Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

# Bakalářská práce


Filip Rozsypal

## Regresní modely pro binární veličiny

Katedra pravděpodobnosti a matematické statistiky

**Vedoucí bakalářské práce:** Mgr. Pavel Ranocha
**Studijní program:** Matematika
**Studijní plán:** Obecná matematika

2006

Rád bych poděkoval Mgr. Pavlu Ranochovi za cenné rady a připomínky k textu během psaní.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 8.8.2006

Filip Rozsypal

# Obsah

Název práce: Regresní modely pro binární veličiny
Autor: Filip Rozsypal
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Pavel Ranocha
e-mail vedoucího: ranocha@karlin.mff.cuni.cz

Abstrakt: Teoretická část popisuje motivaci a odvození zobecněných lineárních modelů (GLM). Soustředí se na logistickou regresi, ukazuje, jak odhadnout parametry pomocí metody maximální věrohodnosti, jak parametry interpretovat a jak v tomto modelu testovat hypotézy.

V druhé části je logistická regrese použita na reálná data z Evropského společenského výzkumu (ESS). Na vzorku z České republiky se snaží zjistit, jaké faktory ovlivnily ochotu lidí jít volit v parlamentních volbách v roce 2002.

Klíčová slova: Zobecněné lineární modely, logistická regrese

Title: Regresion models for binary variables
Author: Filip Rozsypal
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Pavel Ranocha
Supervisor's e-mail address: ranocha@karlin.mff.cuni.cz

Abstract: In theoretical part, this thesis shows the motivation and the derivation of generalized linear models (GLM). The logistic regresion is described in detail, with its coefficient estimation and interpretation. The methods how to test hypotheses in GLM framework are included.

In practical part, the data from European Social Survey (ESS) are examined. On the sample from the Czech Republic we ask which factors influenced someone's decision whether to vote or not in parliamentary elections in 2002.
Keywords: Generalized linear models, logistic regression

# Kapitola 1

# Úvod

Základní modely pro lineární regresi jsou formulovány pro spojité veličiny. V praxi se však často dostaneme do situace, kdy potřebujeme pracovat s veličinami, které nabývají pouze několika málo hodnot. Pro tyto případy jsou formulovány speciální metody (například pro veličiny nabývající jen dvou hodnot, tj. binární, nebo obecně kategoriální).

Kategoriální data se v praxi vyskytují velice často. Takovou proměnnou může být stav pacienta po aplikaci nového léku (např. 1 odpovídá zlepšení stavu, 0 stav beze změn) nebo počet nehod na úseku nějaké silnice za týden. Tímto způsobem se dá kódovat dosažené vzdělání jedince (1 základní, 2 střední, 3 vysokoškolské, 4 postgraduální) nebo i politická preference (např. $-1$ levicově orientován, 0 středový volič, 1 pravicový volič). Diskrétní povaha veličin vtsupuje do hry vždy, když není možné popisovanou veličinu přesně měřit nebo kvantifikovat nebo je možno měřit pouze jakýsi interval, ve kterém se hodnota nachází. Proto se tento způsob modelování uplatňuje často v sociálních vědách, které jsou ze své povahy méně exaktní. V současné době se tyto modely mimo jiné velmi často používají v bankovnictví, ale například i v zoologii, lékařství či sociologii.

V této práci se zaměříme na popis základních metod pro modelování binárních kategoriálních proměnných. Blíže představíme tzv. logistickou regresi. V druhé části pak tato metoda bude aplikována na data získaná při sociologockém sběru dat z European Social Survey (dále jen jako ESS, viz [4]).

Pro anlýzu dat použijeme program R, viz [5].

# Kapitola 2

# Teoretická část

## 2.1 Motivace

### 2.1.1 Povaha kategoriálních dat

Kategoriální data můžeme rozdělit podle toho, zda existuje nějaká škála či stupnice na seřazení jednotlivých kategorií. Příkladem situace, kdy tomu tak je, může být dosažené vzdělání (třeba když chceme zkoumat vliv dosaženého vzdělání na plat jedince). Do této skupiny spadají také data, která vznikla zdiskrétněním určité hypoteticky spojité veličiny (např. již zmíněný příjem jedince uvažovaný v kategoriích po desetitisících korun). Tyto proměnné se nazývají *ordinální*.

Druhou skupinou jsou proměnné *nominální*. U těchto veličin žádné takové přirozené seřazení „od nejmenší do největší" neexistuje. Příkladem může být rasový původ (Často se objevuje v amerických datech, např. bílý, hispánský, afroamerický. Takovéto kategoriální proměnné se používají třeba při testování rovných příležistostí a rasové diskriminace.), nebo náboženského vyznání (atheista, katolík, protestant, pravoslavný, muslim, ...).

V této práci budeme budeme sestavovat modely, jejichž vysvětlovaná proměnná bude nabývat pouze dvou hodnot. Takovýmto veličinám se říká binární, protože jejich výsledek se dá „zakódovat" pouze pomocí znaků $0 - 1$.

### 2.1.2 Rozdělení v pozadí

Uvažujme situaci, kdy $X_i$, $i = 1, \ldots, n$, je výběr z alternativního rozdělení. Tedy provádíme celkem $n$ identických nezávislých náhodných pokusů, které mají jen dva možné výstupy. Ty označme 0 pro neúspěch a 1 pro úspěch. Pravděpodobnost, že v jakémkoliv pokusu nastane úspěch, označíme $\pi$. Máme tedy

$$\forall i = 1 \ldots n : \ \mathsf{P}[X_i = 1] = \pi, \ \mathsf{P}[X_i = 0] = 1 - \pi.$$

Chceme zkoumat celkový počet úspěchů. Označíme-li $X = \sum_{i=1}^{n} X_i$, pak náhodná veličina $X$ označuje celkový počet úspěchů v $n$ pokusech. Veličina $X$ má binomické rozdělení (značíme $\mathsf{Bi}(n,\pi)$). Pravděpodobnost, že ve všech $X_i$ bude právě $x$ úspěchů (a tedy $n - x$

neúspěchů), je dána jako

$$\mathsf{P}[X = x] = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Zobecněním binomického rozdělení pro náhodné veličiny nabývající více hodnot je tzv. *multinomické rozdělení*. Zde předpokládáme, že výsledkem každého z $n$ pokusů je jeden z možných výsledků $A_j, j = 1 \ldots k$, přičemž pro každý jednotlivý pokus je pravděpodobnost, že nastane výsledek $A_j$, rovna $p_j$, tedy

$$\forall i = 1 \ldots n : \ \mathsf{P}[X_i = A_j] = p_j, \quad \sum_{j=1}^{k} p_j = 1.$$

Vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$, jehož jednotlivé složky $Y_j, j = 1, \ldots, k$, jsou definovány jako počet pokusů, kdy výsledkem byl jev $A_j$, má pak multinomické rozdělení. Potom pravděpobnost, že vektor $\boldsymbol{Y}$ nabyde hodnot $\boldsymbol{y} = (y_1, \ldots, y_k)$ je dána jako

$$\mathsf{P}[\boldsymbol{Y} = \boldsymbol{y}] = \begin{cases} \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k} & \text{když} \quad \sum_{i=1}^{n} y_i = n \\ 0 & \text{jinak.} \end{cases}$$

Binomické rozdělení je tedy případem multinomického pouze pro dva možné výsledky. Je dobré si povšimnout, že jednotlivé složky vektoru $\boldsymbol{Y}$ nejsou na sobě nezávislé. Nejlépe je to vidět na binomickém případě, počet „úspěchů“ a „neúspěchů“ je spolu svázán tak, že jejich součet musí dát dohromady celkový počet pokusů. Totéž platí samozřejmě i pro multinomické rozdělení. Dalším užitečným rozdělením je rozdělení *Poissonovo*, které má jediný parametr $\lambda > 0$.

