UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Aplikovaná logistická regrese

Vedoucí diplomové práce:
Mgr. Jana Vrbková, Ph.D.
Rok odevzdání: 2012

Vypracoval:
Petr Dokoupil
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořil tuto diplomovou práci samostatně za vedení Mgr. Jany Vrbkové, Ph.D. a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 30. března 2012

Poděkování

Rád bych na tomto místě poděkoval vedoucí diplomové práce Mgr. Janě Vrbkové, Ph.D. za obětavou spolupráci i za čas, který mi věnovala při konzultacích. Díky své vysoké odbornosti a laskavému přístupu mi byla velikou oporou při psaní této práce. Také bych rád poděkoval své přítelkyni, že se mnou měla v této těžké době trpělivost a podporovala mě ve studiu.

Obsah

Úvod

Logistická regresní analýza vyšetřuje vztahy mezi množinou vysvětlujících proměnných a závisle proměnnu (binární, nominální, ordinální).

Pro binární závisle proměnnou D, která může nabývat jen 2 hodnot (D = 1 znamená přítomnost určité události např. nemoci, D = 0 naopak nepřítomnost), vektor $\mathbf{X} = (X_1, \ldots, X_k)$ vysvětlujících proměnných a podmíněnou pravděpodobnost výskytu události $P(\mathbf{X}) = P(D = 1|\mathbf{X})$ má základní logistický model pravděpodobnosti tvar

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \boldsymbol{\beta}'\mathbf{X})}},$$

kde α je absolutní člen (intercept) a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ vektor neznámých parametrů.

Z důvodu jednodušších výpočtů se více používá logitový tvar logistického modelu pravděpodobnosti tzv. logitový model

$$logit P(\mathbf{X}) = \log\left[\frac{P(\mathbf{X})}{1 - P(\mathbf{X})}\right] = \alpha + \boldsymbol{\beta}'\mathbf{X}.$$

Tento model patří do daleko větší třídy lineárních modelů, a to konkrétně do tzv. zobecněných lineárních modelů, ve kterých tzv. linková funkce $g(\mu) = g(ED)$ vytváří vazbu mezi náhodnou (stochastickou) složkou a mezi deterministickou (systematickou) složkou náhodné proměnné D. Výše použitá logitová funkce *logit* není jedinou možnou linkovou funkcí, má však výhodu v interpretovatelnosti na rozdíl od dalších využívaných linkových funkcí (*probit*, $log - log$ funkce).

Pro nominální závisle proměnné s G kategoriemi (přirozeně neseřaditelnými) lze logitový model rozšířit na tzv. zobecněný kategoriální logitový model tvaru

$$\log\left[\frac{P(D = g|\mathbf{X})}{P(D = 0|\mathbf{X})}\right] = \alpha_g + \sum_{i=1}^{g} \beta_{gi} X_i = \alpha_g + \boldsymbol{\beta}_g'\mathbf{X}, \qquad g = 1, 2, \ldots, G - 1,$$

kde $\alpha_1, \ldots, \alpha_g$ jsou absolutní členy a $\beta_1, \ldots, \beta_g$ jsou vektory neznámých regresních parametrů.

Odhady parametrů výše uvedených modelů získáváme metodou maximální věrohodnosti. Prakticky je potom tato metoda realizována prostřednictvím iteračních algoritmů. V případě mnou využité procedury LOGISTIC v softwaru SAS (resp. úlohy Logistic Regression v SAS EG) se jedná o dva algoritmy: Newton-Raphsonova metoda a Fisherova skórovací metoda, příp. Firthova penalizační metoda v případě separace vstupních dat.

Ve své práci nejprve seznámím čtenáře s nezbytnými matematickými pojmy analýzy dat metodami logistické regrese a poté využiji software SAS Enterpirse Guide k řešení 2 reálných příkladů v oblastech, kde není tolik používána. Je to oblast sportovní a oblast společensko-politická.

Téma logistická regrese najdeme převážně v cizojazyčné literatuře. Většina studijních materiálů v češtině se dotýká této problematiky jen velmi letmo. Tato práce by tedy mohla poskytnout i širší seznámení s logistickou regresí, zejména pak multinomickou logistickou regresí, studentům, kteří preferují literaturu v českém jazyce.

Doufám, že tato práce bude pro čtenáře přínosem.

1 Zobecněný lineární model

Logistický regresní model je speciálním případem zobecněného lineárního modelu GLM (z angl. Generalized Linear Model), proto nejprve pojednám v této kapitole právě o tomto modelu a souvisejících pojmech jako je např. třída hustot exponenciálního typu, linková funkce a její kanonický tvar apod.

1.1 Lineární model

Dříve než přistoupím k definici zobecněného lineárního modelu, připomenu definici a několik tvrzení, které se týkají klasického lineárního modelu.

Definice 1.1. *Nechť* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *je náhodný vektor,* $\mathbf{X}$ *daná matice typu* $n \times k$*, kde* $k < n$ *a* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ *je vektor parametrů. Řekneme, že* $\mathbf{Y}$ *se řídí lineárním modelem, jestliže*

1. $E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$
2. $var\mathbf{Y} = \boldsymbol{V}$ *existuje a nezávisí na* $\boldsymbol{\beta}$*.*

Tuto skutečnost zapisujeme $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \boldsymbol{V})$*.*

V teorii lineárních modelů nám jde zejména o odhad vektoru parametrů lineárního modelu $\boldsymbol{\beta}$, přičemž požadujeme, aby takový odhad měl jisté speciální vlastnosti.

Definice 1.2. *Řekneme, že vektor* $\mathbf{b}$ *je nejlepší nestranný lineární odhad (BLUE) vektoru* $\boldsymbol{\beta}$*, jestliže:*

1. *existuje matice* $\boldsymbol{U}$ *typu* $k \times n$ *taková, že* $\mathbf{b} = \boldsymbol{UY}$*,*
2. $\mathbf{b}$ *je nestranným odhadem* $\boldsymbol{\beta}$*, tj. platí* $E\mathbf{b} = \boldsymbol{\beta}$*,*
3. *je-li* $\mathbf{b}^*$ *jiný nestranný lineární odhad* $\boldsymbol{\beta}$*, pak musí platit* $var\mathbf{b}^* - var\mathbf{b} \geq 0$*.*

Poznámka 1.1. *Je-li splněn předpoklad 1. v Def. 1.2, pak mluvíme o* $\mathbf{b}$ *jako o lineárním odhadu* $\boldsymbol{\beta}$*.*

Poznámka 1.2. *Výraz* $var\mathbf{b}^* - var\mathbf{b} \geq 0$ *v bodě 3. v Def 1.2. je třeba chápat jako tvrzení, že rozdíl variančních matic je pozitivně semidefinitní, tj. pro* $\forall$ *vektor* $\mathbf{c} \in \mathbb{R}^k$, $\mathbf{c} \neq 0$ *platí*

$$\mathbf{c}'(var\mathbf{b}^* - var\mathbf{b})\mathbf{c} \geq 0.$$

Pokud je hodnost $h(\mathbf{X})$ matice $\mathbf{X}$ rovna počtu jejích sloupců (matice má tzv. plnou hodnost ve sloupcích), je určení BLUE vektoru $\boldsymbol{\beta}$ založeno na následující větě.

Věta 1.1. *Nechť* $h(\mathbf{X}) = k$ *a* $\mathbf{V} = var\mathbf{Y}$ *je regulární matice. Potom BLUE* $\mathbf{b}$ *pro* $\boldsymbol{\beta}$ *je určen vztahem*

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{Y} \tag{1}$$

a má varianční matici

$$var\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{2}$$

Důkaz: Odhad (1) je lineární odhad. Je to také odhad nestranný, poněvadž platí

$$\begin{aligned} E\mathbf{b} &= E\left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\right] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}E\mathbf{Y} = \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

Mějme $\mathbf{b}^*$ jiný lineárně nestranný odhad $\boldsymbol{\beta}$, $\mathbf{b}^* = \mathbf{U}\mathbf{Y}$, potom

$$var\mathbf{b}^* = \mathbf{U}var\mathbf{Y}\mathbf{U}' = \mathbf{U}\mathbf{V}\mathbf{U}' = \mathbf{U}\mathbf{V}^{1/2}\mathbf{V}^{1/2}\mathbf{U}'$$

Matice $\mathbf{V}^{1/2}$ existuje, protože $\mathbf{V}$ je regulární.

$$\begin{aligned} var\mathbf{b} &= \left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right] var\mathbf{Y} \left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right]' = \\ &= \underbrace{(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\,\underbrace{\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}}_{\mathbf{V}^{-1}}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}}_{\mathbf{I}} = \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \end{aligned}$$

Platí

$$var\mathbf{b}^{\alpha} - var\mathbf{b} = \mathbf{U}\mathbf{V}^{1/2}\mathbf{V}^{1/2}\mathbf{U}' - (\mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{U})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{U}) \geq 0,$$

poněvadž $\mathbf{U}\mathbf{X} = \mathbf{I}$ (a teda i $\mathbf{X}'\mathbf{U}' = \mathbf{I}$, viz Věta 1, s. 132 Anděl, 1985) a pro libovolnou matici $\mathbf{A}$ typu m $\times$ n takovou, že $\mathbf{A}\mathbf{A}'$ je regulární, a matici $\mathbf{P}$ typu n $\times$ k platí tzv. zobecněná Schwarzova nerovnost

$$\mathbf{P}'\mathbf{P} - (\mathbf{A}\mathbf{P})'(\mathbf{A}\mathbf{A}')^{-1}(\mathbf{A}\mathbf{P}) \geq 0.$$

Stačí položit $\mathbf{A} = \mathbf{X}'\mathbf{V}^{-1/2}$ a $\mathbf{P} = \mathbf{V}^{-1/2}\mathbf{U}'$. ■

Kromě samotného odhadu vektoru neznámých parametrů $\boldsymbol{\beta}$ nás často zajímá i odhad lineární kombinace prvků vektoru $\boldsymbol{\beta}$, tj. odhad parametru $\Theta = \mathbf{c}'\boldsymbol{\beta}$, kde $\mathbf{c} \in \mathbb{R}^k$. Platí následující tvrzení.

Věta 1.2. *Nechť $h(\mathbf{X}) = k$ a $var\mathbf{Y} = \mathbf{V}$ je regulární, potom BLUE $\hat{\Theta}$ parametru $\Theta = \mathbf{c}'\boldsymbol{\beta}$ je $\hat{\Theta} = \mathbf{c}'\mathbf{b}$, kde $\mathbf{b}$ je BLUE $\boldsymbol{\beta}$.*

Důkaz: Viz důkaz Věty 4, s. 133 (Anděl, 1985)

1.2 Vážený průměr

Nechť $Y_1, \ldots, Y_n$ jsou takové nezávislé veličiny, že

$$EY_i = \beta, \qquad varY_i = \sigma_i^2 > 0,$$

kde β je neznámý parametr a $\sigma_1^2, \ldots, \sigma_n^2$ jsou známá čísla, což lze zapsat jako

$$E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}, \qquad var\mathbf{Y} = \mathbf{V},$$

kde

$$\mathbf{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$$

Odhad b neznámého parametru β potom získáme dle Věty 1.1 jako

$$b = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

Protože

$$\mathbf{V}^1 = \begin{pmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n^2} \end{pmatrix}$$

je

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}) = (1,\ldots,1) \begin{pmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n^2} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

a

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \frac{1}{\sum_{i=1}^{n} \sigma_i^2} = varb$$

Potom tedy

$$b = \frac{1}{\sum_{i=1}^{n} \sigma_i^2} \sum_{j=1}^{n} \frac{Y_j}{\sigma_j^2}.$$

Parametr b představuje vlastně vážený průměr veličin $Y_1,\ldots,Y_n$ s váhami $\sigma_1^{-2},\ldots,\sigma_n^{-2}$.

2 Rozdělení pravděpodobnosti exponenciálního typu

Teorie zobecněných lineárních modelů je založena na rodině distribucí exponenciálního typu. Každou hustotu $f(z|\zeta)$ patřící do této množiny lze zapsat ve tvaru

$$f(z|\zeta,\xi) = \exp\big[\underbrace{t(z)u(\zeta)}_{\text{interakční komponenta}} + \underbrace{\log(r(z)) + \log(s(\zeta))}_{\text{aditivní komponenta}}\big],$$

kde $r(\cdot)$ a $t(\cdot)$ jsou reálné funkce proměnné z nezávislé na ζ, $s(\cdot)$ a $u(\cdot)$ jsou reálné funkce proměnné ζ nezávisející na z, přičemž $r(z) > 0$ a $s(\zeta) > 0$ pro $\forall z, \zeta$.

Bylo dokázáno, že rozdělení tohoto typu mají řadu výhodných vlastností, např. existují všechny momenty pro náhodnou veličinu s takovouto distribucí.

V případě, že $t(z) = z$ pro $\forall z$, potom říkáme, že hustota je v kanonické formě vzhledem k náhodně proměnné Z. Podobně, je-li $u(\zeta) = \zeta$ pro $\forall\zeta$ říkáme, že je hustota v kanonické formě vzhledem k parametru ζ.

Často se proto setkáváme (a dále v textu budeme uvažovat) se zápisem hustoty exponenciálního typu $f(y,\theta)$ rozdělení transformované náhodné veličiny $Y = t(z)$ právě v kanonickém tvaru

$$f(y,\theta) = \exp\left(y\theta - b(\theta) + c(y)\right) \tag{3}$$

a o $\theta = u(\zeta)$ hovoříme jako o kanonickém parametru.

O členu $b(\theta)$ se často hovoří jako o tzv. „normalizační konstantě“. Je to jediný člen nezávisející na vlastních datech.

Vztah $\theta = u(\zeta)$ je tzv. kanonický link poskytující vazbu mezi vyjádřením hustoty v původní a v kanonické formě. Vyjádření hustoty ve tvaru (3) přitom není jednoznačné.

Často potřebujeme vyjádřit podobně sdruženou hustotu náhodného vektoru $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ tvořícího náhodný výběr z rozdělení exponenciálního typu, tj.

$$f(y,\theta) = \exp\left[\sum_{i=1}^{n} y_i\theta - nb(\theta) + \sum_{i=1}^{n} c(y_i)\right].$$

Pokud hustota závisí na vektoru parametrů $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, pak můžeme psát

$$f(y,\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{k} (y\theta_j - b(\theta_j)) + c(y)\right].$$

Příklad 2.1. *Binomické rozdělení* $Bi(n,p)$

$$f(y;n,p) = \binom{n}{y} p^y (1-p)^{n-y} = \exp\left[\log\binom{n}{y} + y\log(p) + (n-y)\log(1-p)\right] =$$

$$= \left[\underbrace{y\log\left(\frac{p}{1-p}\right)}_{y\theta} - \underbrace{(-n\log(1-p))}_{b(\theta)} + \underbrace{\log\binom{n}{y}}_{c(y)}\right]$$

Potřebujeme vyjádřit

$$b(\theta) = -n\log(1-p)\Big|_{\theta=\log\frac{p}{1-p}}$$

$$e^\theta = \frac{p}{1-p} \Rightarrow p = \frac{e^\theta}{1-e^\theta}$$

$$-n\log(1-p) = -n\log\left(\frac{1}{1+e^\theta}\right) = -n\log 1 + n\log(1+e^\theta) =$$

$$= n\cdot\log(1+e^\theta) = b(\theta)$$

$$\Rightarrow \qquad f(y,\theta) = \exp[y\theta - b(\theta) + c(y)]$$

$$kde \qquad \theta = \log\left(\tfrac{p}{1-p}\right),$$

$$b(\theta) = n\cdot\log(1+e^\theta),$$

$$c(y) = \log\tbinom{n}{y}.$$

3 Lineární struktura a linkové funkce

Standardní lineární model lze vyjádřit ve tvaru

$$\mathbf{V} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$$E\mathbf{V} = \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta},$$

kde $\mathbf{V} = (V_1, \ldots, V_n)'$ je náhodný výběr, tj. vektor stejně rozdělených nezávislých náhodných veličin se střední hodnotou θ. Vektor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ reprezentuje náhodnou složku modelu, přičemž $\epsilon_i \sim \mathbb{N}(0, \sigma)$, i = 1, ..., n.

