10121359557-port/tema/672-jak-na-spici-hiv/

[4] CIPRA, Tomáš. *Finanční ekonometrie.* 2., upr. vyd. Praha: Ekopress, 2013. ISBN 978-80-86929-93-4.

[5] ČERŇANSKÁ, Marie. *Probit model.* Praha, 2012. Univerzita Karlova.

[6] FAČEVICOVÁ, Kamila. *Použití logistické regrese pro diagnostiku výskytu rakoviny prostaty.* Olomouc, 2012. Diplomová práce. Univerzita Palackého v Olomouci. Vedoucí práce Mgr. Ondřej Vencálek, Ph.D.

[7] Generalized Linear Mixed Models. HEDEKER, Donald. *Encyclopedia of statistics in behavioral science.* Hoboken, N.J.: John Wiley, c2005, pp. 729–738. ISBN 0470860804.

[8] HRON, Karel; KUNDEROVÁ, Pavla. *Základy počtu pravděpodobnosti a metod matematické statistiky.* Univerzita Palackého v Olomouci, 2013.

[9] LEPŠ, Jan; ŠMILAUER, Petr. *Mnohorozměrná analýza ekologických dat.* České Budějovice, 2000. Jihočeská Univerzita v Českých Budějovicích.

[10] LEVÁKOVÁ, Marie. *Zobecněné lineární smíšené modely.* Brno, 2011. Diplomová práce. Masarykova Univerzita. Vedoucí práce RNDr. Marie Forbelská, Ph.D.

[11] O viru HIV a nemoci AIDS. *Nezkreslená věda - vzdělávací cyklus Akademie věd ČR* [online]. 2014 [cit. 2017-08-30]. Dostupné z: http://www.otevrena-veda.cz/sd/novinky/videogalerie/nezkreslena-veda/140522-nezkreslena-veda-4-dil-o-viru-hiv-a-nemoci-aids.html

[12] Typy proměnných. *Interaktivní učebnice statistiky* [online]. [cit. 2016-10-10]. Dostupné z: http://iastat.vse.cz

[13] Zpráva o výskytu a šíření HIV/AIDS za rok 2018. *Státní zdravotní ústav* [online]. [cit. 2018-04-08]. Dostupné z: http://www.szu.cz/tema/prevence/zprava-o-vyskytu-a-sireni-hiv-aids-za-rok-2018