Pravděpodobnost, že náhodná veličina $X$ mající toto rozdělení (značíme jako $X \sim \mathsf{Po}(\lambda)$) nabyde konkrétní hodnoty $k$, se rovná

$$\mathsf{P}[X = k] = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k \in \mathbb{N}.$$

Používáme ho mimo jiné v případě, kdy v pozadí tušíme určité binomické rozdělení s velmi malou pravděpodobností „úspěchu“ a velmi velkým počtem pokusů, přičemž tento počet pokusů je tak velký, že je komplikované ho přesně určit. Agresti [1] uvádí jako příklad počet smrtelných nehod za jeden den na dálnicích v Itálii. Počet lidí pohybujících se na dálnici je velmi velký, proto v podstatě neexistuje horní hranice pro počet možných úmrtí. Současně je však pravděpodobnost nehody malá. To odpovídá faktu, že Poissonovo rozdělení je limitním případem binomického. Máme-li totiž

$$X_n \sim \mathsf{Bi}(n, p_n), \quad p_n \in (0, 1),$$

a $\lim_{n\to\infty} np_n = \lambda < \infty$, potom platí

$$\lim_{n\to\infty} \mathsf{P}[X_n = k] = \mathsf{P}[X = k] \quad \text{pro} \quad k = 0, 1, \ldots \quad \text{a} \quad X \sim \mathsf{Po}(\lambda),$$

odvození například viz [3], strana 66.

### 2.1.3 Metoda maximální věrohodnosti

V dalším textu budeme používat pro odhadování parametrů *metodu maximální věrohodnosti*. Představme stručně tuto metodu.

Nechť je $\boldsymbol{X} = (X_1, \ldots, X_n)$ náhodný výběr z rozdělení, které je závislé na parametru $\theta$, který je z parametrického prostoru $\Theta$. Parametr $\theta$ může být i vícerozměrný.

Hustotu vektoru $\boldsymbol{X}$ označme jako $l(\boldsymbol{x}, \theta)$. Díváme-li se na $l(\boldsymbol{x}, \theta)$ jako na funkci parametru $\theta$, pak $l(\theta)$ nazveme *věrohodnostní funkcí*.

Hodnotu $\hat{\theta}$, která maximalizuje věrohodnostní funkci pro dané $\boldsymbol{x}$, nazveme maximálně věrohodný odhad parametru $\theta$. Jde tedy o funkci náhodného vektoru $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Uveďme si příklad výpočtu pro alternativní rozdělení rozdělení, $\mathsf{Alt}(p)$. Mějme tedy náhodný výběr $(X_1, \ldots, X_n)$ z alternativního rozdělení s neznámým parametrem $p$, který chceme odhadnout. Veličiny $X_i$ nabývají pouze hodnot 0 a 1, pravděpodobnost, že se tak stane, lze zapsat jako

$$l(x, \theta) = l(x, p) = \mathsf{P}[X = x] = p^x (1-p)^{n-x}.$$

U diskrétních rozdělení uvažujeme hustoty vzhledem k čítací míře.

Jelikož pracujeme s výběrem z nějakého rozdělení, je hustota vektoru rovna součinu marginálních hustot jednotlivých veličin. Platí tedy

$$l(\boldsymbol{x}, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{n-x_i}.$$

Často bývá výhodné pracovat s logaritmem věrohodnostní funkce, tzv. *logaritmickou věrohodnostní funkcí* $L(\theta)$. V našem případě má tvar

$$L(p) = \log\left(\prod_{i=1}^{n} p^{x_i}(1-p)^{n-x_i}\right) = \sum_{i=1}^{n} \left(x_i \log p + (1-x_i)\log(1-p)\right).$$

Logaritmus je rostoucí funkce a tak $L(\theta)$ nabývá maximum ve stejném bodě jako $l(\theta)$. Zderivujeme $L(p)$ a položíme rovno nule. Dostaneme

$$\frac{\partial L(p)}{\partial p} = \sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p}\right)$$

$$0 = \sum_{i=1}^{n} \left(\frac{(1-p)x_i - (1-x_i)p}{p(1-p)}\right) = \frac{\sum_{i=1}^{n} x_i - np}{p(1-p)}.$$

Maximálně věrohodný odhad parametru $p$ je tedy roven

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}. \tag{2.1}$$

Že se jedná o extrém a o maximum lze nahlédnout třeba z toho, že $L(\theta)$ je hladká funkce a na otevřeném intervalu $(0, 1)$ se jedná o jediný podezřelý bod. Nyní se již stačí podívat na znaménko první derivace v pravém a levém okolí podezřelého bodu.

Výsledky, které v dalším textu budeme používat, často využívají předpoklad regularity (viz [2], strana 106). Alternativní rozdělení tento požadavek splňuje.

### 2.1.4 Poměr šancí

V dalším textu se nám bude velmi hodit následující úvaha. Mějme pokus, který může dopadnout buď výsledkem 1 (úspěch) nebo 0 (neúspěch). Máme tedy náhodnou veličinu $X$ s nám už dobře známým alternativním rozdělením. Označme $\mathsf{P}[X = 1] = \pi$. Potom logicky $\mathsf{P}[X = 0] = 1 - \pi$. Pro pravděpodobnost $\pi$ definujeme *poměr šancí* (*odds ratio*) $\Omega$ jako

$$\Omega = \frac{\pi}{1 - \pi}.$$

Poměr šancí je vždy nezáporný. Pokud $\Omega > 1$, potom je úspěch pravděpodobnější než neúspěch. Je-li např. $\pi = 0{,}75$, potom $\Omega = 3$. To můžeme interpretovat tak, že na jeden neúspěšný pokus očekáváme 3 pokusy úspěšné. Obráceně platí

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

## 2.2 Zobecněné lineární modely

### 2.2.1 Lineární pravděpodobností model

Začněme jednoduchým příkladem. Máme nějaký předmět a zajímá nás, zda se poškodí, když ho pustíme na zem z určité výšky (například testování nerozbitnosti mobilních telefonů). Je na snadě předpokládat, že z čím větší výšky předmět vypustíme, tím je pravděpodobnější, že se poškodí. V případě, že bychom použili standardní lineární regresi, mohli bychom uvažovat model

$$\pi(x) = \alpha + \beta x,$$

kde $\pi(x)$ je pravděpodobnost, že se předmět rozbije, je-li puštěn z výšky $x$.

Potom parametr $\beta$ charakterizuje, jak se zvětší pravděpodobnost rozbití, zvýšíme-li $x$ o 1. Takovýto model je však často příliš zjednodušující. Jednak je modelovaná reakce na zvětšení $x$ o jednotku stejná, ať už je $x$ jakékoliv. Je však rozumné očekávat, že zvýšíme-li výšku z 1 cm na 2 cm (kdy je $\pi(x)$ ještě téměř 0), bude změna v pravděpodobnosti jiná (menší), než když se posuneme například z 30 na 31. Posuneme-li se naopak ve výšce několika metrů (kdy je $\pi(x)$ již téměř 1), můžeme očekávat opět velmi malou odezvu v $\pi(x)$.

Druhým důvodem je fakt, že pokud budeme $x$ dostatečně zvětšovat, vzroste $\pi(x)$ nad 1, což z definice pravděpodobnosti není možné. Obdobně by se nám při jiném experimentu mohlo stát, že pokud snížíme $x$ dostatečně, klesne[1] $\pi(x)$ pod 0. Na druhou stranu, pokud se omezíme na vhodný interval pro $x$, může lineární pravděpodobnostní model vcelku dobře fungovat. Jeho velkou výhodou je jednoduchá interpretace parametru $\beta$.

Musíme se tedy podívat po nějakém obecnějším rámci. Ten poskytují *zobecněné lineární modely*, budeme používat zkratku GLM pocházející z anglického *Generalized Linear Model*.

Zobecněné lineární modely nám umožní provádět regresi jednak pro širší třídu rozdělení vysvětlované proměnné. Za druhé nám pomohou vyřešit situaci, kdy střední hodnota

[1] Záleží samozřejmě na znaménku parametru $\beta$.

vysvětlované proměnné závisí na proměnných vysvětlujících nějakým nelineárním vztahem.

Každý GLM model se skládá ze tří částí: *náhodné složky*, *systematické složky* a *linkové funkce*[2] (v anglické literatuře bývají označovány jako *random component*, *systematic component* a *link function*).