Matice $\mathbf{X}$ typu n × k je tzv. matice pozorovaných hodnot (design matrix), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ je vektor neznámých odhadovaných parametrů. Součin $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\theta}$ tvoří systematickou (nenáhodnou) složku modelu. Tento model zapisujeme

$$\mathbf{V} \sim \mathbb{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

Mějme nyní spojitou funkci $g(\cdot)$ takovou, že

$$g(\boldsymbol{\mu}) = \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}.$$

Tuto funkci nazýváme linková funkce, protože poskytuje vazbu (link) mezi lineárním prediktorem $\mathbf{X}\boldsymbol{\beta}$ a střední hodnotou $\boldsymbol{\mu}$ sledované závislé veličiny, která nemusí být (a často není) normálně rozdělená.

Zobecněný lineární model (GLM) má potom tyto komponenty:

1. Stochastická (náhodná) komponenta:

   Náhodná veličina reprezentovaná v modelu náhodným výběrem $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ se střední hodnotou $\boldsymbol{\mu}$,

2. Systematická (nenáhodná) složka:

   Lineární prediktor $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, kde vysvětlující proměnné $\mathbf{X} = (X_1, \ldots, X_k)'$ ovlivňují pozorovanou závisle proměnnou, reprezentovanou $\mathbf{Y}$ jen a pouze prostřednictvím funkce $g(\cdot)$,

3. Linková funkce:

   Specifikuje vazbu mezi stochastickou (náhodnou) složkou a systematickou (nenáhodnou) složkou modelu:

$$g(\boldsymbol{\mu}) = \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta},$$

$$g^{-1}(g(\boldsymbol{\mu})) = g^{-1}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\mu} = E\mathbf{Y}.$$

   Přehled základních linkových funkcí pro vybraná rozdělení je uveden v tabulce 1.

Rozdělení		Kanonický vztah: $\theta = g(\mu)$	Inverzní vztah: $\mu = g^{-1}(\theta)$
Poissonovo		$\log(\mu)$	e^{θ}
Binomické	logit:	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{e^{\theta}}{1+e^{\theta}}$
	probit:	$\Phi^{-1}(\mu)$	$\Phi(\theta)$
	cloglog:	$\log(-\log(1-\mu))$	$1-e^{(-e^{\theta})}$
Normální		μ	θ
Gamma		$-\frac{1}{\mu}$	$-\frac{1}{\theta}$
Negativní binomické		$\log(1-\mu)$	$1-e^{\theta}$

Tabulka 1: Základní linkové funkce (Φ označuje distribuční funkci normovaného normálního rozdělení).

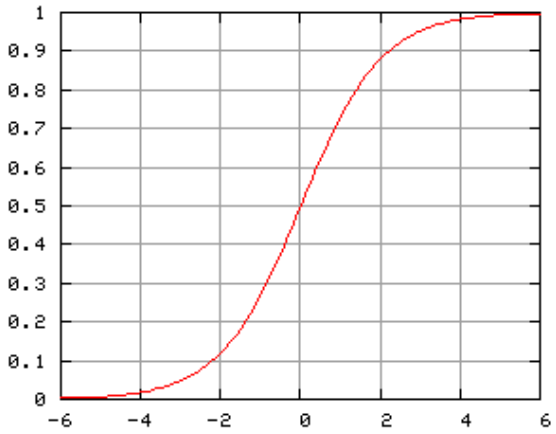
4 Standardní logistická regrese

4.1 Logistická funkce

Logistická regrese vychází z tvaru logistické funkce:

$$F(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

Hodnoty této funkce se nachází mezi 0 a 1, tvar křivky je rostoucí, S-ovitý, v nekonečnu se křivka přimyká k 1 a v mínus nekonečnu k 0. Tyto vlastnosti nám dovolují uvažovat logistickou funkci jako distribuční funkci, konkrétně jde o distribuční funkci logistického rozdělení pravděpodobnosti. Její graf je na obrázku 1.



Obrázek 1: Logistická distribuční funkce

Pro svůj protáhlý S-ovitý tvar je logistická funkce oblíbená hlavně v lékařství, zvláště pak v epidemiologických studiích (KLEINBAUM, David G. and Mitchel KLEIN, 2010). Proměnná z zde reprezentuje index, v němž jsou skloubeny vlivy několika různých rizikových faktorů. Hodnota $F(z)$ následně představuje riziko dané hodnotami z.

4.2 Definování proměnných

Dále budeme uvažovat nezávisle (vysvětlující) proměnné $X_1, X_2, X_3, \ldots, X_k$ a jednu závisle (vysvětlovanou) proměnnou D, značení D je odvozeno od angl. slova disease (v češtině nemoc, onemocnění). Vysvětlující proměnné mohou mít

binární charakter (mohou nabývat pouze hodnot 1 a 0), například výsledek vyšetření může být ohodnocen jako normální stav (= 0) nebo abnormální stav (= 1). Vysvětlující proměnné mohou však mít i spojitý charakter, například věk pacienta. Jedna vysvětlující proměnná může také vyjadřovat kombinaci více sledovaných vlivů.

Závislou proměnnou D budeme v případě standardní logistické regrese uvažovat pouze v binomickém tvaru, tj. nabývající hodnot 0 nebo 1. Úkolem je určit vztah mezi vysvětlujícími proměnnými a vysvětlovanou proměnnou.

$$X_1, X_2, X_3, \ldots, X_k \Rightarrow D$$

4.3 Logistický model

Přechod od logistické funkce k logistickému modelu provedeme tak, že proměnnou z vyjádříme jako lineární kombinaci vysvětlujících proměnných a neznámého vektoru parametrů $(\alpha, \beta_1, \ldots, \beta_k)'$.

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \alpha + \sum_{i=1}^{n} \beta_i X_i \tag{5}$$

Pokud z dosadíme do logistické funkce, dostaneme vyjádření tvaru

$$F(z) = \frac{1}{1 + e^{(-z)}} = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{n} \beta_i X_i)}} = \frac{1}{1 + e^{-(\alpha + \boldsymbol{\beta}'\mathbf{X})}}, \tag{6}$$

kde $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_k)$ a $\mathbf{X} = (X_1, \ldots, X_k)$.

Pravděpodobnost výskytu události D můžeme popsat pomocí podmíněné pravděpodobnosti

$$P(D = 1 | X_1, X_2, X_3, \ldots, X_k) \tag{7}$$

Pokud poskládáme všechny tyto informace dohromady, dostáváme definici logistického modelu.

Definice 4.1. *Logistický model pravděpodobnosti v základním tvaru definujeme jako*

$$P(D=1|X_1,X_2,X_3,\ldots,X_k) \overset{\text{ozn.}}{=} P(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i X_i)}} \tag{8}$$

kde α *a* $\beta_i, i = 1,\ldots,k$ *jsou neznámé parametry,* $X_1, X_2, X_3, \ldots, X_k$ *jsou nezávisle proměnné a* D *je binomická závisle proměnná.*

Pozn.: $P(D=0|\mathbf{X}) = 1 - P(D=1|\mathbf{X})$

4.4 Logitová transformace modelu

Klasický logistický regresní model je speciálním případem zobecněného lineárního regresního modelu pro případ, kdy modelujeme pravděpodobnost pro dichotomickou (binární) proměnnou, tj. pracujeme s binomickým rozdělením pravděpodobnosti, jehož hustota je hustotou exponenciálního typu, jak bylo uvedeno a odvozeno výše. Pro toto rozdělení je k dispozici několik linkových funkcí (viz tabulka 1). Dále budu pracovat pouze s logitovou funkcí pro její snadnou interpretovatelnost.

Definice 4.2. *Logistický model pravděpodobnosti v logitovém tvaru definujeme jako*

$$logitP(\mathbf{X}) = \alpha + \sum_{i=1}^{k}\beta_i X_i,$$

kde $P(\boldsymbol{X}) = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i X_i)}} = \frac{1}{1+e^{-(\alpha+\boldsymbol{\beta'X})}}$, α *a* β_i, $i = 1,\ldots,k$ *jsou neznámé parametry a* $X_1, X_2, X_3, \ldots, X_k$ *jsou vysvětlující proměnné.*

Můžeme tedy psát

$$logitP(\mathbf{X}) = \log\left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right] = \alpha + \sum_{i=1}^{k}\beta_i X_i = \alpha + \boldsymbol{\beta'}\mathbf{X}. \tag{9}$$

Logitový tvar logistického modelu vyjadřuje přirozený logaritmus podílu pravděpodobnosti $P(\mathbf{X}) = P(D=1|\mathbf{X})$, že sledovaná událost D nastane oproti pravděpodobnosti $1 - P(\mathbf{X}) = 1 - P(D=1|\mathbf{X}) = P(D=0|\mathbf{X})$, že sledovaná událost

D nenastane. Samotný podíl nazýváme poměr šancí a značíme jej OR (z anglického odds ratio). Máme-li například $p = P(\boldsymbol{X}) = 0,2$, potom podíl vypadá následovně $\frac{p}{1-p} = \frac{0,2}{0,8} = \frac{1}{4} = 0,25$. Výsledný *logit* je přirozeným logaritmem o základu e z této hodnoty poměru šancí, tj. $\log(\frac{p}{1-p}) = \log(0,25) = -1,386$. Aby nedošlo k nejasnostem, uveďme definice OR.

Definice 4.3. *Poměr šancí (OR) definujeme předpisem*

$$OR_x = \frac{P(\mathbf{X})}{1 - P(\mathbf{X})}.$$

Tato hodnota nám udává míru rizika.

Pro názornost bych uvedl výpočet OR pro tabulku četností 2 x 2, viz tabulka 2.

	E = 1	E = 0
D = 1	a	b
D = 0	c	d

Tabulka 2: Tabulka četností 2x2

D zastupuje závisle proměnnou, E zastupuje vysvětlující proměnnou a a, b, c, d, jsou skupinové četnosti. Odhad $\widehat{OR}$ poměru šancí OR pak spočítáme prostřednictvím vzorce

$$\widehat{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{c \cdot b}. \tag{10}$$

Vztah (10) můžeme přepsat pomocí podmíněných pravděpodobností (resp. jejich odhadů) takto

$$\widehat{OR} = \frac{\hat{P}(D = 1|E = 1)/\hat{P}(D = 1|E = 0)}{\hat{P}(D = 0|E = 1)/\hat{P}(D = 0|E = 0)}, \tag{11}$$

kde např. $\hat{P}(D = 1|E = 1)$ je odhad pravděpodobnosti, že jev D nastal za podmínky, že nastal jev E.

Definice 4.4. *Rizikový poměr, RR (risk ratio), definujeme jako poměr dvou pravděpodobností nastání jevu u dvou vybraných vzorků.*

$$RR_{\overline{\mathbf{X}},\mathbf{X}} = \frac{\frac{1}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i\cdot\overline{X_i})}}}{\frac{1}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i\cdot X_i)}}} = \frac{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i\cdot\overline{X_i})}}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i\cdot X_i)}},$$

kde $\overline{\mathbf{X}}$ *označuje soubor hodnot pro pvní vzorek a* $\mathbf{X}$ *označuje soubor hodnot pro druhý vzorek.*

Použijeme-li například dva vzorky lišící se pouze v jednom binárním faktoru souboru $X_1, \ldots, X_k$, ukáže hodnota rizikového faktoru kolikrát je vyšší pravděpodobnost nastání jevu v závislosti na tomto faktoru.

Definice 4.5. *Rizikový poměr šancí v logistickém modelu ROR (z anglického risk odds ratio) definujeme předpisem*

$$ROR_{X_1,X_0} = e^{\sum_{i=1}^{k}\beta_i(X_{1i}-X_{0i})}.$$

Rizikový poměr šancí vychází z poměru pravděpodobností dvou porovnávaných skupin, které můžeme nazvat skupina 1 a skupina 0. Obě skupiny můžeme definovat pomocí vektoru $\boldsymbol{X}$ vysvětlujících proměnných X_i, $i = 1, \ldots, k$.

Označme $\mathbf{X}_1$ jako kolekci proměnných X_i, $i = 1, \ldots, k$, které specifikují skupinu 1 a $\boldsymbol{X}_0$ jako kolekci X specifikující skupinu 0, tj.

$$\boldsymbol{X}_1 = (X_{11}, X_{12}, \ldots, X_{1k}),$$

$$\boldsymbol{X}_0 = (X_{01}, X_{02}, \ldots, X_{0k}).$$

Následně budeme aplikovat tvar logistického modelu do obecného předpisu a dostaneme obecný tvar pro rizikový poměr šancí v logistickém tvaru.

Pro

$$P(\boldsymbol{X}) = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i X_i)}},$$

a

$$OR_{\boldsymbol{X}_1} = \frac{P(\boldsymbol{X}_1)}{1 - P(\boldsymbol{X}_1)} = e^{\alpha + \sum_{i=1}^{k} \beta_i X_{1i}},$$

$$OR_{\boldsymbol{X}_0} = \frac{P(\boldsymbol{X}_0)}{1 - P(\boldsymbol{X}_0)} = e^{\alpha + \sum_{i=1}^{k} \beta_i X_{0i}},$$

máme

$$\begin{aligned} ROR_{\boldsymbol{X}_1, \boldsymbol{X}_0} = \frac{OR_{\boldsymbol{X}_1}}{OR_{\boldsymbol{X}_0}} = \frac{e^{\alpha + \sum_{i=1}^{k} \beta_i X_{1i}}}{e^{\alpha + \sum_{i=1}^{k} \beta_i X_{0i}}} = \\ = e^{(\alpha + \sum_{i=1}^{k} \beta_i X_{1i}) - (e^{\alpha + \sum_{i=1}^{k} \beta_i X_{0i}})} = \\ = e^{[\alpha - \alpha + \sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})]} = e^{\sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})} \end{aligned}$$

Na základě tohoto vyjádření můžeme rizikový poměr šancí také definovat jiným vztahem, který nazýváme multiplikatívní vztah,

$$ROR_{\boldsymbol{X}_1, \boldsymbol{X}_0} = \prod_{i=1}^{k} e^{\beta_i (X_{1i} - X_{0i})}. \tag{12}$$

Zde můžeme vidět, že v logistické regresi každá vysvětlující proměnná $X_i, i = 1, \ldots, k$ přispívá k OR multiplikativně.

Pokud budeme uvažovat jednu vysvětlující proměnnou X v binomickém tvaru (0, 1), tj. $X_1 = 1, X_0 = 0$, dostaneme

$$ROR = e^{\beta(1-0)} = e^{\beta}. \tag{13}$$

5 Odhady parametrů logistického modelu

Pro výpočet odhadů parametrů zobecněných lineárních modelů, mezi něž logistický regresní model patří, se v současnosti používají různé iterační algoritmy poskytující tzv. maximálně věrohodné odhady. Otcem myšlenek, které vedly k vytvoření metody maximální věrohodnosti, je Daniel Bernoulli (1700-1782).

Z důvodu její výpočetní náročnosti nebyla tato metoda dlouho používána. Nejpoužívanější byla metoda nejmenších čtverců (MNČ). Teprve až s nástupem počítačů a vyvinutím příslušných softwarů se metoda maximální věrohodnosti velice rozšířila. Pokud předpokládáme u závisle proměnné normální rozdělení, poskytuje metoda maximální věrohodnosti a metoda nejmenších čtverců stejné výsledky (SEHNALOVÁ, Michala, 2009).