### 2.2.2 Náhodná složka

*Náhodná složka* je tvořena vektorem náhodných veličin $(Y_1, \ldots, Y_n)$, který reprezentuje vysvětlovanou proměnnou. Veličiny $Y_i$ mají speciální hustotu exponeniálního typu, tj. jejich hustota lze vyjádřit ve tvaru[3]

$$f(x, \theta) = C(\theta) e^{xQ(\theta)} u(x), \tag{2.2}$$

kde $C(\theta)$, $Q(\theta)$ a $u(x)$ jsou nějaké funkce. Do této skupiny patří celá řada důležitých rozdělení a o těchto rozděleních lze dokázat mnoho zajímavých výsledků (viz [2], strana 169). Ukažme, že alternativní rozdělení patří do této rodiny. Nechť tedy $X$ má alternativní rozdělení s parametrem $p$. Potom

$$\begin{aligned} \mathsf{P}[X = x] = f(x, p) = p^x (1-p)^{1-x} &= (1-p) \left( \frac{p}{1-p} \right)^x \\ &= (1-p) \exp\left( x \log \frac{p}{1-p} \right). \end{aligned} \tag{2.3}$$

Výrazu $Q(\theta)$ z rovnice (2.2) říkáme *přirozený parametr* (*natural parametr*). Jak vidíme z výpočtu výše, v alternativním rozdělení je přirozený parametr $Q(p)$ roven

$$Q(p) = \log \frac{p}{1-p},$$

tedy *logaritmus poměru šancí úspěchu*. Tento výraz nazýváme *logit*. Ostatní členy z výrazu (2.2) mají tvar $C(p) = 1 - p$ a $u(x) = 1$. Pod pojmem logit budeme tedy rozumět

$$\mathsf{logit}(p) = \log \frac{p}{1-p}.$$

### 2.2.3 Systematická složka

Mějme matici $\boldsymbol{X}$ obsahující hodnoty vysvětlujících proměnných pro všech $n$ pozorování. Mějme celkem $m + 1$ vysvětlujících proměnných, dimenze $\boldsymbol{X}$ je tedy $n \times (m + 1)$. Jako $\boldsymbol{\beta}$ označme vektor parametrů. Potom jako *systematickou složku* GLM označme vektor

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

Pro vektor $\boldsymbol{\eta}$ tedy zřejmě platí

$$\eta_i = \sum_{j=0}^{m} \beta_j x_{ij}, \qquad i = 1, \ldots, n.$$

[2] Někdy také spojovací.

[3] Uvedená rovnice platí pro rozdělení s jedním parametrem, ale lze zobecnit i na dva parametry.

Tato lineární kombinace je označována též jako *lineární prediktor*. V situaci, kdy pracujeme jen s absolutním členem a jednou vysvětlující proměnnou, se zápis zjednoduší na

$$\eta_i = \alpha + \beta x_i.$$

Matice $\boldsymbol{X}$ má v tomto případě první sloupec tvořený samými jedničkami. Tento sloupec reprezentuje absolutní člen, intercept. V druhém sloupci jsou hodnoty vysvětlující proměnné pro jednotlivá pozorování.

### 2.2.4 Linková funkce

Poslední částí GLM modelu je *linková* funkce, označme ji $g$. Linková funkce spojuje dvě předchozí části dohromady. Označme

$$\mu_i = \mathsf{E}[Y_i], \qquad i = 1, \ldots, n.$$

Potom $\eta_i = g(\mu_i)$ a požadujeme, aby $g$ byla monotónní diferencovatelná funkce. Máme tedy

$$g(\mathsf{E}[Y_i]) = g(\mu_i) = \sum_{j=1}^{m} \beta_j x_{ij}.$$

Nejjednodušším případem linkové funkce je identita, tedy $g(\mu) = \mu$. V praxi často se vyskytujícím příkladem linkové funce je tzv. *kanonický link*. Pro ten platí

$$g(\mathsf{E}[Y_i]) = g(\mu_i) = Q(\theta_i).$$

#### Konkrétní případy GLM

Nyní vidíme, že výše uvedený *lineární pravděpodobnostní model* je speciálním případem GLM, kde náhodná složka má alternativní rozdělení a spojovací funkce je identita $g(y) = y$.

*Klasická lineární regrese* je také speciálním případem GLM. Stačí vzít za linkovou funkci identitu a uvažovat, že náhodná složka má normální rozdělení.

My se budeme zabývat *logistickou regresí*. Ta používá náhodnou složku s alternativním rozdělením a logit jako link. V případě jednoho regresoru a absolutního členu lze zapsat ve tvaru

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

Jejím zobecněním je tzv. *zobecněná logistická regrese*. Ta používá náhodnou složku s multinomickým rozdělením a logitovou linkovou funkci.

Pro vysvětlovanou veličinu mající alternativní rozdělení jsou jako linkové funkce vhodé například také kvantilové funkce absolutně spojitých rozdělení. Nabízí se tedy použít

kvantilovou funkci normálního rozdělení. Model využívající tento link se obvykle nazýva *probit*. Máme tedy

$$\pi(x) = \Phi(\alpha + \beta x),$$

kde $\Phi$ je distribuční funkce normálního rozdělení $\mathsf{N}(0, 1)$.

V bodě $-\frac{\alpha}{\beta}$ se nalézá místo, kde je nárůst v pravděpodobnosti největší a parametr $\beta$ určuje, jednak zda bude pravděpodobnost *růst* či *klesat* (kladné hodnoty $\beta$ znamenají, že s nárůstem $x$ roste i $\pi(x)$ a obráceně) a *strmost nárůstu* (čím je $\beta$ větší absolutní hodnotě, tím je nárůst/pokles strmější).

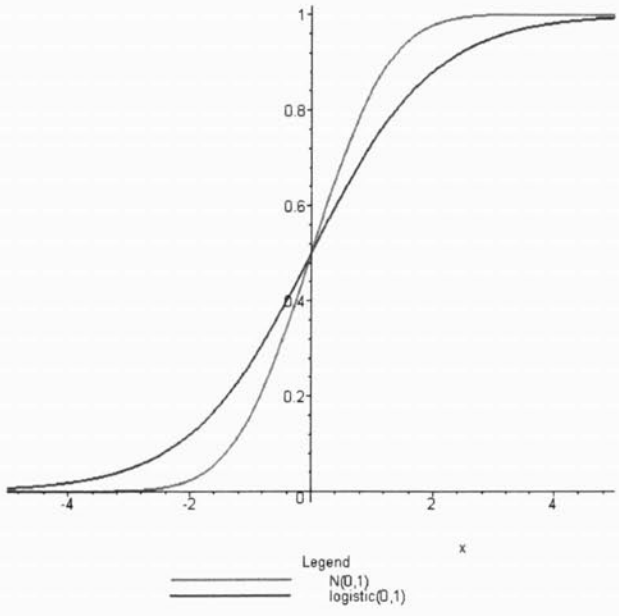
Mají-li veličiny $Y_i$ Poissonovo rozdělení s parametrem $\mu$, bývá zvykem modelovat logaritmus jejich středních hodnot:

$$\log \mu = \alpha + \beta x.$$

Logaritmus se používá mimo jiné z toho důvodu, že střední hodnota je kladná, její logaritmus pak stejně jako pravá strana může nabývat i záporných hodnot. Protože používáme logaritmus pro modelování lineární pravé strany, model se nazývá *loglineární*. Model si můžeme přepsat do exponenciálního tvaru

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}.$$

Pro srovnání logistické a normální distribuční funkce je zde obrázek 2.1.




Obrázek 2.1: Porovnání průběhu logistické a normální distribuční funkce

Je vidět, že při vhodně zvoleném rozptylu u normálního rozdělení, nebo naopak koeficietů u logistické funkce, si jsou grafy velmi podobné. V některých situacích může být vhodné použít nesymetrickou linkovou funkci. Příkladem může být tzv. *Log-Log* linková funkce (viz [1], strana 104) mající tvar

$$g(\pi(x)) = \log(-log(1 - \pi(x))).$$

Shrnutí různých GLM modelů přináší následující tabulka.

| model | náhodná složka | linková funkce |
|---|---|---|
| Lineární regrese | normální | identita |
| Lineárná pravděpodobnostní model | alternativní | identita |
| Logistická regrese | alternativní | logit |
| Zobecněná logistická regrese | multinomická | logit |
| Probit | alternativní | kvantilová normální |
| Loglinearní | Poissonovská | logaritmus |

## 2.3 Logistická regrese

Nejprve si krátce zopakujme pojmy, se kterými budeme pracovat. Náhodná složka $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ obsahuje posloupnost náhodých veličin $Y_i$ majících alternativní rozdělení s parametrem $\pi(\boldsymbol{x}_i)$ zavisejícím na $i$-tém řádku matice regresorů $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{im})$. Pro střední hodnotu alternativního rozdělení platí

$$\mathsf{E}[Y_i] = \pi(\boldsymbol{x}_i).$$

Jako link budeme používat logit

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

Všimněme si, že se jedná o kanonický link, jelikož

$$g(\mathsf{E}[Y_i]) = \log \frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} = Q(\pi(\boldsymbol{x}_i)).$$

Matice $\boldsymbol{X}$ dimenze $n \times (m+1)$ obsahuje hodnoty $m$ vysvětlujících proměnných a absolutní člen pro $n$ prozorování.

Máme tedy

$$\log \frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} = \sum_{j=0}^{m} \beta_j x_{ij}, \tag{2.4}$$

$$\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} = \exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right),$$

$$\pi(\boldsymbol{x}_i) = \frac{\exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right)}. \tag{2.5}$$

### 2.3.1 Interpretace parametrů

Jak můžeme interpretovat parametr $\boldsymbol{\beta}$? Mějme jedno pozorování, označme si ho jako $\boldsymbol{x} = (x_0, \ldots, x_m)$. Model si můžeme napsat ve tvaru

$$\Omega(x) = \frac{\pi(x)}{1-\pi(x)} = \exp\left(\sum_{j=0}^{m} \beta_j x_j\right) = \prod_{j=0}^{m} \left(e^{\beta_j}\right)^{x_j}.$$

Mějme druhé pozorování $\tilde{\boldsymbol{x}}$, které se od prvního liší pouze v $l$-té složce, a to přesně o jednotku:

$$\tilde{\boldsymbol{x}} = (x_0, \ldots, x_{l-1}, x_l + 1, x_{l+1}, \ldots, x_m) = (\tilde{x}_0, \ldots, \tilde{x}_m).$$

Jaká bude změna poměru šancí?

$$\frac{\Omega(\pi(\tilde{\boldsymbol{x}}))}{\Omega(\pi(\boldsymbol{x}))} = \frac{\frac{\pi(\tilde{\boldsymbol{x}})}{1-\pi(\tilde{\boldsymbol{x}})}}{\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}} = \frac{\exp\sum_{j=0}^{m}\beta_j\tilde{x}_j}{\exp\sum_{j=0}^{m}\beta_j x_j} = \exp\sum_{j=0}^{m}\beta_j(\tilde{x}_j - x_j) = e^{\beta_l} \tag{2.6}$$

Jednotkový nárůst v $x_l$ má tedy multiplikativní dopad velikosti $e^{\beta_l}$ na poměr šancí $\Omega$, tedy v bodě $\tilde{\boldsymbol{x}}$ je poměr šancí rovnen hodnotě v $\boldsymbol{x}$ vynásobené $e^{\beta_l}$.

Můžeme také spočítat, jaký má vliv změna jednoho regresoru přímo na pravděpodobnost $\pi(x)$. Dostaneme

$$\begin{aligned}
\frac{\partial\pi(\boldsymbol{x})}{\partial x_i} &= \frac{\partial}{\partial x_i}\frac{\exp\left(\sum_{j=0}^{m}\beta_j x_j\right)}{1+\exp\left(\sum_{j=0}^{m}\beta_j x_j\right)} \\
&= \frac{\beta_i\left(\exp\left(\sum_{j=0}^{m}\beta_j x_j\right)\left(1+\exp\left(\sum_{j=0}^{m}\beta_j x_j\right)\right) - \left(\exp\left(\sum_{j=0}^{m}\beta_j x_j\right)\right)^2\right)}{\left(1+\exp\left(\sum_{j=0}^{m}\beta_j x_j\right)\right)^2} \\
&= \beta_i\pi(\boldsymbol{x})\left(1-\pi(\boldsymbol{x})\right).
\end{aligned}$$

Můžeme si všimnout symetrie kolem bodu $\pi = 1/2$. Z chování znaménka derivace můžeme usoudit, že v bodě $\pi = 1/2$ se nachází místo, kde pravděpodobnost nejvíce roste (pokud $\beta_i > 0$).

Můžeme si také všimnout, že zafixováním všech vysvětlujících proměnných až na $i$-tou můžeme izolovat vliv změny právě $i$-té vysvětlující proměnné na proměnnou vysvětlovanou.

### 2.3.2 Odhad parametrů

Nyní odvodíme, jak vypadají maximálně věrohodné odhady parametrů $\boldsymbol{\beta}$. Jelikož jsou na sobě jednotlivé $Y_i$ nezávislé, sdružená hustota vektoru $\boldsymbol{Y}$ je rovna

$$\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} &\left(\frac{\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}\right)^{y_i} \times \\
&\times\left(1 - \frac{\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}\right)^{1-y_i}.
\end{aligned}$$

Pro snadnější zápis použijeme (2.5) a přejdeme od $\boldsymbol{\beta}$ k $\pi(\boldsymbol{x})$, čímž dostaneme

$$f(\boldsymbol{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n}\pi(\boldsymbol{x}_i)^{y_i}\left(1-\pi(\boldsymbol{x}_i)\right)^{1-y_i}. \tag{2.7}$$

Tuto sdruženou hustotu upravíme tak, aby se nám s ní lépe pracovalo. Konkrétně

$$f(\boldsymbol{y},\boldsymbol{\beta}) = \left(\prod_{i=1}^{n}(1-\pi(\boldsymbol{x}_i))\right)\left(\prod_{i=1}^{n}\exp\left(\log\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right)^{y_i}\right)\right)$$
$$= \left(\prod_{i=1}^{n}(1-\pi(\boldsymbol{x}_i))\right)\exp\left(\sum_{i=1}^{m} y_i \log\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right)\right).$$

Do poslední rovnice dosadíme do první části z (2.5) a do druhé z (2.4) a získáme

$$f(\boldsymbol{y},\boldsymbol{\beta}) = \frac{\exp\left(\sum_{i=1}^{n} y_i \sum_{j=0}^{m}\beta_j x_{ij}\right)}{\prod_{i=1}^{n}\left(1+\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)\right)}.$$

Logaritmus věrohodností rovnice je pro $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_m)$ tedy roven

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

Maximálně věrohodné odhady dostaneme, pokud tuto rovnice zderivujeme a položíme rovnu 0. Máme

$$\frac{\partial L}{\partial \beta_l} = \sum_{i=1}^{n} y_i x_{il} - \sum_{i=1}^{n} x_{il}\left(\frac{\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}\right) = 0, \quad l = 0,\ldots,m. \tag{2.8}$$

Označíme-li

$$\hat{\pi}_i = \frac{\exp\left(\sum_{j=0}^{m}\hat{\beta}_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m}\hat{\beta}_j x_{ij}\right)},$$

pak můžeme napsat věrohodnostní rovnice ve tvaru

$$\sum_{i=1}^{n} y_i x_{il} - \sum_{i=1}^{n}\hat{\pi}_i x_{il} = 0, \qquad l = 0,\ldots,k,$$

což se dá napsat maticově jako jako

$$\boldsymbol{X}'\boldsymbol{y} \;=\; \boldsymbol{X}'\hat{\boldsymbol{\pi}}.$$

Podobný výsledek bychom obdrželi pro každý GLM model používající kanonickou linkovou funkci (viz [1]). Ačkoliv výsledek vypadá jednoduše, díky vztahu (2.5) pro $\pi$ se jedná o nelineární rovnice a ty je třeba řešit většinou numericky. Nejčastěji používaná metoda je Newton-Raphsonova (viz [1], strana 114).

### 2.3.3 Testování hypotéz o parametrech modelu

#### Fisherova informace

Při hledání optimálního modelu se budeme opírat o asymptotické testy založené na maximálně věrohodných odhadech. Dá se dokázat ([2], strana 150), že za jistých předpokladů konverguje rozdělení maximálně věrohodných odhadů k normálnímu rozdělení s rozptylem závisejícím na Fisherově informační matici.