Metoda maximální věrohodnosti je založená na tom, že odhady neznámých parametrů distribuce pravděpodobnosti uvažované náhodné veličiny se vyberou tak, aby hodnoty hustoty (pravděpodobnostní funkce) v bodech náhodného výběru byly maximální. Tuto metodu můžeme používat ve velmi rozmanitých situacích a odhady získané tímto způsobem mají velmi dobré vlastnosti.

5.1 Maximálně věrohodné odhady

Nechť $\boldsymbol{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z diskrétního rozdělení s pravděpodobnostní funkcí $p(x, \boldsymbol{\Theta})$, resp. ze spojitého rozdělení s hustotou $f(x, \boldsymbol{\Theta})$, kde

$$\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_m)' \in \boldsymbol{\Omega} \tag{14}$$

je vektor neznámých parametrů, který může nabývat jen hodnot z nějakého parametricého prostoru $\boldsymbol{\Omega} \subset \mathbb{R}^{\mathbf{m}}$.

Omezme se pro jednoduchost na jednorozměrný parametr θ. Pro každé pevné $\mathbf{x} \in \mathbb{R}^n$ lze $p(\mathbf{x}, \theta)$, resp. $f(\mathbf{x}, \theta)$, chápat jako funkce proměnné θ. Pro tuto funkci budeme používat označení $L(\theta)$ (z angl. likelihood - věrohodnost) a budeme ji nazývat *věrohodnostní funkce*. Pro libovolnou dvojici $(\mathbf{x}, \theta)$ samozřejmě platí (díky

nezávislosti náhodných veličin X_i a tomu, že jsou stejně rozdělené)

$$L(\theta) = \prod_{i=1}^{n} p(x_i, \theta), \ resp. \ L(\theta) = \prod_{i=1}^{n} f(x_i, \theta). \tag{15}$$

Jestliže existuje takový bod $\hat{\theta} \in \mathbf{\Omega}$, že pro všechny $\theta \in \mathbf{\Omega}$ platí,

$$L(\mathbf{x}, \theta) \leq L(\mathbf{x}, \hat{\theta}),$$

potom říkáme, že $\hat{\theta}$ je *odhad neznámého parametru θ získaný metodou maximální věrohodnosti.*

Často je výhodnější maximalizovat místo funkce $L(\theta)$ její logaritmus $\log L(\theta)$ (*logaritmická funkce věrohodnosti*). Když si uvědomíme, že logaritmus součinu je rovný součtu logaritmů, maximálně věrohodný odhad $\hat{\theta}$ parametru θ se zpravidla stanoví řešením rovnice

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{\partial \log \prod_{i=1}^{n} f(x_i, \theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} \log f(x_i, \theta)}{\partial \theta} = 0,$$

které říkáme *věrohodnostní rovnice.*

Všechny další úvahy jsou založeny na předpokladu, že odhadujeme parametry hustoty, která patří do tzv. regulárního systému hustot.

Definice 5.1. *Řekneme, že systém hustot $\{f(\mathbf{x}, \theta), \theta \in \mathbf{\Omega}\}$ je regulární, jsou-li splněny tyto podmínky:*

1. *Množina $\mathbf{\Omega}$ je neprázdná a otevřená.*

2. *Množina $\boldsymbol{M} = \{\mathbf{x} : f(\mathbf{x}, \theta) > 0\}$ nazávisí na θ.*

3. *Pro skoro všechna $\mathbf{x} \in \boldsymbol{M}$ (vzhledem k μ) existuje konečná parciální derivace*

$$f'(\mathbf{x}, \theta) = \frac{\partial f(\mathbf{x}, \theta)}{\partial \theta}.$$

4. *Pro všechna $\theta \in \mathbf{\Omega}$ platí $\int_M f'(\mathbf{x}, \theta) d\mu(\mathbf{x}) = 0$.*

5. Integrál

$$J_n(\theta) = \int_M \left[\frac{f'(\mathbf{x},\theta)}{f(\mathbf{x},\theta)}\right]^2 f(\mathbf{x},\theta)d\mu(\mathbf{x})$$

je konečný a kladný.

Funkce $J_n(\theta)$ se nazývá Fisherova míra informace.

Věta 5.1. *Nechť systém hustot $f(\mathbf{x},\theta), \theta \in \mathbf{\Omega}$ je regulární. Jestliže pro skoro všechna $\mathbf{x} \in \boldsymbol{M}$ (vzhledem k μ) existuje*

$$f''(\mathbf{x},\theta) = \frac{\partial^2 f(\mathbf{x},\theta)}{\partial\theta^2},$$

a jestliže pro všechna $\theta \in \mathbf{\Omega}$ platí

$$\int_M f''(\mathbf{x},\theta)d\mu(\mathbf{x}) = 0,$$

pak

$$J(\theta) = -\int_M \frac{\partial^2 \log f(\mathbf{x},\theta)}{\partial\theta^2} f(\mathbf{x},\theta)d\mu(\mathbf{x}).$$

Důkaz: Pro skoro všechna $\theta \in \mathbf{\Omega}$ platí

$$\frac{\partial^2 f}{\partial\theta^2} = \frac{f''}{f} - \left(\frac{f'}{f}\right)^2.$$

Odtud dostáváme

$$J(\theta) = \int_M \left(\frac{f'}{f}\right) f d\mu = \int_M \left(\frac{f''}{f}\right) f d\mu - \int_M \frac{\partial^2 \log f}{\partial\theta^2} f d\mu =$$
$$= -\int_M \frac{\partial^2 \log f}{\partial\theta^2} f d\mu$$

Nyní zobecníme Fisherovu míru informace na Fisherovu informační matici pro mnohorozměrný parametr $\mathbf{\Theta} = (\theta_1, \ldots, \theta_m)'$

Definice 5.2. *Nechť náhodný vektor* $\boldsymbol{X} = (X_1, \ldots, X_n)'$ *má hustotu* $f(\mathbf{x}, \boldsymbol{\Theta})$ *vzhledem k nějaké* σ*-konečné míře* μ*. Předpokládejme, že platí:*

1. $\boldsymbol{\Theta} \in \boldsymbol{\Omega}$*, kde* $\boldsymbol{\Omega}$ *je neprázdná otevřená množina v* $\mathbb{R}_m$*.*
2. *Množina* $\boldsymbol{M} = \{\mathbf{x} | f(\mathbf{x}, \boldsymbol{\Theta}) > 0\}$ *nazávisí na* $\boldsymbol{\Theta}$*.*
3. *Pro skoro všechna* $\mathbf{x} \in \boldsymbol{M}$ *(vzhledem k* μ*) a pro všechna* $i = 1, \ldots, m$ *existují parciální derivace*
$$f'_i(\mathbf{x}, \boldsymbol{\Theta}) = \frac{\partial f(\mathbf{x}, \boldsymbol{\Theta})}{\partial \theta_i}.$$
4. *Pro každé* i *a pro všechna* $\boldsymbol{\Theta} \in \boldsymbol{\Omega}$ *platí* $\int_M f'_i(\mathbf{x}, \boldsymbol{\Theta}) d\mu(\mathbf{x}) = 0$*.*
5. *Pro každou dvojici* (i, j) *existuje konečný integrál*
$$\boldsymbol{J}_{ij}(\boldsymbol{\Theta}) = \int_M \left[\frac{f'_i(\mathbf{x}, \boldsymbol{\Theta}) f'_j(\mathbf{x}, \boldsymbol{\Theta})}{f^2(\mathbf{x}, \boldsymbol{\Theta})} \right] f(\mathbf{x}, \boldsymbol{\Theta}) d\mu(\mathbf{x}),$$
je konečný a kladný.
6. *Matice* $J(\boldsymbol{\Theta}) = \|J_{ij}(\boldsymbol{\Theta})\|_{i,j=1}^m$ *je pozitivně definitní pro každé* $\boldsymbol{\Theta} \in \boldsymbol{\Omega}$*.*

Lze vyslovit podobné tvrzení jako pro Fisherovu míru informace, tj. pro jednorozměrný parametr.

Věta 5.2. *Nechť systém hustot* $f(\mathbf{x}, \boldsymbol{\Theta}), \boldsymbol{\Theta} \in \boldsymbol{\Omega}$ *je regulární. Předpokládejme, že pro skoro všechna* $\mathbf{x} \in \boldsymbol{M}$ *(vzhledem k* μ*) existuje derivace*
$$f''_{ij}(\mathbf{x}, \boldsymbol{\Theta}) = \frac{\partial^2 f(\mathbf{x}, \boldsymbol{\Theta})}{\partial \theta_i \theta_j}, i, j = 1, \ldots, m,$$
a jestliže pro všechna $\boldsymbol{\Theta} \in \boldsymbol{\Omega}$ *platí*
$$\int_M f''_{ij}(\mathbf{x}, \boldsymbol{\Theta}) d\mu(\mathbf{x}) = 0, i, j = 1, \ldots, m.$$
Pak platí
$$J_{ij}(\boldsymbol{\Theta}) = -\int_M \frac{\partial^2 \log f(\mathbf{x}, \boldsymbol{\Theta})}{\partial \theta_i \theta_j} f(\mathbf{x}, \boldsymbol{\Theta}) d\mu(\mathbf{x}), i, j = 1, \ldots, m.$$

Důkaz: Tvrzení dokážeme analogicky jako ve větě 5.1.

Následující tvrzení se týká existence řešení věrohodnostní rovnice a rozdělení maximálně věrohodných odhadů.

Věta 5.3. *Nechť systém hustot* $\{f(x,\theta), \theta \in \Omega\}$ *je regulární. Předpokládejme, že existuje* $f''(x,\theta) = \partial^3 f(x,\theta)/\partial\theta^3$ *pro skoro všechna* $x \in M$ *vzhledem k* μ*. Nechť platí*

$$\int_M f''(x,\theta)d\mu(x) = 0.$$

Nechť existuje nezáporná meřitelná funkce $H(x)$ *splňující podmínku*

$$\int_M H(x)f(x,\theta)d\mu(x) < K$$

(kde K *je kladné číslo nezávislé na* θ*), pro kterou platí*

$$\left|\frac{\partial^3 \log f(x,\theta)}{\partial\theta^3}\right| \leq H(x)$$

pro skoro všechna $x \in M$ *vzhledem k* μ*.*

Nechť $\theta_0 \in \Omega$ *je skutečná hodnota parametru a nechť* $\epsilon \in (0,1)$ *je dané číslo. Pak ke každému dostatečně malému* $\delta > 0$ *existuje takové přirozené* n_0 *(závisející na* ϵ *a* δ*), že pro libovolné* $n \geq n_o$ *má s pravděpodobností alespoň* $(1-\epsilon)$ *věrohodnostní rovnice kořen* θ_n^* *splňující nerovnost* $|\theta_n^* - \theta_0| < \delta$*.*

Jestliže pro každé dostatečně velké n *existuje takový kořen* θ_n^* *věrohodnostní rovnice, že* $\theta_n^* \to \theta_0$ *podle pravděpodobnosti, pak náhodná veličina*

$$n^*(\theta_n^* - \theta_0)$$

má asymptoticky normální rozdělení $\mathbb{N}(0, [J(\theta_0)]^{-1})$*.*

Důkaz: Viz důkaz Věty 10, s. 268 (Anděl, 1985)

Obdobná věta platí i pro případ mnohorozměrného parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$. Je-li řešení $\boldsymbol{\theta}_n^*$ věrohodnostních rovnic konzistentním odhadem skutečné hodnoty parametru $\boldsymbol{\theta}_0$, pak $\sqrt{n}(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0)$ má asymptoticky normální rozdělení $\mathbb{N}(0, [\boldsymbol{J}(\boldsymbol{\theta}_0)]^{-1})$,

kde $\boldsymbol{J}(\boldsymbol{\theta}_0)$ je Fisherova informační matice. Přitom věrohodnostními rovnicemi rozumíme systém

$$\sum_{i=1}^{n} \frac{\partial \log f(X_i, \theta)}{\partial \theta_k} = 0, \qquad k = 1, 2, \ldots, m.$$

5.2 Podmíněná logistická regrese

Při odhadování parametrů logistického regresního modelu můžeme kromě klasické (nepodmíněné) metody maximální věrohodnosti využít i tzv. podmíněnou metodu. Potom hovoříme o podmíněné logistické regresi.

Při rozhodování o tom, zda použít místo klasické regrese podmíněnou logistickou regresi, je důležitý poměr mezi počtem pozorování a počtem parametrů. Nepodmíněná metoda se doporučuje, pokud je počet parametrů malý vzhledem k počtu pozorování. Běžně se snažíme, pokud je to možné, používat nepodmíněnou metodu. Podmíněná metoda je doporučována, pokud máme velký počet parametrů vzhledem k počtu pozorování (velikosti náhodného výběru). Ovšem určit co je malý a co velký počet, je těžké. Oba tyto pojmy jsou relativní a není mezi nimi přesně stanovena hranice. Rozhodnutí, kterou metodu použít, je na řešiteli úlohy.

Výběr jedné z těchto dvou metod je ovlivněn rovněž možnostmi počítačového programu, který chceme pro výpočet maximálně věrohodných odhadů použít.

Funkci věrohodnosti pro podmíněnou metodu lze schematicky zapsat

$$L_C = \frac{P(\text{pozorované hodnoty})}{P(\text{všechny možné kombinace})}$$

Budeme-li mít k dispozici celkem n pozorování, přičemž v m_1 případech $\mathbf{X}_1, \ldots, \mathbf{X}_{m_1}$ nastane sledovaný jev a v $n-m_1$ případech $\mathbf{X}_{m_1+1}, \ldots, \mathbf{X}_n$ sledovaný jev nenastane, můžeme psát

$$L_C = \frac{\prod_{l=1}^{m_1} P(\mathbf{X}_l) \prod_{l=m_1+1}^{n} [1 - P(\mathbf{X}_l)]}{\sum_u \left\{ \prod_{l=1}^{m_1} P(\mathbf{X}_{ul}) \prod_{l=m_1+1}^{n} [1 - P(\mathbf{X}_{ul})] \right\}},$$

kde u označuje počet všech možných konfigurací n pozorování v m_1 případech, kdy sledovaný jev nastane a v $n - m_1$ případech, kdy sledovaný jev nenastane.

Analogický zápis pro nepodmíněnou metodu je pak tvaru

$$L_U = \prod_{l=1}^{m_1} P(\mathbf{X}_l) \prod_{l=m_1+1}^{n} \left[1 - P(\mathbf{X}_l)\right].$$

Nahradíme-li v předchozích vyjádřeních obecný zápis pravděpodobnosti regresním, pak máme pro parametry $\alpha, \beta_1, \ldots, \beta_k$ logistického regresního modelu

$$L_C = \frac{\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^{k} \beta_i X_{li}\right)}{\sum_u \left[\prod_{l=1}^{n} \exp\left(\sum_{i=1}^{k} \beta_i X_{lui}\right)\right]}, \tag{16}$$

$$L_U = \frac{\prod_{l=1}^{n} \exp\left(\alpha + \sum_{i=1}^{k} \beta_i X_{il}\right)}{\prod_{l=1}^{n} \left[1 + \exp\left(\alpha + \sum_{i=1}^{k} \beta_i X_{il}\right)\right]}. \tag{17}$$

Všimněme si, že v případě podmíněné metody (17) se parametr α (absolutní člen) zkrátil. Tento parametr tedy nelze prostřednictvím podmíněné metody odhadnout.

5.3 Iterační algoritmy pro výpočet maximálně věrohodných odhadů

Jelikož pro účely své práce budu využívat zejména statistický software SAS, chtěl bych popsat základní iterační algoritmy, které jsou dostupné v tomto programu.

Základním algoritmem je Fisherova skórovací metoda. Alternativou je pak Newton-Raphosonova metoda. Oba algoritmy poskytují stejné odhady parametrů, ale odhadovaná kovarianční matice odhadů parametrů se mírně liší. To je způsobeno tím, že Fisherova skórovací metoda při výpočtech využívá očekávanou informační matici, zatímco Newton-Raphsonova metoda pozorovanou informační matici.

V případě binárního logistického modelu je očekávaná a pozorovaná informační matice totožná, což vede ke shodě odhadovaných kovariančních matic pro oba algoritmy. V softwaru SAS můžeme specifikovat, jakou metodu z výše uvedených chceme použít a můžeme také vybrat Firthovu penalizační metodu v případě separace vstupních dat.

Pro multinomickou regresi je k dispozici jen Newtonův-Raphsonův algoritmus. Firthova penalizační metoda je v současné době v softwaru SAS dostupná jen pro binární logistické modely.

Popis obou iteračních metod provedu již konkrétně pro speciální případ zobecněných lineárních modelů - logistickou regresi tak, jak jsou výpočetní postupy popsány v manuálech k softwaru SAS, poněvadž právě tento software budu později v praktických příkladech používat.

Při popisu iteračních algoritmů předpokládejme logistický regresní model v následujících tvarech ($\mathbf{X}\ldots$ hodnoty vysvětlovaných proměnných):

- klasická (binární) logistická regrese:

$$g(\pi) = logit(\pi) = \log\left(\frac{P(\mathbf{X})}{1 - P(\mathbf{X})}\right) = \alpha + \boldsymbol{\beta}'\mathbf{X},$$

  kde $P(\mathbf{X}) = P(Y = 1|\mathbf{X})$, Y nabývá jen hodnot 0 a 1, α je tzv. absolutní člen, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)'$ je vektor neznámých regresních parametrů (koeficientů vysvětlujících proměnných),

- podmíněná logistická regrese (kumulativní model):

$$g\left(P(Y \leq i|\mathbf{X})\right) = \alpha_i + \boldsymbol{\beta}'\mathbf{X}, \qquad i = 1, \ldots, k,$$

  kde $\alpha_1, \ldots, \alpha_k$ jsou absolutní členy, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)'$ je vektor neznámých regresních parametrů, Y může nabývat G hodnot, které lze uspořádat,

- multinomická logistická regrese:

$$\log\left[\frac{P(D = g|\mathbf{X})}{P(D = 0|\mathbf{X})}\right] = \alpha_g + \boldsymbol{\beta}_g'\mathbf{X}, \qquad g = 1, 2, \ldots, G - 1,$$

kde $\alpha_1, \ldots, \alpha_g$ jsou absolutní členy a $\beta_1, \ldots, \beta_g$ jsou vektory neznámých regresních parametrů.

5.3.1 Fisherova skórovací metoda

Uvažujme multinomickou proměnnou $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{k+1,j})$ takovou, že

$$Z_{ij} = \begin{cases} 1 & \text{jestliže } Y_j = i \\ 0 & \text{jinak} \end{cases}$$

Jestliže π_{ij} označuje pravděpodobnost, že j-té pozorování odpovídá hodnotě i, očekávaná hodnota $\boldsymbol{Z}_j$ je $\boldsymbol{\pi}_j = (\pi_{ij}, \ldots, \pi_{k+1,j})'$, kde $\pi_{k+1,j} = 1 - \sum_{i=1}^{k} \pi_{ij}$. Kovariančí matici vektoru proměnných $\boldsymbol{Z}_j$ označíme jako $\boldsymbol{V}_j$, což je kovarianční matice multinomické náhodné proměnné pro jeden pokus s vektorovým parametrem $\boldsymbol{\pi}_j$. Nechť $\boldsymbol{\beta}$ je vektor regresních parametrů, jinými slovy $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_s)'$. Nechť $\boldsymbol{D}_j$ je matice parciálních derivací $\boldsymbol{\pi}_j$ s ohledem na $\boldsymbol{\beta}$. Rovnice odhadu pro regresní parametry je pak

$$\sum_j \boldsymbol{D}_j' \boldsymbol{W}_j (\boldsymbol{Z}_j - \boldsymbol{\pi}_j) = 0,$$

kde $\boldsymbol{W}_{ij} = w_j f_j \boldsymbol{V}_j^-$, w_j jsou váhy a f_j frekvence j-tého pozorování, $\boldsymbol{V}_j^-$ je zobecněná inverze k $\boldsymbol{V}_j$. Procedura LOGISTIC volí $\boldsymbol{V}_j^-$ jako inverzní matici k diagonální matici s diagonálními prvky $\boldsymbol{\pi}_j$.

Pro počáteční hodnotu $\boldsymbol{\beta}^{(0)}$ je pak $(m+1)$-ní iterace maximálně věrohodného odhadu $\boldsymbol{\beta}$ získána ze vztahu

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left(\sum_j \boldsymbol{D}_j' \boldsymbol{W}_j \boldsymbol{D}_j \right)^{-1} \sum_j \boldsymbol{D}_j' \boldsymbol{W}_j (\boldsymbol{Z}_j - \boldsymbol{\pi}_j),$$

kde $\boldsymbol{D}_j$, $\boldsymbol{W}_j$ a $\boldsymbol{\pi}_j$ jsou vyčísleny v bodě $\boldsymbol{\beta}^{(m)}$. Výraz za znaménkem plus značí velikost kroku. Iterační schéma pokračuje dokud není dosáhnuto konvergence, tj. dokud $\boldsymbol{\beta}^{(m+1)}$ není dostatečně blízko k $\boldsymbol{\beta}^{(m)}$. Potom maximálně věrohodný odhad $\boldsymbol{\beta}$ je $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$.

Kovarianční matice $\hat{\boldsymbol{\beta}}$ je odhadována podle následujícího vztahu:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \left(\sum_j \widehat{\boldsymbol{D}}_j' \widehat{\boldsymbol{W}}_j \widehat{\boldsymbol{D}}_j \right)^{-1} = \hat{\boldsymbol{I}}^{(-1)},$$

kde $\hat{\boldsymbol{D}}_j$ a $\hat{\boldsymbol{W}}_j$ jsou odhady $\boldsymbol{D}_j$ a $\boldsymbol{W}_j$ vyhodnocené v $\hat{\boldsymbol{\beta}}$. $\hat{\boldsymbol{I}}$ je informační matice, nebo také záporná očekávaná Hessova matice vyhodnocená v $\hat{\boldsymbol{\beta}}$.

5.3.2 Newton-Raphsonova metoda

Pro kumulativní modely (podmíněná logistická regrese), kdy máme vektorový parametr $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_s)'$ a pro zobecněné logitové modely pak $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_k')'$.

Nadefinujeme vektor gradientů $\boldsymbol{g}$ a Hessovu matici $\boldsymbol{H}$ následovně:

$$\boldsymbol{g} = \sum_j w_j f_j \frac{\partial l_j}{\partial \boldsymbol{\beta}},$$

$$\boldsymbol{H} = \sum_j w_j f_j \frac{\partial^2 l_j}{\partial \boldsymbol{\beta}^2},$$

kde $l_j = \log L_j$ je logaritmická funkce věrohodnosti pro j-té pozorování. S počáteční hodnotou $\boldsymbol{\beta}^{(0)}$ je maximálně věrohodný odhad $\hat{\boldsymbol{\beta}}$ koeficientu $\boldsymbol{\beta}$ získán iteračně dokud není dosaženo konvergence dle vztahu

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \boldsymbol{H}^{-1}\boldsymbol{g},$$

kde $\boldsymbol{H}$ a $\boldsymbol{g}$ jsou vyčísleny v bodě $\boldsymbol{\beta}^{(m)}$.

Kovarianční matice $\hat{\boldsymbol{\beta}}$ je odhadována podle následujícího vztahu:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{I}}^{(-1)},$$

kde informační matice $\hat{\boldsymbol{I}} = -\widehat{\boldsymbol{H}}$ je počítána podle $\widehat{\boldsymbol{H}}$ v $\hat{\boldsymbol{\beta}}$.

5.3.3 Firthova penalizační metoda

Firthova penalizační metoda se používá ke snížení zkreslení odhadů parametrů (Heinze and Schemper; 2002; Firth; 1993). Tato metoda je užitečná v případech separace vstupních proměnných a je alternativou k provedení přesné logistické regrese. Jak už bylo zmiňováno, Firthova metoda je v současné době dostupná v softwaru SAS jen pro binární logistické modely.

Nahrazuje obvyklé skórové rovnice (gradient)

$$g(\beta_j) = \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} = 0 \qquad j = 1, \ldots, p,$$

kde p je počet parametrů v modelu.

Modifikovaná skórová rovnice má tvar

$$g(\beta_j)^* = \sum_{i=1}^{n} \{y_i - \pi_i + h_i(0,5 - \pi_i)\} x_{ij} = 0 \qquad j = 1, \ldots, p,$$

kde h_i jsou i-té diagonální prvky matice $\boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X'WX})^{-1}\boldsymbol{X'W}^{1/2}$ a $\boldsymbol{W} = diag\{\pi_i(1-\pi_i)\}$. Hessova matice není modifikována a optimalizační metody se provádí obvyklým způsobem.

5.4 Existence maximálně věrohodných odhadů v modelech logisitcké regrese

Pravděpodobnostní rovnice pro modely logistické regrese nemusí mít vždy konečné řešení. Existence, konečnost a jedinečnost maximálně věrohodných odhadů pro modely logistické regrese závisí na struktuře dat. Máme 3 vzájemě se vylučučící kategorie: kompletní separace, kvazi-kompletní separace a překrytí. Je dokázáno, že v připadě kompletní separace dat a tzv. kvazi-kompletní separace dat neexistují maximálně věrohodné odhady neznámých parametrů logistického regresního modelu (ALBERT, A. and J. A. ANDERSON, 1984).

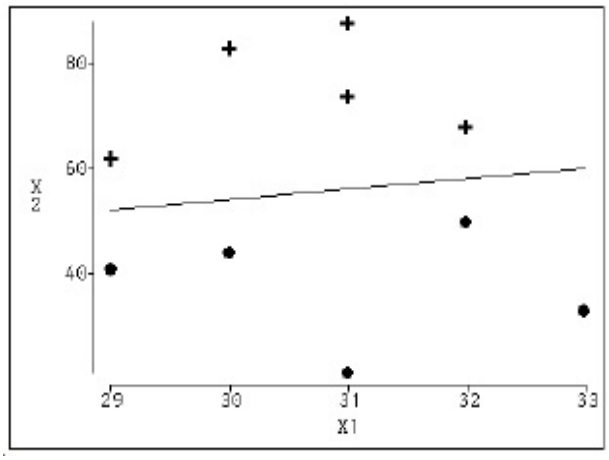
Uvažujme binomickou výstupní proměnnou v modelu logistické regrese. Nechť Y_i, $i = 1, \ldots, n$ je hodnota výstupní proměnné pro i-té pozorování a nechť $\mathbf{X}_i =$

$(1, X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, k$ je vektor vysvětlujících proměnných (zahrnující konstantu 1, která odpovídá absolutnímu členu), kde k je počet vysvětlujících proměnných.

Definice 5.3. *Říkáme, že datové body jsou kompletně separované, jestliže existuje vektor* $\mathbf{b} \in \mathbb{R}^{k+1}$ *takový, že přesně rozděluje všechny pozorování do jejich skupin, tj.*

$$\begin{cases} \mathbf{b}'\mathbf{X}_i > 0 \; Y_i = 0 \\ \mathbf{b}'\mathbf{X}_i < 0 \; Y_i = 1 \end{cases} \qquad i = 1, \ldots, n. \tag{18}$$

Jestliže existuje kompletní separace datových bodů, pak neexistují konečné maximálně věrohodné odhady parametrů modelu logistické regrese a věrohodnostní funkce jde s rostoucí iterací k 0.



Obrázek 2: Příklad kompletní separace datových bodů

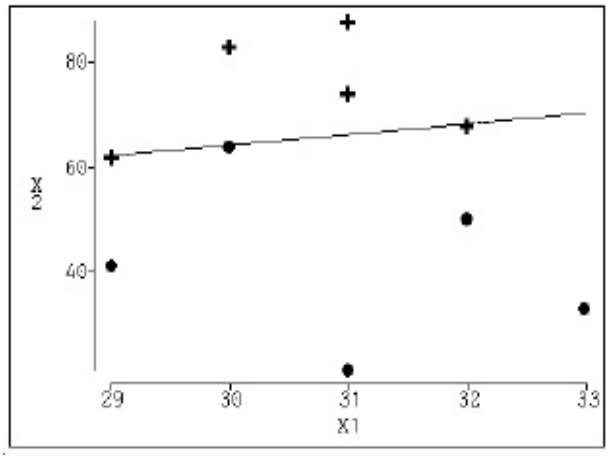
Na obrázku 2 je zobrazena kompletní separace datových bodů pro případ dvou vysvětlujících proměnných X_1 a X_2 a dvou skupin odpovídajících hodnotám vysvětlující proměnné Y, značených jako plus pro $Y_i = 0$ a tečky pro $Y_i = 1$, $i = 1, \ldots, n$.

Definice 5.4. *Říkáme, že datové body jsou kvazi-kompletně separované, jestliže existuje vektor* $\mathbf{b} \in \mathbb{R}^{k+1}$*, pro který platí*

$$\begin{cases} \mathbf{b}'\mathbf{X}_i \geq 0 \; Y_i = 0 \\ \mathbf{b}'\mathbf{X}_i \leq 0 \; Y_i = 1 \end{cases} \qquad i = 1, \ldots, n \tag{19}$$

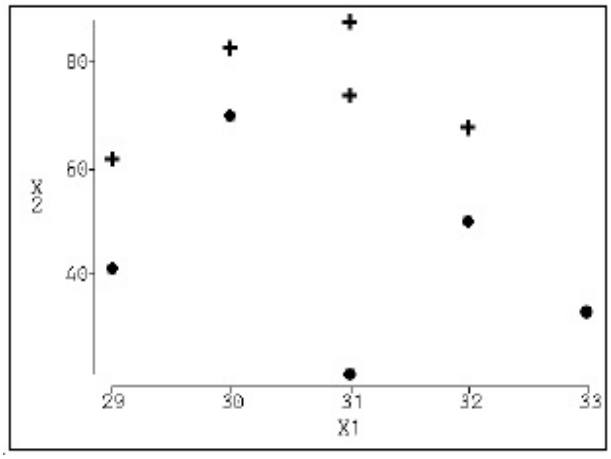
s rovností platící alespoň pro jedno pozorování z každé skupiny.

Jestliže existuje kvazi-kompletní separace datových bodů, pak neexistují konečné maximálně věrohodné odhady parametrů modelu logistické regrese a kovarianční matice se stává neomezenou.



Obrázek 3: Příklad kvazi-kompletní separace datových bodů

Jestliže u datových bodů žádná z předchozích dvou separací neexistuje, pak se datové body nutně překrývají, viz obráezk 4. Maximálně věrohodné odhady parametrů existují a jsou jedinečné.



Obrázek 4: Příklad překrytí datových bodů

Kompletní a kvazi-kompletní separace jsou typické problémy, se kterými se můžeme setkat u malých datových množin. Ačkoli kompletní separace se může vyskytovat u všech typů dat, kvazi-kompletní separace je málo pravděpodobná u skutečně spojitých vysvětlujících proměnných (YING, So, 1995).

6 Posuzování výsledků

Po odhadu parametrů je potřeba zhodnotit dosažené výsledky. Pro toto posuzování využíváme testy hypotéz nebo intervalové odhady parametrů.

6.1 Testy modelů

Při posuzování modelů, ve kterých počítáme odhady neznámých parametrů metodou maximální věrohodnosti, můžeme využít hodnotu věrohodnostní funkce. Obecně platí, že čím víc máme parametrů, tím je model lepší. Věrohodnostní funkce by tedy měla dosahovat vyšší hodnoty maxima (KLEINBAUM, David G. and Mitchel KLEIN, 2010). Proto můžeme pro letmé posouzení využít maximální hodnotu věrohodnostní funkce.