Spočtěme tedy nejdříve Fischerovu infomační matici $\boldsymbol{J}(\boldsymbol{\beta})$ v modelu logistické regrese. Tato matice je dimenze $(m+1) \times (m+1)$. Konkrétní políčko $(i, j)$ Fisherovy informační matice se dá spočítat jako (viz [2], strana 121)

$$J_{ij}(\boldsymbol{\beta}) = -\mathsf{E}\left[\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right].$$

Během odvozování maximálně věrohodných odhadů jsme v (2.8) již dokázali, že

$$\frac{\partial L}{\partial\beta_l} = \sum_{i=1}^{n} y_i x_{il} - \sum_{i=1}^{n} x_{il}\left(\frac{\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}\right).$$

Další derivace je pak rovna

$$\frac{\partial^2 L}{\partial\beta_l\beta_k} = -\sum_{i=1}^{n} x_{il}x_{ik}\frac{\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)}{\left(1+\exp\left(\sum_{j=0}^{m}\beta_j x_{ij}\right)\right)^2},$$

což stejně jako dříve upravíme na

$$\frac{\partial^2 L}{\partial\beta_l\beta_k} = -\sum_{i=1}^{n} x_{il}x_{ik}\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i)).$$

Posledním krokem je vypočtení střední hodnoty. V tomto výrazu již ale nefiguruje žádná náhodná veličina, a tak platí

$$J_{lk}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{il}x_{ik}\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i)). \tag{2.9}$$

#### Test poměrem věrohodnosti

Předpokládejme, že platí model s parametry z určitého parametrického prostoru $\Theta_1$. Chceme testovat, zda je možné přejít k submodelu s parametry pouze z vlastního podprostoru $\Theta_2$ původního parametrického prostoru $\Theta_1$. Mějme tedy maximálně věrohodný odhad parametrů $\hat{\beta}_1$ v původním prostoru a $\hat{\beta}_2$ v podprostoru $\Theta_2$. Jako nulovou hypotézu máme, že je možné přejít k jednodušímu modelu, tj. koeficienty u vybraných vysvětlujících proměnných jsou rovny nule (nenulové určují podprostor původního parametrického prostoru).

Test *poměrem věrohodností* používá statistiku

$$LR(\hat{\beta}_2, \hat{\beta}_1) = -2(L(\hat{\beta}_2) - L(\hat{\beta}_1)),$$

kde $L$ jsou hodnoty logaritmické věrohodnostní funkce příslušného vektoru parametrů. Tato statistika má asymptoticky $\chi^2$ s počtem stupňů volnosti rovným rozdílu dimenzí původního parametrického prostoru a nového podprostoru. Označme tento rozdíl jako $q$. Nulovou hypotézu tedy zamítáme na hladině $\alpha$ (a tedy nelze přejít k jednodušímu submodelu), když $LR(\hat{\beta}_2, \hat{\beta}_1) \geq \chi^2_q(\alpha)$.

### Waldův test

Mějme model s parametry $\boldsymbol{\beta} = (\beta_m, \ldots, \beta_m)$. Testujeme hypotézu $H_0$: $\boldsymbol{\beta} = \boldsymbol{\beta_0}$ proti alternativě $H_1$: $\boldsymbol{\beta} \neq \boldsymbol{\beta_0}$. Označme dále $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_m)$ odhady získané metodou maximální věrohodnosti. Potom statistika

$$W(\boldsymbol{\beta_0}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})\boldsymbol{J}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{2.10}$$

má asymptoticky $\chi^2$ rozdělení s $m + 1$ stupni volnosti.

Při nulové hypotéze o jednorozměrném paramteru $H_0$: $\beta = \beta_0$ přejde testová statistika na

$$W(\beta) = \frac{(\hat{\beta} - \beta_0)^2}{\sigma^2}, \qquad \sigma^2 = 1/J(\hat{\beta}).$$

Pro určování signifikance parametrů se používá následující modifikace. Dá se ukázat, že za platnosti hypotézy $H_0$: $\beta_0 = 0$ má statistika

$$z = \sqrt{W(\beta)} = \frac{\hat{\beta}}{\sigma}$$

asymptoticky normální rozdělení N(0,1).

Chceme-li testovat hypotézu jen o určitém podvektoru $\boldsymbol{\tau}$ vektoru $\boldsymbol{\beta} = (\boldsymbol{\tau}, \boldsymbol{\phi})$, stále potřebujeme odhadnout celé $\boldsymbol{\beta}$. Vektor $\boldsymbol{\phi}$ tvoří tzv. *rušivý parametr*. Označme si

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_r), \qquad \boldsymbol{\phi} = (\phi_1, \ldots, \phi_s), \qquad \text{přičemž} \quad r + s = m + 1.$$

Fisherovu informační matici rozdělíme na bloky příslušející subvektorům $\boldsymbol{\tau}$ a $\boldsymbol{\phi}$

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{J}_{11} & \boldsymbol{J}_{12} \\ \boldsymbol{J}_{21} & \boldsymbol{J}_{22} \end{pmatrix}.$$

Požadujeme, aby bloky $\boldsymbol{J}_{11}$ a $\boldsymbol{J}_{22}$ byly čtvercové a regulární. Symbolem $\boldsymbol{J}_{11.2}$ označíme

$$\boldsymbol{J}_{11.2} = \boldsymbol{J}_{11} - \boldsymbol{J}_{12}\boldsymbol{J}_{22}^{-1}\boldsymbol{J}_{21}.$$

V této situaci je třeba modifikovat testovou statistiku pro $W$. Použijeme značení z [2]. Maximálně věrohodný odhad, který není vázán žádnými dalšími podmínkami, označíme jako

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\tau}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix}.$$

Maximálně věrohodný odhad, který je omezen hypotézou $H_0^*$: $\boldsymbol{\tau} = \boldsymbol{\tau}_0$, označíme jako

$$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{\tau}_0 \\ \tilde{\boldsymbol{\phi}} \end{pmatrix}.$$

V případě testu s rušivými parametry přejde statistika (2.10) na

$$W^*(\boldsymbol{\beta_0}) = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0)\boldsymbol{J}_{11.2}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0).$$

Statistika $W^*(\boldsymbol{\beta_0})$ má potom asymptoticky $\chi^2_r$ rozdělení.

### Raův skórový test

Použijeme-li značení z předcházejícího odstavce, pak Raův skórový test je založen na statistice

$$LM(\boldsymbol{\beta_0}) = [\boldsymbol{U}(\boldsymbol{\beta}_0)]'[\boldsymbol{J}(\boldsymbol{\beta}_0)]^{-1}[\boldsymbol{U}(\boldsymbol{\beta}_0)],$$

kde

$$\boldsymbol{U}(\boldsymbol{\beta}_0) = \left(\frac{\partial L}{\beta_0} \ldots, \frac{\partial L}{\beta_m}\right),$$

a víme, že

$$\frac{\partial L}{\partial \beta_l} = \sum_{i=1}^{n} y_i x_{il} - \sum_{i=1}^{n} x_{il} \left( \frac{\exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right)} \right).$$

Statistika $LM(\boldsymbol{\beta}_0)$ má při $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ asymptoticky $\chi^2$ rozdělení s $m+1$ stupni volnosti. Modifikovaná statistika pro test s rušivými parametry má tvar

$$LM^*(\boldsymbol{\beta_0}) = [\boldsymbol{U}_1(\tilde{\boldsymbol{\beta}})]'[\boldsymbol{J}_{11.2}(\tilde{\boldsymbol{\beta}})]^{-1}[\boldsymbol{U}_1(\tilde{\boldsymbol{\beta}})], \tag{2.11}$$

kde $\boldsymbol{U}_1$ je podvektor vektoru $\boldsymbol{U}$ odpovídající vektoru $\boldsymbol{\tau}$. Statistika $LM^*(\boldsymbol{\beta_0})$ má pak asymptoticky $\chi^2_r$ rozdělení.

Tyto testy a jejich odvození lze nalézt v [2], strana 151, 176, testy s rušivými parametry pak na stranách 183 až 186.

Je-li k dispozici pro testování jisté hypotézy více různých testů, můžeme se někdy setkat s následují chybou. Nejprve se vypočtou hodnoty všech statistik a porovnají se s příslušnými kritickými hodnotami. Jedna ze statistik překročí kritickou hodnotu. Chyby se dopustíme, když nyní vybereme tu statistiku, která překročila kritickou hodnotu a na jejím základě konstatujeme, že hypotéza neplatí.

Chyby se dopouštíme proto, že ve skutečnosti používáme maximum ze statistik. Maximum pak již obecně nemá stejné rozdělení jako jednotlivé statistiky a tedy nemůžeme rozhodnout na původní kritické hladině.

V aplikační části této práce budeme používat Waldův test pomocí statistiky $z$ a to z důvodu její implementace do softwaru, který jsme při zpracování dat použili.

### 2.3.4 Postup při sestavování modelu

#### Diskrétní vysvětlující proměnné

Předpokládejme, že chceme do modelu zahrnout diskrétní proměnnou. Nejdříve se musíme rozhodnout, zda se jedná o ordinální nebo nominální diskrétní proměnnou, viz sekce 2.1.1.

Pokud se jedná o ordinální, musíme se rozhodnout, jak budeme proměnnou škálovat. Přirozeně, jednotlivé hodnoty musí být seřazeny podle své ordinální stupnice (například vzdělání od základního až po postgraduální). Jednotlivým hladinám však můžeme přiřadit různé koeficienty a tak zohlednit, že hladiny nemusí být „ekvidistatní".

Příkladem může být modelování nezaměstanosti. Jako jednu z vysvětlujících proměnných máme vzdělání respondenta. Nechť nám dotazník uvádí jen 3 hodnoty: bez základního, vysokoškolské a postgraduální vzdělání. Víme, že mezi lidmi bez základního vzdělání je obrovská nezaměstnanost. Při přechodu z hladiny bez *základního vzdělání* do kategorie *vysokoškolák* je však tento skok určitě vyšší, než když se člověk při přechodu z *vysokoškolák* do *postgraduálně vzdělaný*. Z tohoto pohledu může být výhodné proměnnou vzdělání kódovat třeba čísly (1, 5, 6) a zdůraznit tak přechody mezi jednotlivými hladinami vzdělání.

V našem modelu jsme nezaznemanali potřebu zdůrazňovat rozdíly mezi hladinami u vysvětlujících proměnných. Všude je tak použito ekvidistatní kódování.

U nominálních diskrétních proměnných musíme postupovat jinak. Dejme tomu, že máme máme proměnnou, která značí, že respondent běloch, černoch, indián nebo eskymák. Bude nás zajímat, jak se jednotlivé skupiny obyvatel v závislosti na svém rasovém původu liší v přístupu k práci v České republice.

Protože je u nás drtivá převaha bělochů, budeme za referenční považovat hodnotu *běloch* a budeme ostatní rasy vyhodnocovat v porovnání s bělochy. Pro každou další rasu vytvoříme novou tzv. dummy proměnnou, která bude nabývat hodnoty 1, když je respondent právě té konkrétní rasy a 0, když je jiné rasy. Parametry u jednotlivých dummy proměnných pak ukazují, jak je na tom konkrétní skupina obyvatel v porovnání se skupinou, kterou jsme zvolili jako referenční.

#### Výběr optimálního modelu

V případě, že máme více vysvětlujících proměnných, nastává problém, jak určit nejlepší model. Naše požadavky jsou stejné, jako ve standardní regresi. Požadujeme, aby výsledný model dostatečně přesně vysvětloval naše data, ale současně, aby nebyl počet vysvětlujících proměnných příliš vysoký. Menší počet proměnných také skýtá výhody v jednodušší interpretaci výsledků. Větší počet vysvětlujících proměnných přináší automaticky také mnohem větší počet vzájemných interakcí.

V zásadě existují dva motivy pro statistická šetření. Za prvé, potvrzovací studie mají za cíl potvrdit či vyvrátit jistý vliv. V tomto případě bývá výhodnější sestavovat model jednodušeji, porovnáváme model s tímto vlivem (proměnnou) a bez něho. Druhou situací je případ, kdy nemáme jasný teoretický základ, zkoumáme větší množství proměnných a hledáme, zda nějaké nemají náhodou na vysvětlovanou proměnnou vliv.

Někdy je možné získat základní orientaci znázorněním do grafu (pro spojité prediktory) nebo kontigenční tabulkou (pro diskrétní prediktory). Obecně neexistuje vždy nejlepší

metoda, jak vybírat či naopak vyřazovat z modelu proměnné. V praxi se nejvíce používají dva přístupy: *forward selection* a *backward elimination.*

První spočívá v tom, že začínáme s modelem obsahujícím pouze absolutní člen. Do tohoto minimálního modelu pak přidáváme postupně další proměnné. Tak pokračujeme až do okamžiku, kdy se přidáním další proměnné již nezvýší kvalita modelu. Oproti tomu zpětná eliminace funguje přesně obráceně. Začíná s modelem s co nejvíce proměnnými a v každém kroku odstraní proměnnou, která nejméně přispívá ke schopnosti modelu vysvětlovat data. Proces končí v situaci, kdy by vyřazení kterékoliv další proměnné vedlo k podstatně horšímu modelu.

Pro rozhodování se používají testy uvedené v sekci 2.3.3. Jak tedy v praxi postupujeme? Začínáme se všemi proměnnými, které považujeme za relevantní. Test poměrem věrohodnosti má nulovou hypotézu, že lze přejít k jednodušímu submodelu, tedy že lze vynechat nějakou vysvětlující proměnnou. Pokud pomocí Waldova testu nezamítneme nulovost nějakého parametru, můžeme, stejně jako v případě LR testu, vyřadit tuto vysvětlujicí proměnnou z modelu. Podobně pro Raův test.

Signifikanci testujeme následujícím způsobem. Mějme hypotézu $H_0 : \beta_k = 0$, přičemž ostatní $\beta_j$ jsou brány jako rušivé parametry. Můžeme použít libovolný test z části 2.3.3. Jako nesignifikantní označíme ty proměnné, u kterých příslušný test nezamítá nulovou hypotézu na hladině $\alpha = 0{,}05$.

Software často uvádí $p$-value, což je nejmenší hladina, na které test nezamítá $H_0$. Jako nesignifikanntí tedy označíme ty proměnné, u kterých vyšla $p$-value větší než 0,05.

Pokud jsou v modelu nějaké nesignifikantní proměnné, obyvykle se vyřadí ta s nejvyšší $p$-value. Tato procedura se opakuje až do té doby, kdy jsou všechny parametry signifikantní. Jinými slovy, nejde přejít k jednoduššímu submodelu.

V praxi je dávána přednost spíše zpětné eliminaci, díky její větší přímočarosti. Metody nemusí vést ke stejnému modelu a výsledkem není vždy smysluplný model (viz [1], strana 214). Problémy nastávají zejména v situacích, kdy jsou vysvětlující proměnné navzájem korelovány.

# Kapitola 3

# Aplikační část

## 3.1 Data

European Social Survey je celoevropský sociologický průzkum. Dává si za cíl zaznamenat a vysvětlit, jakým způsobem ovlivňuje měnící se prostředí postoje a chování lidí. V současné době na něm participuje celkem 26 zemí, mimo jiné i Česká republika.

Pro tuto práci jsme použili data z druhého kola šetření a pouze pro Českou republiku. Průzkum se konal během roku 2004. Průzkumníci se ptali respondentů na zhruba 330 otázek. Data jsou volně dostupná na stránkách ESS, viz [11].

### 3.1.1 Vybrané proměnné

Výběr možných proměnných byl opravdu široký. Hledali jsme nějakou vhodnou proměnou nabývající jen binárních hodnot. Nakonec byla vybrána proměnná vote, která indikuje, zda respondent volil v posledních celonárodních volbách. V našem případě se jedná o volby do Poslanecké sněmovny Parlamentu České republiky v roce 2002.

Většina proměnných je z povahy dotazníkového šetření diskrétní. Nominální diskrétní proměnné zakódujeme pomocí dummy proměnných, viz odstavec 2.3.4.

Jako vysvětlující proměnné jsme vybrali:

- mmbprty, která ukazuje, zda je respondent členem nějaké politické strany. Tato proměnná je také pouze binární. Její kódování je 0: respondent není členem žádné politické strany, 1: je členem.
- lrscale zaznamenává politickou orientaci tázaného. Škála od 0 do 10 po 1 odpovídá rozmezí od extrémní levice až do extémní pravice. Tuto diskrétní proměnnou budeme považovat za ordinální.
- Proměnná happy ukazuje, jak se respondent cítí šťastný. Kódování je od 0 do 10 po 1 od extrémně nešťastných do extrémně šťastných. Tuto proměnnou považujeme za ordinální.
- rlgdgr se ptá, jak moc je respondent nábožensky založen. Škála od 0 do 10 po 1 odpovídá rozmezí od naprostého atheisty do velice nábožného člověka. Tuto diskrétní proměnnou považovažujeme za ordinální.