6.1.1 Test poměrem věrohodností

Pro lepší posouzení, jestli použít model s interakcí, nebo bez ní, nebo jestli zahrnout do modelu další parametr, můžeme použít test poměrem věrohodností - LR test (z anglického Likelihood Ratio Test). Pro tento test potřebujeme vypočítat odhady parametrů pro oba posuzované modely a potřebujeme znát maximální hodnotu věrohodnostní funkce. Tímto způsobem porovnáváme dva typy modelů. První typ je úplný (full) model, který zahrnuje všechny parametry. Druhý typ je redukovaný (reduced) model, což je speciální tvar plného modelu. Z plného modelu dostaneme redukovaný model tak, že položíme jeden nebo více parametrů rovny nule. V podstatě testujeme nulovost parametrů, které jsou obsaženy pouze v plném modelu. Je možné takto testovat i více parametrů zároveň.

Testová statistika PV má následující tvar

$$PV = -2\log L_{reduced} - (-2\log L_{full}) = -2\log\frac{L_{reduced}}{L_{full}}, \tag{20}$$

kde $L_{reduced}$ je věrohodnostní funkce redukovaného modelu a L_{full} je věrohodnostní funkce plného modelu.

Tato statistika má při velkých výběrech χ^2 -rozdělení s počtem stupňů volnosti, který odpovídá rozdílu počtu parametrů dvou testovaných modelů.

$$PV \sim \chi^2{}_{k_2-k_1}(0, 1-\alpha).$$

Je-li hodnota L_2 hodně větší než hodnota L_1, pak jejich poměr je velmi malý, v limitním případě se blíží 0. Logaritmus z tohoto poměru je tedy záporný, v limitě se blíží k $-\infty$. Po vynásobení -2 se dostáváme do $+\infty$. Má-li tedy parametr v plném modelu velký vliv (není vhodné jej vynechat) je hodnota LR testu vysoká, pozitivní. Pokud je ale vliv parametrů v plném modelu malý, zanedbatelný, jsou hodnoty L_2 a L_1 téměř stejné. Jejich poměr je v limitě roven 1. Logaritmus 1 je nulový a i po vynásobení -2 dostáváme v limitě nulu.

6.1.2 Waldův test

Waldův test se používá pro testování významnosti jednoho parametru. Pro tento test, na rozdíl od LR testu, nepotřebujeme odhady parametrů dvou modelů. Postačí odhadnout parametry pouze jednoho modelu, a ty pak můžeme testovat. U tohoto testu je poměrně důležité, aby výběr byl dostatečně velký. Samozřejmě tato hodnota není přesně specifikována. Testová statistika pro Waldův test, tj. pro nulovou hypotézu $H_0 : \hat{\boldsymbol{\beta}} = 0$ má přibližně normované normální rozdělení $\mathbb{N}(0, 1)$ a následující tvar

$$Z = \frac{\hat{\beta}}{s_{\hat{\beta}}}, \tag{21}$$

kde $\hat{\beta}$ je maximálně věrohodný odhad testovaného parametru a $s_{\hat{\beta}}$ je odhad standardní chyby testovaného parametru.

Lze použít i testovou statistiku Z^2, která má přibližně χ^2 - rozdělení s jedním stupněm volnosti.

Statistika pro test poměrem věrohodností (LR test) má při velkém výběru přibližně stejné hodnoty jako testová statistika Waldova testu. To znamená, že při dostatečně velkém výběru je jedno, kterou testovací statistiku použijeme.

Ovšem pokud je výběr malý, mohou tyto testy dávat velice rozdílné výsledky. Ve většině situací se dává přednost testu poměrem věrohodností před Waldovým testem. Analýza prostřednictvím Waldova testu je ale naopak výpočetně jednodušší, protože stačí spočítat odhady pouze jednoho modelu.

6.2 Intervalové odhady

6.2.1 Intervalové odhady pro jeden parametr bez interakce

Obecný tvar intervalových odhadů neznámých parametrů modelu lze jednoduše odvodit ze znalosti rozdělení odhadů. Za předpokladu, že odhady mají normální rozdělení, mají intervalové odhady následující tvar

$$\beta_i \in \left\langle \hat{\beta}_i - u\left(1-\frac{\alpha}{2}\right) s_{\hat{\beta}_i}, \hat{\beta}_i + u\left(1-\frac{\alpha}{2}\right) s_{\hat{\beta}_i} \right\rangle,$$

kde $\hat{\beta}_i$ jsou maximálně věrohodné odhady neznámých parametrů, $s_{\hat{\beta}_i}$ jsou odhady standardní chyby neznámých parametrů a $\left(1-\frac{\alpha}{2}\right)$ je kvantil normovaného normálního rozdělení $\mathbb{N}(0,1)$.

Kromě intervalových odhadů pro jednotlivé parametry jsou někdy požadovány i konfidenční intervaly pro poměry šancí OR. Je to zvláště u studií zabývajících se medicínskými problémy, kde spíše než odhady parametrů jsou důležité odhady pravděpodobnosti onemocnění nebo přežití.

Např. budeme předpokládat, že vysvětlující veličina, pro kterou počítáme odhad poměru šancí, má binární charakter (nabývá jen hodnot 0 a 1). Poměr šancí poté vypočítáme podle příslušného vzorce uvedeného v kapitole 4.4 a od něj odvodíme tvar pro intervalový odhad.

$$X_i \in \{0,1\} \qquad \Rightarrow \qquad \widehat{OR}_{X_i} = e^{\hat{\beta}_i}$$

$$OR_{X_i} \in \left\langle e^{\hat{\beta}_i - u\left(1-\frac{\alpha}{2}\right)s_{\hat{\beta}_i}}, e^{\hat{\beta}_i + u\left(1-\frac{\alpha}{2}\right)s_{\hat{\beta}_i}} \right\rangle$$

Pokud budeme mít jiné kódování veličiny X, upravíme příslušný vzorec podle použitého kódu.

Například pokud budeme používat kódování (-1, 1) upravíme vzorec následujícím způsobem

$$X_i \in \{0,1\} \qquad \Rightarrow \qquad \widehat{OR}_{X_i} = e^{(X_1 - X_0)\hat{\beta}_i} = e^{(1-(-1))\hat{\beta}_i} = e^{2\hat{\beta}_i}$$

$$OR_{X_i} \in \left\langle e^{2[\hat{\beta}_i - u\left(1-\frac{\alpha}{2}\right)s_{\hat{\beta}_i}]}, e^{2[\hat{\beta}_i + u\left(1-\frac{\alpha}{2}\right)s_{\hat{\beta}_i}]} \right\rangle$$

Podrobněji o výpočtech bodových odhadů a konfidenčních intervalů pro OR pojednám později.

7 Multinomická logistická regrese

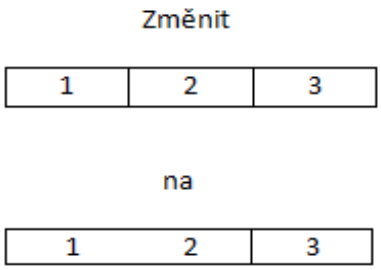
7.1 Přehled

V této kapitole se zaměřím na standardní model logistické regrese rozšířený tak, aby zvládnul výstupní proměnné, které mají více než dvě kategorie. Multinomická logistická regrese se používá v případech, kdy výstupní proměnná má nominální tvar, tzn. že nemá žádné přirozené uspořádání. Pokud má nějaké přirozené uspořádání, lze použít i ordinální logistickou regresi.

Až doteď jsme se bavili o modelu, který zahrnoval závislou proměnnou v binárním tvaru, jako například absence (= 0) či výskyt(= 1) choroby. Nicméně může nastat případ, kdy máme více jak 2 úrovně pro výstupní proměnnou. V této kapitole bych chtěl formulovat podobu a charakteristiky modelu pro takový případ multinomického výstupu.

Typickým příkladem výstupu multinomické logistické regrese je rozdělení pacientů podle pooperačních stavů - bez potíží, střední potíže a závažné potíže. Nebo nejvíce vyhovující léčba pacienta vybraná ze tří a více možností (KLEINBAUM, David G. and Mitchel KLEIN, 2010).

Jedním z přístupů analýzy dat s vícehodnotovým výstupem může být výběr správného bodu rozdělení, binarizace výstupní proměnné, viz obrázek 5. Poté lze jednoduše aplikovat metody standardní logistické regrese diskutované dříve.



Obrázek 5: Výběr bodu rozdělení

Nevýhodou tohoto přístupu je ztráta detailu v popisování výstupů. Například v našem případě nemůžeme už dále porovnávat 1 s 2. Tato ztráta detailu může

ovlivnit vyvozené závěry.

Alternativním přístupem je ponechání původní podoby dat a využití modelů speciálně vyvinutých pro polynomické výstupy. Záleží přitom na škále, na které jsou výstupní proměnné měřeny, tedy jestli jsou nominální nebo ordinální.

Nominální proměnná jednoduše označuje rozdílné kategorie. Příkladem mohou být různé podtypy rakoviny, například u rakoviny děložní sliznice adenokarcinom, adenosquamózní typ a jiný typ.

Ordinální proměnná má přirozené uspořádání mezi jednotlivými stupni. Příkladem mohou být již dříve uvedené pooperační stavy.

7.2 Příklad multinomické logistické regrese se třemi kategoriemi

V této kapitole bych chtěl představit příklad modelu multinomické logistické regrese s binární vstupní (vysvětlující) proměnnou a výstupní proměnnou (odpovědí) D, která má 3 kategorie, což je nejjednodušší případ multinomického modelu.

V příkladu použiji data z Mezinárodního institutu pro rakovinu Black/White Cancer Survival Study (1995). Předpokládáme, že chceme vyhodnotit vztah mezi věkovou skupinou pacientek a typem rakoviny děložní sliznice. Vstupní proměnná je kódována jako 0 pro věkovou skupinu 50-64 nebo 1 pro věkovou skupinu 65-79. Typy rakoviny jsou kódovány jako 0 pro adenokarcinom, 1 pro adenosquamózní typ a 2 pro jiný typ. Mezi výstupními proměnnými neexistuje žádné upořádání a kódování je nahodilé. Data jsou prezentována v tabulce 3.

	50-64 $X_1 = 0$	65-79 $X_1 = 1$
Adenokarcinom D = 0	77	109
Adenosquamózní typ D = 1	11	34
Jiný typ D = 2	18	39

Tabulka 3: Četnosti typů rakoviny dle věkové kategorie

Při použití multinomické logistické regrese musí být jedna z kategorií výstupní proměnné označena jako referenční kategorie a všechny ostatní jsou s ní srovnávány. Výběr referenční kategorie může být náhodný a je na uvážení výzkumníka, kterou zvolí. Změna referenční kategorie nemá vliv na změnu tvaru modelu, ale ovlivňuje interpretaci odhadů parametrů v modelu.

V našem příkladu byla jako referenční kategorie zvolena skupina Adenakarcinom. Tudíž se budeme zabývat modelováním 2 hlavních srovnání. Chceme srovnat pacientky s výstupní proměnnou Adenosquamózní typ (kategorie 1) a pacientky s výstupní proměnnou Adenokarcinom (kategorie 0). Také budeme srovnávat pacientky s výstupem 2 a 0.

Pokud budeme uvažovat tato dvě srovnání, přibližný poměr šancí OR můžeme vypočítat z dané tabulky. Poměr šancí srovnávající Adenosquamózní typ a Adenokarcinom je výsledek podílu dvou součinů, podobně vypočítáme i poměr šancí pro kategorie Jiný typ a Adenokarcinom, tj.

$$\widehat{OR}_{1vs.0} = \frac{77 \times 34}{109 \times 11} = 2,18,$$

$$\widehat{OR}_{2vs.0} = \frac{77 \times 39}{109 \times 18} = 1,53.$$

V multinomické logistické regresi se třemi výstupními proměnnými musíme použít dvě logitové transformace modelu. A protože v našem příkladě máme tři kategorie výstupní proměnné a jednu vstupní proměnnou (X_1 věk), tak model vyžaduje i dvě regresní vyjádření

$$(1) \qquad \log\left[\frac{P(D=1|X_1)}{P(D=0|X_1)}\right] = \alpha_1 + \beta_{11}X_1, \tag{22}$$

$$(2) \qquad \log\left[\frac{P(D=2|X_1)}{P(D=0|X_1)}\right] = \alpha_2 + \beta_{21}X_1. \tag{23}$$

7.2.1 Poměr šancí (OR) se třemi kategoriemi

Potřebujeme spočítat dva poměry šancí. Jeden porovnávající kategorii 1 (Adenosquamózní typ) s kategorií 0 (Adenokarcinom) a jeden porovnávající kategorii 2 (Jiný typ) s kategorií 0 (Adenokarcinom). Ty spočítáme podobným způsobem jako u standardní logistické regrese

$$OR_1 = \frac{[P(D=1|X_1=1)/P(D=0|X_1=1)]}{[P(D=1|X_1=0)/P(D=0|X_1=0)]}, \tag{24}$$

$$OR_2 = \frac{[P(D=2|X_1=1)/P(D=0|X_1=1)]}{[P(D=2|X_1=0)/P(D=0|X_1=0)]}. \tag{25}$$

Při použití výše definovaných logitových transformací a nahrazení dvou hodnot X_1 hodnotami 0 a 1 můžeme vidět, že OR_1 se rovná $e^{\beta_{11}}$ a OR_2 se rovná $e^{\beta_{21}}$, tj.

$$OR_1 = \frac{\exp[\alpha_1 + \beta_{11}(1)]}{\exp[\alpha_1 + \beta_{11}(0)]} = e^{\beta_{11}} \tag{26}$$

$$OR_2 = \frac{\exp[\alpha_2 + \beta_{21}(1)]}{\exp[\alpha_2 + \beta_{21}(0)]} = e^{\beta_{21}} \tag{27}$$

Speciální případ binomického prediktoru můžeme zobecnit tak, aby zahrnoval kategoriální nebo spojité prediktory. Pro srovnání jakýkoliv dvou skupin prediktorů ($X_1 = X_1^{**}$ vs. $X_1 = X_1^{*}$) je vzorec pro poměr šancí OR roven

$$OR_g = \exp[\beta_{g1}(X_1^{**} - X_1^{*})] \tag{28}$$

kde g indikuje skupinu závisle (výstupní) proměnné (kategorie 1 nebo 2) srovnávanou s referenční proměnnou (kategorie 0).

7.2.2 Počítačový výstup

Výsledky pro polynomický model zkoumání histologického podtypu a věku pacienta jsou prezentovány v tabulce 4. Výsledky jsou získány úlohou Logistic Regression využívající proceduru LOGISTIC ve statistickém softwaru SAS.

Proměnná	Odhad	Standardní chyba	Značení
Intercept 1	-1,4534	0,2618	$\hat{\alpha}_2$
Intercept 2	-1,9459	0,3223	$\hat{\alpha}_1$
AGEGP	0,4256	0,3215	$\hat{\beta}_{21}$
AGEGP	0,7809	0,3775	$\hat{\beta}_{11}$

Tabulka 4: Výsledky polynomického modelu

Jsou zde 2 sady odhadů parametrů odpovídající dvěma logitovým transformacím (22), (23). Výstup je seřazen v sestupném pořadí s $\hat{\alpha}_2$ označené jako Intercept 1 a $\hat{\alpha}_1$ označené jako Intercept 2. Jestliže by D = 2 byla navržena jako referenční kategorie, výstup by pak měl vzestupné pořadí. Vysvětlující proměnná je označena AGEGP.