- Proměnná mnyavth zaznamenává, zda respondent souhlasí s tvrzením, že „když chce člověk vydělat peníze, nemůže se vždy chovat poctivě". Na škále od 1 do 5 po 1 značí čím vyšší hodnota, tím větší nesouhlas s uvedeným tvrzením. Tuto diskrétní proměnnou budeme považovat za ordinální.
- gndr zaznemenává pohlaví tázaného. Jedná se o binární veličinu. Muž je označen jako 0, žena 1.
- yrbrn zaznemenává rok narození tázaného. Toto je téměř spojitá proměnná.
- domicil ukazuje, v jak velkém sídle či obci respondent bydlí (velké město až samota). Na škále od 1 do 5 po 1 značí čím vyšší hodnota, tím menší sídlo.
- hhmodwl se ptá, jestli respondent bydlí ve vlastním domě či bytě. 1 značí ano, 0 že respondent nebydlí ve vlastním domě či bytě. Toto je pouze binární proměnná.
- rmhhus zaznamenává, kolik pokojů má byt, kde tázaný bydlí.
- edlvcz zachycuje nejvyšší dosažené vzdělání respondenta. Škála od 0 do 10 po 1 odpovídá rozmezí od nedokončeného základního 0 až do postgraduálního vzdělání 10. Jedná se o ordninální diskrétní proměnnou.
- Proměnná marital obsahuje současný rodinný stav tázaného. Toto je nominální diskrétní proměnná, a tak její hodnoty překódujeme do dummy proměnných, viz sekce 2.3.4. Protože původní proměnná nabývala celkem pěti hodnot, nové dummy proměnné budou celkem čtyři. V marital je o respondentovi obsaženo, zda je
  1. ženatý nebo vdaná. Tuto hodnotu bude považovat za základní, jelikož je v datech nejvíce zastoupena (asi 56%). Nebudeme pro ni tedy vytvářet žádnou dummy proměnnou,
  2. žije odděleně, ale je stále ženatý nebo vdaná. Dummy proměnná d.sep je v tomto případě 1, v jiných případech je rovna 0,
  3. rozvedený(á), d.div je 1, jindy je tato proměnná rovna 0,
  4. vdovec nebo vdova, d.wid je 1, jindy je tato proměnná rovna 0,
  5. nikdy nebyl ženatý či vdaná. Příslušná dummy proměnná se nazývá d.nmar. Je 1, když respondent nikdy nebyl ženatý nebo respondentka vdaná. Jindy je tato proměnná rovna 0.
- yrspdwk zachycuje, kolik let strávil tazáný v pracovním poměru. Tuto proměnnou budeme považovat za spojitou.

### 3.1.2 Celkový popis dat

Data pro Českou republiku obsahovala celkem 3027 pozorování. Tato data jsme však museli upravit. Všechny výše uvedené proměnné mají kromě uvedených možností odpovědí ještě možnosti *refusal*, *do not know* a *no answer*. Všechna pozorování, kde se v nějaké

proměnné nějaká z těchto odpovědí vyskytla, jsme z datového souboru odstranili. Počet pozorování po tomto očištění je roven 1268. Takovýto počet pozorování je celkem vysoký a měl by nám zaručit dobré chování asymtotických testů.

## 3.2 Hledání optimálního modelu

Při uvádění tabulek výsledků nám pro přehlednost použe následující značení. Proměnné, které by vyšly signifikantní na hladině $\alpha = 0{,}001$, označíme symbolem $***$, pro hladinu $\alpha = 0{,}01$ použijeme $**$, pro hladinu $\alpha = 0{,}05$ použijeme $*$ a pro $\alpha = 0{,}1$ použijeme tečku.

### 3.2.1 Maximální model

Začínáme s modelem obsahujícím všechny vybrané proměnné. Software nám dává výstup, který shrnujeme v tabulce 3.1.

| Proměnná | Parametr | Směr. odchylka | $z$ | $Pr(>\|z\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 36,160181 | 18,775058 | 1,926 | 0,054108 | . |
| mmbprty | 2,716097 | 0,737083 | 3,685 | 0,000229 | *** |
| lrscale | 0,033604 | 0,027950 | 1,202 | 0,229260 | |
| happy | 0,127622 | 0,034673 | 3,681 | 0,000233 | *** |
| rlgdgr | 0,029813 | 0,023924 | 1,246 | 0,212703 | |
| mnyacth | 0,169644 | 0,054577 | 3,108 | 0,001881 | ** |
| yrbrn | -0,019868 | 0,009486 | -2,094 | 0,036220 | * |
| domicil | 0,156115 | 0,063699 | 2,451 | 0,014253 | * |
| rmhhus | 0,100991 | 0,060571 | 1,667 | 0,095451 | . |
| edlvcz | 0,153381 | 0,030824 | 4,976 | 6,49e-07 | *** |
| yrspdwk | 0,017421 | 0,010447 | 1,668 | 0,095402 | . |
| gndr | -0,047917 | 0,133524 | -0,359 | 0,719695 | |
| hhmodwl | -0,110865 | 0,142791 | -0,776 | 0,437505 | |
| d.nmar | -0,226216 | 0,204091 | -1,108 | 0,267687 | |
| d.sep | 0,233397 | 0,418215 | 0,558 | 0,576791 | |
| d.div | -0,314046 | 0,207277 | -1,515 | 0,129746 | |
| d.wid | -0,598355 | 0,208547 | -2,869 | 0,004116 | ** |
| hinctnt | -0,019689 | 0,040274 | -0,489 | 0,624938 | |

Tabulka 3.1: Maximální model

Když se podíváme do posledního sloupce, který obsahuje jednotlivé $p$-value, vidíme, že máme v modelu několik velice signifikantních a několik velice nesignifikantních proměnných. Dále budeme pokračovat podle postupu, který jsme naznačili v sekci 2.3.4.

### 3.2.2 Eliminace nesignifikantních proměnných

Vybereme tedy proměnnou s největší $p$-value. To je d.div s $p$-value 0,733818. Proceduru odstraňování nesignifikantních proměnných několikrát zopakujeme. Postup shrňme v tabulce 3.2. Celkem jsme z původních sedmnácti vysvětlujících proměnných devět vyřadili

při zpětné eliminaci. V modelu nám zůstalo 8 proměnných, které jsou již signifikantní na hladině $\alpha = 0{,}05$, téměř polovina dokonce na hladině $\alpha = 0{,}001$. Tyto proměnné jsou tedy vysoce signifikanní. Následující tabulka ukazuje, jak tato procedura běžela. Ve sloupci *Signifikance* je uvedena $p$-value, se kterou byla proměnná z modelu vyřazena.

| Krok | Proměnná | Signifikance |
|---|---|---|
| 1. | gndr | 0,719695 |
| 2. | hinctnt | 0,631110 |
| 3. | d.sep | 0,575961 |
| 4. | hhmodwl | 0,417002 |
| 5. | d.nmar | 0,263536 |
| 6. | lrscale | 0,214991 |
| 7. | d.div | 0,158273 |
| 8. | rlgdgr | 0,158080 |
| 9. | yrspdwk | 0,12684 |

Tabulka 3.2: Postupné vyřazování nesignifikantních proměnných

### 3.2.3 Optimální model

Po devíti krocích se dostáváme do situace, kdy jsou již všechny proměnné na hladině $\alpha = 0{,}05$ signifikantní. Výsledný model shrnuje tabulka 3.3.

| Proměnná | Parametr | Směr. odchylka | $z$ | $Pr(>\|z\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 67,875628 | 9,059400 | 7,492 | 6,77e-14 | *** |
| mmbprty | 2,664171 | 0,732263 | 3,638 | 0,000274 | *** |
| happy | 0,140955 | 0,033270 | 4,237 | 2,27e-05 | *** |
| mnyacth | 0,169919 | 0,053740 | 3,162 | 0,001568 | ** |
| yrbrn | -0,035935 | 0,004641 | -7,743 | 9,75e-15 | *** |
| domicil | 0,152021 | 0,062393 | 2,436 | 0,014831 | * |
| rmhhus | 0,134408 | 0,053917 | 2,493 | 0,012671 | * |
| edlvcz | 0,149925 | 0,030365 | 4,937 | 7,92e-07 | *** |
| d.wid | -0,565076 | 0,196868 | -2,870 | 0,004100 | ** |

Tabulka 3.3: Optimální model

## 3.3 Interpretace výsledků

Důležitým krokem v každám statistickém šetření je interpretace výsledků. Co nám tedy výstupy ze statistického softwaru říkají?