Rovnice logitové transformace modelu kategorie 2 vs. kategorie 0 má tvar

$$\log\left[\frac{P(D=2|X_1)}{P(D=0|X_1)}\right] = -1,4534 + (0,4256)AGEGP.$$

Exponováním odhadu regresního parametru $\hat{\beta_{21}}$ pro věkovou kategorii (AGEGP) v tomto modelu poskytuje odhadovaný poměr šancí 1,53, tj.

$$\widehat{OR}_2 = e^{\hat{\beta}_{21}} = e^{0.4256} = 1.53.$$

Druhá rovnice logitové transformace modelu kategorie 1 vs. kategorie 0 má tvar

$$\log\left[\frac{P(D=1|X_1)}{P(D=0|X_1)}\right] = -1,9459 + (0,7809)AGEGP.$$

Exponováním odhadu regresního parametru $\hat{\beta_{11}}$ pro věkovou kategorii (AGEGP) v tomto modelu poskytuje odhadovaný poměr šancí 2,18, tj.

$$\widehat{OR}_1 = e^{\hat{\beta}_{11}} = e^{0.7809} = 2,18.$$

Všimněme si, že poměry šancí z polynomického modelu jsou stejné jako ty, které jsme počítali na začátku z dat z tabulky před modelováním. Ve speciálním

případě, kde je jen jedna binomická vstupní proměnná, je hrubý odhad poměru šancí shodný s odhadem získaným z polynomického modelu (nebo z modelu standardní logistické regrese).

Poměry šancí můžeme interpretovat tak, že pro ženy staršího věku (65-79 let), kterým byla diagnostikována rakovina děložní sliznice, v porovnání k ženám mladšího věku (50-64 let) je více pravděpodobné, že jejich tumor bude kategorizován jako kategorie Jiný typ než jako Adenokarcinom ($\widehat{OR}_2 = 1,53$) a je ještě více pravděpodobné, že jejich tumor bude klasifikován jako Adenosquamózní typ než jako Adenokarcinom ($\widehat{OR}_1 = 2,18$).

7.3 Posuzování modelu a výsledků se třemi kategoriemi

Výsledky získané využitím modelu multinomické logistické regrese se třemi kategoriemi výstupní proměnné lze hodnotit podobně jako u standardní logistické regrese s binární výstupní proměnnou. Kromě odhadů OR a jejich konfidenčních intervalů, nás zajímají i výsledky testování hypotéz o parametrech modelu.

7.3.1 95% interval spolehlivosti pro OR

Výpočet intervalů spolehlivosti pro OR je analogický situaci ze stardardní logistické regrese. Pro jednu vysvětlující proměnnou s úrovněmi X_1^{**} a X_1^* z našeho příkladu má vzorec pro výpočet hranic 95% intervalu spolehlivosti klasický tvar

$$\exp\left\{\hat{\beta}_{g1}(X_1^{**} - X_1^*) \pm u\left(1 - \frac{\alpha}{2}\right)(X_1^{**} - X_1^*)s_{\hat{\beta}_{g1}}\right\}, \qquad g = 1, 2. \tag{29}$$

S využitím výsledků získaných úlohou Logistic Regression v softwaru SAS (viz tabulka 3), kde standardní chyby pro odhad parametrů věkové skupiny (AGEGP) jsou $s_{\hat{\beta}_{21}}$=0,3215 a $s_{\hat{\beta}_{11}}$=0,3775, můžeme určit dva 95 % intervaly spolehlivosti

$$OR_2 \in \left\langle e^{(-0,20454)}; e^{1,05574}\right\rangle = (0,82; 2,87),$$

$$OR_1 \in \left\langle e^{0,041}; e^{1,5208}\right\rangle = (1,04; 4,58).$$

7.3.2 Test poměrem věrohodností (LR test)

Stejně jako u standardní logistické regrese i tady můžeme použít test poměru věrohodností - LR test pro posouzení významu vysvětlující proměnné v našem modelu. Musíme mít na paměti, že spíše než testování jednoho neznámého parametru pro vysvětlující proměnné, se nyní testují dva parametry současně, tj. vektor parametrů (pro každé porovnání $D = 2$ vs. $D = 0$ a $D = 1$ vs. $D = 0$ jeden parametr). Tato skutečnost souvisí se stupněm volnosti spojeným s testem.

V našem příkladu máme výstupní proměnnou se třemi kategoriemi a pouze jednu vysvětlující proměnnou. V modelu máme 2 absolutní členy (α_1, α_2) a 2 beta koeficienty (β_{11}, β_{12}).

Pokud nás zajímá testování významnosti parametrů β_{11}, β_{12}, začneme plným modelem

$$\log\left[\frac{P(D=g|X_1)}{P(D=0|X_1)}\right] = \alpha_g + \beta_{g1}X_1, \qquad g = 1, 2$$

a poté ho srovnáme s redukovaným modelem obsahujícím pouze absolutní člen

$$\log\left[\frac{P(D=g)}{P(D=0)}\right] = \alpha_g, \qquad g = 1, 2.$$

V nulové hypotéze předpokládáme, že parametry β_{11} a β_{12} jsou oba rovny nule, tj.

$$H_0 : \beta_{11} = \beta_{21} = 0.$$

Test poměrem věrohodností se vypočítá podle následného vztahu

$$-2\log L_{reduced} - (-2\log L_{full}) \sim \chi^2_2,$$

kde $L_{reduced}$ je věrohodnostní funkce redukovaného modelu a L_{full} je věrohodnostní funkce plného modelu.

Výsledek má χ^2 rozdělení, stupně volnosti se rovnají počtu parametrů z nulové hypotézy.

Aplikujeme-li opět tento postup na náš příklad, dostaneme hodnotu 514,4 pro redukovaný model a 508,9 pro plný model polynomické logistické regrese. Rozdíl je tedy 5,5. P-hodnota pro tuto hodnotu testovací statistiky s rozdělením χ^2_2 je 0,06. Došli jsme k závěru, že AGEGP je statisticky významné na hladině významnosti 10%, ale ne na 5%-ní hladině významnosti.

7.3.3 Waldův test

LR test umožňuje hodnocení vlivu vysvětlující proměnné na všechny úrovně vysvětlované proměnné současně. Je ovšem možné, že vysvětlující proměnná by mohla mít významný vliv jen vzhledem k jedné úrovni vysvětlované proměnné. V této situaci můžeme provést Waldův test.

Stanovíme nulovou hypotézu pro každou úroveň - testujeme nulovost parametrů β_{11} a β_{12}, tj.

$$H_0 : \beta_{11} = 0, \qquad \text{resp.} \qquad H_0 : \beta_{21} = 0.$$

Testovací statistika Waldova testu je počítána stejně jako je uvedeno v kapitole 6.1.2 a má přibližně normované normální rozdělení, tj.

$$Z = \frac{\hat{\beta}}{s_{\hat{\beta}}} \sim N(0, 1).$$

Jestliže v našem ilustrativním příkladu testujeme nulovou hypotézu pro srovnání kategorie Adenosquamózní typ vs. Adenokarcinom (tedy kategorie 1 vs. 0), tj. $\beta_{11} = 0$, má Waldova statistika hodnotu 2,07 a odpovídající P-hodnota je rovna 0,04.

Dále pro nulovou hypotézu pro srovnání kategorie Jiný typ vs. Adenokarcinom (tedy kategorie 2 vs. 0), tj. $H_0 : \beta_{21} = 0$, je hodnota Waldovy statistiky rovna 1,32 a odpovídající P-hodnota 0,19.

Na hladině významnosti 0,05 zamítáme nulovou hypotézu $H_0 : \beta_{11} = 0$, ale nezamítáme $H_0 : \beta_{21} = 0$. Došli jsme k závěru, že kategorie věku (proměnná AGEGP) je statisticky významná pro srovnání kategorie 1 vs. 0. Ale není významná pro srovnání kategorie 2 vs. 0.

Při modelování multinomické logistické regrese musíme buď oba parametry odpovídající vysvětlující proměnné β_{11}, β_{12} zachovat, nebo oba vypustit. I když je jen jeden z parametrů významný, musí být oba parametry zachovány, pokud má vysvětlující proměnná v modelu zůstat.

7.4 Zobecnění modelu polymonické regrese na G výstupů a k prediktorů

7.4.1 Rozšíření na k prediktorů

Rozšíření modelu v našem ilustrativním příkladu na k vysvětlujících proměnných je poměrně jednoduché. Můžeme přidat k proměnných ke každému srovnání. Logitová transformace srovnání kategorie 1 a kategorie 0 je rovna α_1 plus součet všech k nezávislých proměnných krát jejich β_1 koeficienty. Podobně i u druhého srovnání (kategorie 2 a 0).

$$(1) \qquad \log\left[\frac{P(D=1|\mathbf{X})}{P(D=0|\mathbf{X})}\right] = \alpha_1 + \sum_{i=1}^{k} \beta_{1i} X_i, \tag{30}$$

$$(2) \qquad \log\left[\frac{P(D=2|\mathbf{X})}{P(D=0|\mathbf{X})}\right] = \alpha_2 + \sum_{i=1}^{k} \beta_{2i} X_i. \tag{31}$$

Postup pro počítání poměrů šancí, intervalů spolehlivosti a pro testování nulových hypotéz zůstává stejný.

Konkrétně předpokládejme, že chceme uvážit účinky užívání estrogenu a kouření, stejně jako jsme uvažovali věk pacientek, tj. to, zda mají vliv na histologický podtyp rakoviny ($D = 0, 1, 2$). Model tedy nyní obsahuje 3 prediktory: $X_1 =$ AGEGP, $X_2 =$ ESTROGEN a $X_3 =$ SMOKING.

Obě nové vysvětlující proměnné jsou kódovány opět jako binomické proměnné. Proměnná ESTROGEN je označená jako 1 pro ty ženy, které ho někdy užily a 0 pro ty, co ho nikdy neužily. Proměnná SMOKING je označená jako 1 pro pravidelné kuřačky a 0 pro nekuřačky.

Logitová transformace srovnání kategorie Adenosquamózní typ ($D = 1$) s typem Adenokarcinom ($D = 0$) je rovna následujícímu vzorci

$$(1) \qquad \log\left[\frac{P(D=1|\mathbf{X})}{P(D=0|\mathbf{X})}\right] = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3. \tag{32}$$

Podobně se vytvoří i logitová transformace srovnání kategorie Jiný typ ($D = 2$) s typem Adenokarcinom ($D = 0$)

$$(2) \qquad \log\left[\frac{P(D=2|\mathbf{X})}{P(D=0|\mathbf{X})}\right] = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3. \tag{33}$$

U příkladu, z kterého jsem čerpal buhužel nebyla uvedená data, byl uveden pouze výstup v programu SAS (KLEINBAUM, David G. and Mitchel KLEIN, 2010).

Výstup v programu SAS je ukázán v tabulce č. 5. Jsou zde 2 parametry β pro každý prediktor v modelu. Model tedy obsahuje 8 parametrů včetně absolutních členů.

Proměnná	Odhad	Odchylka	Značení
Intercept 1	-1,2032	0,3190	$\hat{\alpha}_2$
Intercept 2	-1,8822	0,4025	$\hat{\alpha}_1$
AGEGP	0,2823	0,3280	$\hat{\beta}_{21}$
AGEGP	0,9871	0,4118	$\hat{\beta}_{11}$
ESTROGEN	-0,1071	0,3067	$\hat{\beta}_{22}$
ESTROGEN	-0,6439	0,3436	$\hat{\beta}_{12}$
SMOKING	-1,7913	1,0460	$\hat{\beta}_{23}$
SMOKING	0,8895	0,5254	$\hat{\beta}_{13}$

Tabulka 5: Výsledky polynomického modelu se třemi prediktory

Předpokládejme, že nás zajímá účinek vysvětlující proměnné X_1 (AGEGP), za konstantních hodnot proměnných X_2 (ESTROGEN) a X_3 (SMOKING). Poměr šancí pro účinek proměnné X_1 (AGEGP) pro srovnání kategorie 1 vs. 0 je roven

$$\widehat{OR}_1 = \frac{\exp[\hat{\alpha}_1 + \hat{\beta}_{11}(1) + \hat{\beta}_{12}(X_2) + \hat{\beta}_{13}(X_3)]}{\exp[\hat{\alpha}_1 + \hat{\beta}_{11}(0) + \hat{\beta}_{12}(X_2) + \hat{\beta}_{13}(X_3)]} = e^{\hat{\beta}_{11}} = e^{0,9871} = 2,68$$

Poměr šancí pro účinek vysvětlující proměnné X_1 (AGEGP) pro srovnání kategorie 2 vs. 0 je na základě analogického výpočtu roven 1,33.

Interpretace výsledků pro logistický model se třemi prediktory se liší od modelu s jedním prediktorem. Vliv věku na typ rakoviny je nyní odhadován při současném vlivu užití estrogenu a kouření. Jestliže srovnáme oba modely, vliv věku v redukovaném modelu (jen s proměnnou X_1) je slabší pro srovnání Adenosquamózní typ a Adenokarcinom ($\widehat{OR}$ = 2,18 vs. 2,68), ale je silnější pro srovnání Jiný typ a Adenokarcinom ($\widehat{OR}$ = 1,53 vs. 1,33).

Tyto výsledky naznačují, že užití estrogenu a kouření působí jako zkreslující faktory ve vztahu mezi věkovou skupinou a typem rakoviny děložní sliznice.

7.4.2 95% interval spolehlivosti

Intervaly spolehlivosti jsou nyní v našem ilustrativním příkladu počítány použitím standardní chyby odhadu parametrů z modelu se třemi prediktory, tedy 0,4118 pro $\hat{\beta}_{11}$ a 0,3280 pro $\hat{\beta}_{12}$. Tyhle intervaly spolehlivosti počítáme obvyklým vzorcem a pro náš příklad tedy dostáváme

$$OR_1 \in \left\langle e^{0,179972}; e^{1,794228} \right\rangle = (1,20; 6,01),$$

$$OR_2 \in \left\langle e^{(-0,35968)}; e^{0,92608} \right\rangle = (0,70; 2,52).$$

Analogické výpočty bychom provedli pro parametry odpovídající dalším proměnným v modelu.

7.4.3 Test poměrem věrohodností a Waldův test

Postupy pro oba testy jsou stejné jako v předchozích kapitolách, tedy jako pro model multinomické logistické regrese s jednou vysvětlující proměnnou.

Testem poměru věrohodností dojdeme k následujícím výsledkům,

$$-2\log L_{reduced} - (-2\log L_{full}) = 500,97 - 494,41 = 6,56 \sim \chi^2_2.$$

P-hodnota pro tuto hodnotu testovací statistiky má pro χ^2 rozdělení se 2 stupni volnosti hodnotu 0,04. Došli jsme tedy k závěru, že proměnná X_1 (AGEGP) je statisticky významná na hladině 5%.

Waldův test vypadá následovně:

$$H_0 : \beta_{11} = 0,$$

$$Z = \frac{0,9871}{0,4118} = 2,40, \qquad P = 0,02,$$

$$H_0 : \beta_{21} = 0,$$

$$Z = \frac{0,2832}{0,3280} = 0,86, \qquad P = 0,39.$$

Proto na hladině významnosti 5% zamítáme nulovou hypotézu $H_0 : \beta_{11} = 0$, ale nezamítáme $H_0 : \beta_{21} = 0$. Došli jsme tedy k závěru, že kategorizovaný věk pacientek X_1 (AGEGP) je statisticky významný pro srovnání kategorie 1 vs. 0, ale není významný pro srovnání kategorie 2 vs. 0.

Výzkumník se nyní musí rozhodnout, zda proměnnou X_1 (AGEGP) v modelu zachová. Pokud bychom se zajímali o obě srovnání, pak oba parametry β_{11} a β_{21} musí být zachovány, i když je jen jeden parametr statisticky významný.

7.4.4 Rozšíření modelu na G výstupů

Předpokládejme, že výstup má G kategorií (0, 1, 2, ..., G - 1). Nyní zde máme G - 1 možností srovnání s referenční kategorií. Jestliže jsme jako referenční

kategorii zvolili 0, můžeme definovat model multinomické logistické regrese v následující formě

$$\log\left[\frac{P(D=g|\mathbf{X})}{P(D=0|\mathbf{X})}\right] = \alpha_g + \sum_{i=1}^{k} \beta_{gi} X_i, \qquad g = 1, 2, \ldots, G-1. \tag{34}$$

Poměry šancí a odpovídající intervaly spolehlivosti pro G - 1 srovnání $G-1$ kategorií výstupní (závislé) proměnné s referenční kategorií 0 jsou počítány stejně jako v předchozích kapitolách. Máme nyní G - 1 odhadů poměrů šancí a odpovídajících intervalů spolehlivosti pro vliv každé vysvětlující proměnné v modelu.

7.4.5 Test poměrem věrohodností a Waldův test

Oba testy jsou opět počítány obdobně jako v předchozích kapitolách.

Pro test poměrem věrohodností testujeme G - 1 odhadů parametrů současně pro každou vysvětlující proměnnou. Proto pro testování jedné vysvětlující proměnné máme G - 1 stupňů volnosti asymptotické distribuce χ^2 testové statistiky porovnávající redukovaný a plný model:

$$-2\log L_{reduced} - (-2\log L_{full}) \sim \chi^2_{G-1}$$

Můžeme také počítat Waldův test pro zjištění významnosti jednotlivých parametrů β_{g1}, $g = 1, 2, \ldots, G-1$. Máme G - 1 koeficientů, které můžeme testovat pro každou vysvětlující proměnnou. Stejně jako dříve sada koeficientů odpovídající jedné vysvětlující proměnné musí být buď zachována nebo vypuštěna. Testová statistika Waldova testu pro hypotézu $H_0 : \beta_{gi} = 0$, $g = 1, 2, \ldots, G-1.$, $i = 1, \ldots, k$

$$Z = \frac{\hat{\beta}_{gi}}{s_{\hat{\beta}_{gi}}} \sim N(0,1).$$

7.5 Porovnání multinomické a mnohonásobné standardní logistické regrese

Věrohodnostní funkce pro model multinomické logistické regrese využívá data zahrnující všechny kategorie výstupní proměnné v jedné struktuře. Naproti tomu věrohodnostní funkce pro model binomické (standardní) logistické regrese využívá data zahrnující jen dvě kategorie výstupní kategoriální proměnné. Jinými slovy, věrohodnostní funkce používané při odhadu parametrů (fitaci) každého samostatného binomického modelu odpovídajícího dvěma porovnávaným kategoriím se liší od věrohodnostních funkcí používaných při fitaci multinomického modelu, který uvažuje všechny úrovně (kategorie) najednou. V důsledku toho se mohou oba odhady parametrů a standardní chyby odhadů parametrů při srovnání výsledků těchto dvou modelů lišit.

Jen ve speciálním případě multinomické regrese, kdy máme jen jeden dichotimický (binární) prediktor, poskytují mnohonásobné logistické modely stejné odhady parametrů a jejich standardních chyb jako multinomický model (KLEINBAUM, David G. and Mitchel KLEIN, 2010).

8 SAS: Procedura LOGISTIC a SAS EG úloha Logistic Regression

8.1 Popis prostředí SAS Enteprise Guide

V této kapitole bych chtěl čtenáře seznámit se statistickým softwarem SAS, který ve své práci využívám pro praktické příklady a přiblížit jeden z jeho modulů, tzv. tenký klient SAS Enterprise Guide.

SAS je anglická zkratka pro Statistical Analysis System, je to produkt společnosti SAS Institute. Je to plnohodnotný prostředek pro správu, analýzu a prezentaci dat, jehož počátky vývoje spadají do začátku 70. let (North Carolina State University). Řadí se mezi vůbec nejpoužívanější statistické softwary v obchodních i akademických kruzích u nás i v zahraničí.

SAS patří mezi tzv. modulární systémy, tzn. jednotlivé instalace SASu se mohou od sebe podstatně lišit. Modulární skladba systému umožňuje lépe přizpůsobit software potřebám zákazníka. Každý modul systému „umí“ řešit určitý okruh problémů, obsahuje procedury a funkce, které se specializují jen na určitý typ úlohy.

Mimo speciální moduly, jako je např. modul **SAS/OR** pro oblast operačního výzkumu, existují moduly, které najdete snad v každé instalaci, např. modul **SAS/GRAPH** pro grafické úlohy, modul **SAS/ACCESS** pro přístup k databázím.

Při tvorbě programů, u kterých lze předpokládat pozdější přenos z jedné instalace SASu na druhou, je tedy třeba zvážit, jaké konkrétní funkce a procedury budou pro řešení dané úlohy použity, aby se nestalo, že program nebude fungovat, protože určitá procedura či funkce je uložena v modulu, který v dané instalaci chybí. To nehrozí v případě použití procedur a funkcí z modulu **Base SAS**, protože ten musí být součástí každé instalace SASu.

Software SAS na Přírodovědecké fakultě UP Olomouc je v současné době k dispozici jako fakultní licence v rámci tzv. Akademického programu společnosti SAS. Kromě základního balíku modulů SAS je k dispozici ještě doplňkový modul

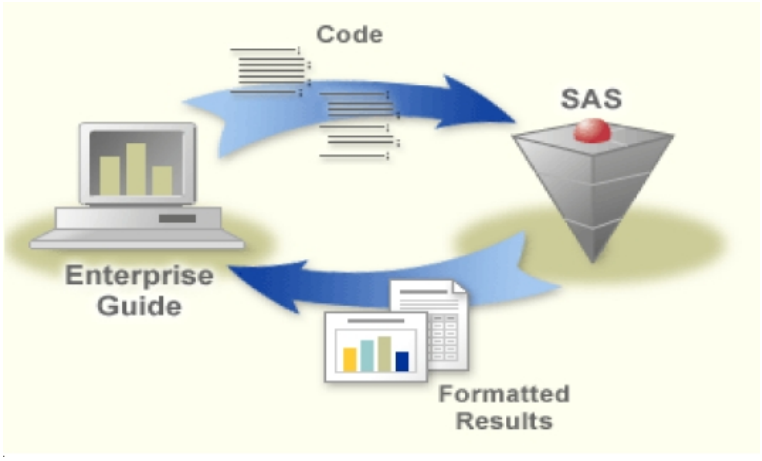
SAS JMP, což je samostatná aplikace umožňující provádění všech základních statistických analýz.

8.1.1 SAS Enterprise Guide

SAS EG není samostatná aplikace. Je to jeden z modulů SASu, tzv. tenký klient, který umožňuje přístup k většině funkcí SASu. Je to vlastně jen jakési intuitivní, vizuální, přizpůsobitelné rozhraní k vlastní instalaci softwaru SAS, v prostředí SAS EG nazývané SAS Server.

SAS EG nemůže bez vlastní instalace SASu vůbec fungovat. Poskytuje transparentní přístup k datům přes pohodlné a přehledné „klikací“ prostředí, ve kterém jsou úlohy pro jednotlivé analýzy připravené ihned k použití. Umožňuje snadný export dat do dalších aplikací. Neomezuje se jen na předdefinované úlohy, ale díky možnosti tvorby vlastních uživatelských skriptů (= SAS programů) a editace již vytvořeného kódu (uživatelem či automaticky předdefinovanou úlohou) se stává univerzálním prostředím pro práci se softwarem SAS.

SAS EG funguje skutečně jako jakýsi „klient“, jehož prostřednictvím vytvoříme SAS program, který je následně odeslán k provedení na SAS Server a výsledky jsou zase zpětně zobrazeny v přehledné formě uživateli klientem SAS EG, viz obrázek 6.



Obrázek 6: Schéma zpracování dat přes SAS Server

Všechny operace v SASu (načítání a manipulace s daty, analýza dat atd.) jsou řízeny programovým kódem.

Program v SASu:

- je sekvencí příkazů prováděných po řadě tak, jak jsou zapsány,
- každý příkaz musí být ukončen středníkem a každý program musí být ukončen příkazem RUN; (někdy je vyžadováno QUIT;),
- příkazy mohou být zapsány velkými a malými písmeny (SAS není tzv. Case sensitive), rozděleny do více řádků a začínat v libovolném sloupci.

SAS načítá data pro provádění analýz ze speciální struktury – SAS data set. SAS data set obsahuje popisné informace, indexy a vlastní data organizovaná podobně jako v relačních databázových systémech, tj. řádky odpovídají pozorováním a sloupce odpovídají proměnným (max. 32767 v jednom data setu). SAS data sety mohou být dočasné (temporary) a stálé (permanent). Platí, že dočasné datové množiny existují pouze od okamžiku svého vytvoření po dobu do ukončení sezení, tj. ukončení programu SAS či SAS EG.

Vlastní data v datasetu mohou být pouze dvojího typu, poněvadž SAS rozeznává pouze numerický datový typ a znakový datový typ:

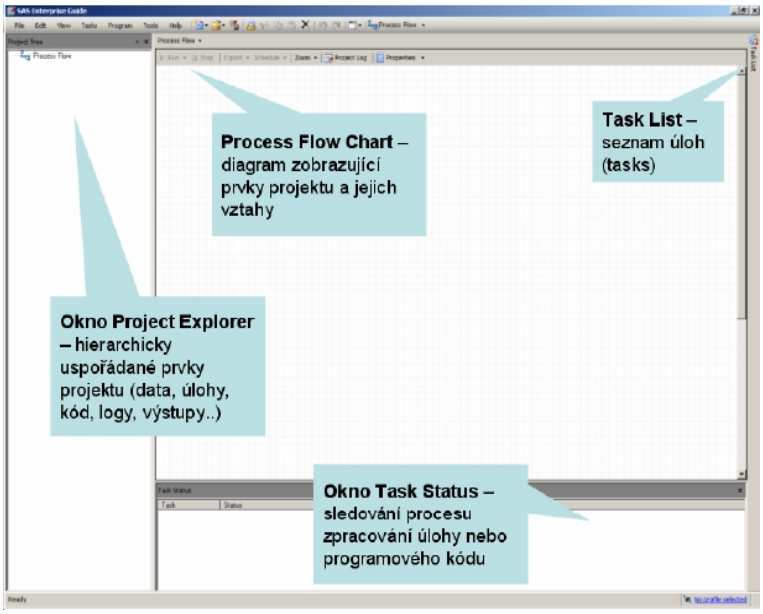
- numerický datový typ reprezentuje kladné nebo záporné číslo, ve kterém se desetinná místa oddělují tečkou, pro oddělení mantisy a exponentu se používá znak E a chybějící hodnota je reprezentována tečkou,
- data znakového datového typu mohou obsahovat číslice i znaky (včetně speciálních, např. $ nebo !), přičemž maximální délka řetězce může být 32767 znaků a chybějící hodnota je reprezentována prázdnou buňkou.

Úloha Import Data v SAS EG umožňuje import nejen tzv. řádkových dat, což jsou obyčejné textové soubory bez formátování, ale i soubory jiných formátů, např. HTML, MS Excel, MS Access atd. U řádkových dat rozlišujeme soubory s pevnou šířkou sloupce a soubory s oddělovači.

Program v SASu má obecně dvě hlavní části – tzv. DATA step a PROC step(s). Část DATA step zajišťuje tvorbu a modifikaci data setů. Část PROC

step(s) řídí zpracování a analýzu dat z data setů prostřednictvím jednotlivých procedur SASu. Programy se ukládají jako soubory s příponou *.sas*.

SAS Enterprise Guide je klasická „okenní“ aplikace. Nachází se zde řádek nabídek, panel nástrojů a několik podoken, viz obrázek 7.



Obrázek 7: Prostředí SAS EG

Po otevření aplikace se systém optá, zda plánujete otevřít již existující nebo vytvořit nový projekt, protože v SAS Enterprise Guide je práce uživatele ukládána vždy do projektu.

Projekt = kolekce dat, úloh (tasks), programového kódu a výstupů. Projekt (a tím i celou práci) je možné uložit jako soubor s příponou *.egp*.

Prvky, které jsou součástí daného projektu (programový kód, log, atd.) lze prohlížet v okně **Project Tree**. **Process Flow** (diagram procesu) zobrazuje jednotlivé prvky projektu a vazby mezi nimi.

Objekty se do diagramů vkládají automaticky po jejich vytvoření, přičemž pořadí objektů určuje i jejich pořadí po spuštění procesu. Za účelem změny pořadí zpracování jednotlivých úloh lze pořadí měnit.

8.1.2 Procedura LOGISTIC

Ve své práci jsem používal zejména úlohu Logistic Regression, která na pozadí využívá jednu z mnoha procedur, kterou SAS nabízí - proceduru LOGISTIC. Tato procedura je součástí modulu **SAS/STAT**.

Procedura LOGISTIC poskytuje nástroj pro analýzu vztahů mezi jednou diskrétní - binární, ordinální nebo nominální závisle proměnnou (tzv. odpovědí - response) a jednou či více vysvětlujícími proměnnými (tzv. efekty - effects).

Prostřednictvím této procedury lze řešit i model podmíněné logistické regrese pro binární závisle proměnnou nebo použít exaktní logistickou regresi.

Dále se pokusím zmínit, dle mého názoru, nejdůležitější aspekty použití procedury LOGISTIC. V příloze č. 4 uvádím kompletní syntaxi této procedury. Celou nápovědu lze získat zdarma na internetových stránkách společnosti SAS Institute (The LOGISTIC Procedure: Syntax, 2012).

Logistická regrese patří mezi zobecněné lineární modely (viz kapitola 1), přičemž lze využít z nabídky 4 linkových funkcí. Funkce logit je výchozí. Pro specifikaci jiné linkové funkce použijeme volbu LINK= v příkazu MODEL (viz tabulka 6)

Volba	Popis
LOGIT	logit: $g(\mu) = \log(\mu/(1-\mu))$
PROBIT	probit: $g(\mu) = \Phi^{-1}(\mu)$
CLOGLOG	complementary log-log: $g(\mu) = \log(-\log(1-\mu))$
GLOGIT	generalized logit: $g(\mu) = \log(\mu_i/\mu_{k+1})$, $i = 1, \ldots, k$

Tabulka 6: Nabídka 4 linkových funkcí

Program SAS dává na výběr ze dvou základních iteračních algoritmů, metodu Fisherova skórování (FISHER, nastavena jako výchozí) a Newton-Raphsonovu (NEWTON) metodu. Obě metody dávají stejný odhad parametrů. Nicméně odhadovaná kovarianční matice odhadů se mírně liší. To je způsobeno tím, že Fisherovo skórování je založeno na očekávané informační matici, zatímco Newton-Raphsonova metoda je založena na pozorované informační matici (viz kapitola 5.3). V případě binárního logitového modelu pozorované a očekávané informační matice jsou totožné. Metody specifikujeme pomocí volby TECHNIQUE= v pří-

kazu MODEL. V případě separace vstupních dat můžeme zvolit volbu FIRTH, pro použití Firthovy penalizační metody (viz kapitola 5.3.3). Pro zobecněné logit modely je dostupná pouze Newton-Raphsonova metoda. Ve výchozím nastavení jsou počáteční hodnoty parametrů modelu nulové. Mohou být specifikovány volbou INEST= v příkazu PROC LOGISTIC.

Pro iterační výpočet jsou k dispozici 4 volby kritéria konvergence v příkazu MODEL, viz tabulka 7.