### 3.3.1 Signifikanní proměnné

Zkusme se zamyslet nad každou signifikantní proměnnou.

Proměnná mmbprty má velice silný vliv na poměr šancí, že jedinec půjde volit. Tento výsledek je asi velice dobře předvídatelný. Lidé, co se politicky angažují v nějaké straně, ji budou asi i volit.

Pro představu, jak silný má tato proměnná vliv, použijeme vztah (2.6). Poměr šancí $\Omega$ je u straníka $e^{\beta_l}$ krát vyšší, než u nestraníka. V našem případě zhruba $e^{2{,}66} = 14{,}3$, tedy více než čtrnáctkrát.

Naopak vliv proměnné happy tak lehce předvídatelný jistě nebyl. Ukázalo se, že spokojení lidé mají větší šanci na to k volbám jít, než lidé nespokojení. Někdo by mohl například naopak předpokládat, že nespokojení lidé budou chtít situaci změnit, a proto půjdou k volbám, což se však nepotvrdilo.

Tento výsledek můžeme interpretovat tak, že nespokojení lidé vůbec systému nedůvěřují a ignorují ho. Nesmíme však zapomínat, že v proměnné happy lidé popisovali svou celkovou spokojenost. Tedy nejen spokojenost s děním ve společnosti, ale i své štěstí v rodině, zdraví atd.

Proměnná mnyacth ukazovala, jak lidé věří společenskému systému, jak ho považují za férový. Kladnou hodnotu tohoto parametru jsme mohli očekávat. Vysvětlení by mohlo být podobné jako u proměnné happy. Toto je však zajímavý výsledek. Mimo jiné znamená, že lidé, kteří jsou se stavem ve společnosti spokojeni, chodí volit, a ti, co nejsou, nikoliv. Důsledkem toho je však to, že ti lidé, kteří jsou spokojeni, zvolí takovou politickou reprezentaci, která bude pokračovat ve stávajícím systému. A tak systém bude dále nastavován tak, že bude vyhovovat lidem, kteří byli volit. Tedy spokojení lidé zůstanou spokojeni a nespokojení lidé zůstanou nespokojeni.

Záporné znaménko u parametru yrbrn znamená (díky použitému kódování), že starší lidé chodí volit více než ti mladší.

Naopak kladné znaménko díky kódování proměnné domicil znamená, že u lidí z menších sídel je šance, že půjdou volit, větší, než u těch, co žijí ve větších městech. Tento výsledek je relativně překvapivý.

Proměnná rmhhus (počet pokojů v domě/bytě) do jisté míry zastupuje proměnnou hinctnt (příjem domácnosti). Také korelační koeficient nám vychází mírně kladný $(0,31)$. Současně když zkusíme ve výše uvedením modelu vyměnit tyto dvě proměnné, hinctnt stále vychází silně nesignifikantní.

Signifikance proměnné edlvcz také asi nikoho nepřekvapí. Tento výsledek můžeme interpretovat tak, že vzdělanější lidé chodí více volit.

Jako jedinná dummy proměnná vzniklá z proměnné reprezentující rodinný stav nám v modelu zůstala proměnná d3.wid. Tedy, že vdovci a vdovy mají menší šanci, že půjdou volit oproti vdaným/ženatým.

### 3.3.2 Nesignifikantní proměnné

Jako zajímavý výsledek můžeme jistě označit nesignifikanci proměnné gndr. Tedy pohlaví respondenta nemá vliv na šanci, že jedinec půjde nebo nepůjde k volbám. Nesignifikance proměnných hhmodwl a hinctnt poukazuje na to, že materiální situace také není determinujícím faktorem pro to, jestli jedinec půjde volit.

Hlouběji se můžeme zamyslet nad proměnnou lrscale. Její nesignifikance říká, že to, že je člověk zaměřen levicově nebo pravicově, nemá vliv na to, jestli k volbám půjde nebo ne.

Průměrná hodnota proměnné lrscale vychází 5,297, kde 5 označuje střed. Tato hodnota je spočítána z dat, které jsou očištěny u všech proměnných o odpovědi *refusal*, *do not know* a *no answer*. Pokud od těchto odpovědí očistíme pouze proměnnou lrscale, tak její průměrná hodnota vyjde ještě vyšší, a sice 5,398. Když se podíváme jen na ty respondenty, co odpověděli, že volili, průměrná hodnota lrscale se ještě více zvýší.

Volby vyhrála levice, ČSSD a KSČM měly dohromady 111 křesel. Průměrná hodnota, když zahrneme pouze lidi, co volili, by tedy měla vyjít rozhodně menší než 5. Tento rozpor můžeme vysvětlit v zásadě třemi způsoby.

Buď je vzorek, se kterým pracujeme, nereprezantativní. Zastoupení jednotlivých skupin obyvatel je ve vzorku tedy jiné než ve společnosti. V tom případě nemůžeme žádné naše závěry učiněné na základě tohoto vzorku zobecnit na celou společnost.

Početně malá, ale velmi extrémně pravicová skupina by mohla vychýlit průměrnou hodnotu proměnné lrscale. Kontrolou histogramu jsme však tuto možnost vyloučili.

Poslední možností je, že lidé do dotazníku systematicky odpovídají jinak, než jak se ve skutečnosti chovají. Konkrétně v tomto případě nepřiznávají, že jsou levicově orientováni. O tomto jevu se někdy mluví v souvislosti s podhodnocením volebního výsledku levicových stran, zejména KSČM. V posledních volbách, v roce 2006, však již k žádnému podcenění nedošlo a volební výsledek komunistů odpovídal jejich předvolebním prognózám.

# Kapitola 4

# Závěr

V této práci jsme popsali zobecnění klasické lineární regrese, které nám umožnilo analyzovat závislost binární proměnné na dalších veličinách.

Zaměřili jsme se na logistickou regresi. Pro ji jsme ukázali jak pomocí metody maximální věrohodnosti odhadnout parametry, jak interpretova výsledky a nastínili jsme pozadí testů, které se používá při hledání optimálního modelu.

Při práci s daty jsme analyzovali vzorek České republiky z European Social Survey. Vysvětovanou proměnnou bylo, zda respondent v minulých volbách do Poslanecké sněmovny volil či nikoliv. Jako vysvětlující proměnné jsme použili nejen řadu demografických, sociálních a ekonomických proměnných, ale i subjektnivní pocity respondenta.

Ukázalo se, že na to, zda člověk půjde volit, mají vliv proměnné zachycující členství v nějaké politické straně, spokojenost, důvěra ve společenské zřízení, velikost obce ve které člověk žije, vzdělání, velikost obydlí a to, když je daná osoba vdova či vdovec.

Naopak u proměnných jako pohlaví, příjem domácnosti, rodiný stav kromě vdovy či vdovce, vlastnictí domova, politická orientace, nábožnosti a počet let strávený v práci se ukázalo, že na to, zda člověk půjde volit, vliv spíše nemají.

Z dat můžeme pojmout podezření, že se lidé stydí za svou levicovou politickou orientaci a do průzkumů ji nepřiznávají.

# Literatura


[1] Agresti Alan: *Categorical Data Analysis*, John Wiley & Sons, 2002, první i druhé vydání

[2] Anděl, Jiří: *Základy matematické statistiky*, Matfyzpress, Praha 2005

[3] Zvára Karel, Štěpán Josef: *Pravděpodobnost a matematická statistika*, Matfyzpress, Praha 2002 2. vydání

[4] *European Social Survey*, http://www.europeansocialsurvey.org/

[5] software R, http://www.R-project.org/