Volba	Popis
ABSFCONV = hodnota	konv. kritérium absolutní funkce; iterační proces je ukončen jestliže: $\lvert l_i - l_{i-1}\rvert <$ hodnota, kde l_i je hodnota věrohod. fce v i-té iteraci
FCONV = hodnota	konv. kritérium relativní funkce; iterační proces je ukončen jestliže: $\frac{\lvert l_i - l_{i-1}\rvert}{\lvert l_{i-1}\rvert + 1E{-}6} <$ hodnota, kde l_i je hodnota věrohod. fce v i-té iteraci
GCONV = hodnota	konv. kritérium relativního gradientu; iterační proces je ukončen jestliže: $\frac{\boldsymbol{g'}_i \boldsymbol{I}_i^{-1} \boldsymbol{g}_i}{\lvert l_{i-1}\rvert + 1E{-}6} <$ hodnota, kde $\boldsymbol{g}_i$ je gradient a $\boldsymbol{I}_i$ je Hessova matice
XCONV = hodnota	konv. kritérium relativního parametru; iterační proces je ukončen jestliže: $\max_j \left\lvert \delta_j^{(i)} \right\rvert <$ hodnota, kde $\delta_j^{(i)} = \begin{cases} \beta_j^{(i)} - \beta_j^{(i-1)} & \lvert \beta_j^{(i-1)} < 0,01 \rvert \\ \frac{\beta_j^{(i)} - \beta_j^{(i-1)}}{\beta_j^{(i-1)}} & \text{jinak} \end{cases}$, $\beta_j^{(i)}$ je odhad j-tého parametru v i-té iteraci

Tabulka 7: Konvergenční kritéria

Zadáme-li více než jedno konvergenční kritérium, optimalizace je ukončena, jakmile je jedno z uvedených kritérií splněno. Pokud není uvedené žádné z kritérií, výchozí hodnota je GCONV = 1E-8.

Pravděpdobnostní rovnice pro logistické regresní modely nemusí mít vždy konečné řešení. Procedura LOGISTIC používá jednoduchý postup rozpoznání kon-

figurace dat, která vede k nekonečným odhadům parametrů. Pokud je dosaženo konvergence v osmi nebo méně iteracích, kontrola pro kompletní nebo kvazi-kompletní separaci neprobíhá. V návaznosti na osmou iteraci je počítána pravděpodobnost sledované závisle proměnné (response) pro každé pozorování. Pokud se předpovídaný výstup rovná pozorovanému výstupu pro každé pozorování, existuje kompletní separace dat a iterační proces je zastaven. Jestliže není detekována kompletní separace dat a pozorování má vysokou ($\geq$ 0,95) predikovanou pravděpodobnost, že nabyde pozorované hodnoty, mohou nastat 2 situace. Za prvé, v datech existuje překrytí a pozorování je pro svou skupinu atypické. V tomto případě iterační proces pokračuje a je zastaven při dosažení maxima. Druhá situace - kvazi-kompletní separace dat. Je-li libovolný prvek diagonály matice disperze pro standardizovaný vektor pozorování (všechny vysvětlující proměnné mají střední hodnotu 0 a roztyl 1) větší než 5000, iterační proces je zastaven.

Pokud je zjištěna kompletní nebo kvazi-kompletní separace dat, zobrazí se varování na výstupu procedury. Proces kontroly můžeme vypnout volbou NONCHECK v příkazu MODEL.

SAS nabízí pět možností výběru efektů. Můžeme je specifikovat pomocí volby SELECTION= v příkazu MODEL. A jsou následující: NONE (kompletní model zahrnující všechny efekty, výchozí možnost), FORWARD (postupně přidává k absolutnímu členu efekty, které jsou významné na hladině testu specifikované volbou SLENTRY=), BACKWARD (postupně z kompletního modelu vyřazuje efekty, které nejsou významné na hladině testu specifikované volbou SLSTAY=), STEPWISE (kombinace předchozích dvou metod), SCORE (používá Furnivalův a Wilsonův algoritmus pro nalezení specifického počtu modelů s nejvyšší χ^2 statistikou pro všechny možné velikosti modelu, od modelu s jedním efektem až po model se všemi vysvětlujícími efekty).

Procedura LOGISTIC vypočte a zobrazí 3 kritétia vhodnosti modelu:

– -2 Log Likelihood:

$$-2LogL = -2\sum_j \frac{w_j}{\sigma^2} f_j \log(\hat{\pi}_j),$$

kde w_j a f_j jsou váhy a četnost hodnot j-tého pozorování, $\hat{\pi}_j$ značí odhadovanou pravděpodobnost, σ^2 je parametr rozptylu, který se rovná 1, pokud není specifikován jinak volbou SCALE=.

– Akaikeho informační kritérium:

$$AIC = -2LogL + 2p,$$

kde p je počet parametrů v modelu.

– Schwarzovo kritérium:

$$SC = -2LogL + p \log\left(\sum_j f_j\right),$$

kde p je počet parametrů v modelu.

AIC a SC se používá pro pozorování několika modelů. Model, který má nižší hodnotu je lepší.

Rozdíl mezi -2 Log L statistikami pro model pouze s absolutními členy a pro úplný model má χ^2 rozdělení s $p - k$ stupni volnosti. Pak testujeme nulovou hypotézu, že všechny vysvětlující efekty jsou rovny nule, kde p je počet parametrů v úplném modelu a k je počet absolutních členů v modelu. Test poměrem věrohodností v tabulce nazvané „Testing Global Null Hypothesis: BETA = 0“ zobrazuje tento rozdíl a odpovídající P-hodnotu pro daný test.

Jsou k dispozici dvě metody výpočtu intervalů spolehlivosti pro regresní parametry. Jedna z nich je založena na iteračním výpočtu a ta druhá je založena na asymptotické normalitě odhadů parametrů. Ta není tak časově náročná jako první iterační metoda, ale není tak přesná, zejména při malém počtu vzorků. Metodu výpočtu intervalů spolehlivosti pro parametry modelu zadáme volbou CLPARM= v příkazu MODEL. Standardně se konfidenční intervaly počítají neiterační metodou.

Příkazem ODDSRATIO nastavujeme požadavek výpočtu poměrů šancí OR pro jednotlivé proměnné, a to i tehdy, jsou-li v interakci s dalšími proměnnými.

Je-li proměnná spojitá, potom respektuje hodnotu specifikovanou v příkazu UNITS, kterým nastavujeme velikost jednotky pro výpočet OR. Příkaz má následující syntaxi:

$$UNITS < var1 = list1 < var2 = list2 \cdots >>< \backslash volby >,$$

kde var1, var2 jsou názvy vysvětlujících proměnných a list1, list2 je seznam jednotek oddělených mezerami; každá jednotka v seznamu má jednu z následujících forem:

- číslo
- SD nebo -SD
- číslo SD,

přičemž číslo je různé od 0 a SD je výběrová směrodatná odchylka odpovídající veličiny.

Je-li proměnná kategoriální (klasifikační, specifikovaná v příkazu CLASS), potom OR porovnává šance pro každý pár úrovní (unikátních hodnot proměnné). Pokud je kategoriální proměnná v interakci se spojitou proměnnou, potom je OR vypočteno defaultně vzhledem k aritmetickému průměru této proměnné. Je-li kategoriální proměnná v interakci s nějakou další kategoriální proměnnou, potom je OR standardně vypočteno pro každou úroveň této proměnné. Vypočtené poměry šancí nezávisí na parametrizaci klasifikační proměnné.

Parametrizace klasifikační proměnné se nastavuje volbou PARAM= v příkazu CLASS. Výchozí volbou je hodnota EFFECT. Způsob parametrizace ovlivňuje konstrukci matice $\mathbf{X}$ v logistickém regresním modelu.

Rozdíl si můžeme ukázat na příkladu kategoriální proměnné barva se 4 úrovněmi: černá, bílá, červená a modrá, která vystupuje v logistickém regresním modelu jako vysvětlující proměnná. Je-li tato proměnná kódována jako efekt (PARAM = EFFECT) s hodnotou „modrá“ jako referenční kategorií, potom v modelu vystupují celkem 3 proměnné (vždy o 1 méně než je počet kategorií), viz tabulka 8.

Barva	X_1	X_2	X_3
bílá	1	0	0
černá	0	1	0
červená	0	0	1
modrá	-1	-1	-1

Tabulka 8: Parametrizace klasifikační proměnné kódované jako efekt

Potom

$$\begin{aligned} logit(\text{bílá}) &= \alpha + (X_1 = 1)\beta_1 + (X_2 = 0)\beta_2 + (X_3 = 0)\beta_3 = \\ &= \alpha + \beta_1, \\ logit(\text{modrá}) &= \alpha + (X_1 = -1)\beta_1 + (X_2 = -1)\beta_2 + (X_3 = -1)\beta_3 = \\ &= \alpha - \beta_1 - \beta_2 - \beta_3. \end{aligned}$$

Logaritmus OR pro kategorii „bílá“ vzhledem k referenční kategorii „modrá“ je

$$\begin{aligned} logit(\psi(\text{bílá}, \text{modrá})) &= logit(\text{bílá}) - logit(\text{modrá}) = \\ &= 2\beta_1 + \beta_2 + \beta_3, \end{aligned}$$

kde ψ označuje poměr šancí vzhledem k danému faktoru.

Je-li klasifikační proměnná kódována jako referenční (PARAM = REF, resp. PARAM = REFERENCE), pak je v modelu nahrazena proměnnými dle schématu uvedeného v tabulce 9.

Barva	X_1	X_2	X_3
bílá	1	0	0
černá	0	1	0
červená	0	0	1
modrá	0	0	0

Tabulka 9: Parametrizace klasifikační proměnné kódované jako refenční

Logaritmus OR pro kategorii „bílá“ vzhledem k referenční kategorii „modrá“ je

$$\begin{aligned} logit(\psi(\text{bílá}, \text{modrá})) &= logit(\text{bílá}) - logit(\text{modrá}) = \\ &= \alpha + (X_1 = 1)\beta_1 + (X_2 = 0)\beta_2 + (X_3 = 0)\beta_3 - \\ &- \alpha + (X_1 = 0)\beta_1 + (X_2 = 0)\beta_2 + (X_3 = 0)\beta_3 \\ &= \beta_1. \end{aligned}$$

Pro zobecněný logitový model jsou výpočty poměrů šancí provedeny obdobně ($G - 1$ poměrů šancí vzhledem ke $G - 1$ logitovým modelům pro G kategorií závisle proměnné).

Ve výstupu procedury LOGISTIC jsou bodové odhady OR doplněny standardně konfidenčními intervaly. Bodové odhady i odpovídající konfidenční intervaly jsou odvozeny z odhadů regresních parametrů a jejich intervalů spolehlivosti s využitím exponenciální funkce.

Pro spojitou vysvětlující proměnnou korespondují OR s jednotkovým přírůstkem rizikového faktoru, přičemž jednotku je možné specifikovat příkazem UNITS v proceduře LOGISTIC (viz výše).

9 Praktická část

9.1 Motivace

V následující kapitole ukážu aplikaci výše zmiňovaných vzorců, vztahů a testů v multinomické regresi. Jak už jsem dříve zmiňoval, logistická regrese se převážně používá ve zdravotnictví, zejména pak v epidemiologických studiích. Chtěl bych ale ukázat, že využití této funkce je mnohem širší. Proto jsem si záměrně vybral 2 příklady aplikace logistické regrese, které jsou z úplně jiného oboru. A to ze sportovní a společensko-politické oblasti.

9.1.1 Hráči basketbalu

Základní úlohou tohoto příkladu je pomoci začínajícímu trenérovi basketbalu, který zatím nemá moc zkušeností, vybrat tu správnou herní pozici pro své hráče na základě několika sledovaných charakteristik.

V basketbalu jsou celkem 3 základní herní posty:

- **rozehrávač** = hráč většinou menšího vzrůstu, velmi pohyblivý a rychlý, s čímž souvisí dobrá kondiční připravenost; měl by mít dobré periferní vidění a vlastnost řídit hru,
- **křídlo** = hráč, u nějž vzrůst nehraje velkou roli; fyzická a kondiční připravenost by měla být co nejvíce vyvážená; křídla jsou často nejlepší střelci družstva, tedy psychická odolnost je nezbytnou vlastností,
- **pivot** = patří mezi nejvýšší hráče týmu (přes 200 cm); nemusí být tak motoricky nadaný; kondiční příprava nebývá zrovna nejlepší, ale o to musí být fyzicky zdatnější.

Uvažujme situaci, kdy je v týmu nějaký 215 cm vysoký basketbalista. V tomto případě trenér nemusí dlouho váhat a ihned takového hráče postaví na pozici pivota. Pokud ale máme v týmu více hráčů vysokých okolo 198 cm, může mít trenér ze začátku problémy je správně do týmu zařadit. Tito basketbalisté mohou hrát

na pozici křídla, menšího pivota či rozehrávače. Jako takovou výjimku, která potvrzuje pravidlo, mohu uvést jednoho z nejlepších rozehrávačů historie basketbalu Srba Dejana Bodirogu, který měřil 205 cm a vážil 110 kg. Trenér bude muset tyto hráče pozorovat při trénincích a utkáních týdny, aby je pak mohl správně zařadit. Proto jsem se snažil využít metody multinomické logistické regrese a vytvořit model, který by mohl pomoci mladému trenérovi při rozhodování a urychlil tak proces klasifikace hráčů. Budeme uvažovat, že trenér bude vycházet ze zkušeností starších trenérů, kteří už jednotlivé hráče mají rozřazené.

Alternativní úlohou tohoto příkladu, kdy využijeme standardní logisitckou regresi, by pak byla situace, kdy trenér hledá hráče jen na určitý post, např. křídlo. Potom na základě sledovaných charakteristik může rozhodnout, zda hráče na hledanou pozici přijme či ne, popř. jaká charakteristika by měla při výběru hrát větší roli.

9.1.2 Volba prezidenta

V tomto praktickém příkladě jsem zjišťoval šance zvolení jednotlivých kandidátů na post prezidenta ČR u voličů dle různých socio-ekonomických charakteristik. Cílem bylo zjistit, jaká skupina lidí by volila daného kandidáta a jaké šance mají jednotliví kandidáti.

Pro jednoduchost jsem se rozhodl, že budu pří průzkumu zohledňovat pouze 4 kandidáty na prezidenta, tedy budu uvažovat pouze 4 varianty výstupu. Důvodem byla lepší interpretovatelnost výsledků a rovněž to, že v době, kdy jsem formuloval svůj dotazník, viz příloha č.2, byli potvrzení kandidáti pouze 3 a jednoho jsem na základě rozsáhlých diskuzí přidal sám.

Respondenti měli na výběr z těchto 4 kandidátů na prezidenta ČR:

- Jan Fišer (A)
- Karel Schwarzenberg (B)
- Jana Bobošíková (C)
- Jan Švejnar (D)

9.2 Basketbal

Data pro svůj praktický příklad jsem získal z výsledků zdravotních sportovních testů profesionálních basketbalistů hrajících v nejvyšší české basketbalové soutěži, Mattoni Národní basketbalová lize. Ty jsem následně zanesl do tabulky, viz příloha č.1.

U sportovců byly sledovány nejrůznější charakteristiky, které jsem považoval za důležité při rozhodování o tom, na jaký post se hráč nejvíce hodí. Konkrétně to byly tyto charakteristiky:

- **věk**,
- **výška**,
- **hmotnost**,
- **BMI** - Body Mass Index (Index tělesné hmotnosti), určuje stupeň obezity a je počítán pomocí vzorce hmotnost v kg / [výška v m]^2; normální váhu určují hodnoty BMI mezi 18,5 a 25 kg/m^2,
- **% tuku** - procento tělesného tuku; průměrné hodnoty u basketbalistů by se měly pohybovat mezi 7-13%,
- **VO2 max** - aerobní kapacita; je to hraniční hodnota dosažení spotřeby kyslíku (maximální množství kyslíku za minutu, které může organismus využít při intenzívním fyzickém zatížení); jedná se o ukazatel tělesné zdatnosti; ideální hodnota by se měla pohybovat kolem 60 ml/kg,
- **VK(l)** - vitální kapacita plic; je závislá na fyzické zdatnosti člověka; měří se pomocí spirometru, do kterého vydechneme co největší množství vzduchu po maximálním nádechu,
- **TF klid** - tepová frekvence za minutu v klidovém stavu,
- **TF max** - maximální tepová frekvence za minutu,

- **ANP** - anaerobní práh (odhad z VO2 max převedený na TF); je to horní hranice, na které je ještě organismus schopen udržovat stabilní hladinu zakyselení a laktátu recyklací a využívat tuk jako palivový zdroj; nad hranicí anaerobního prahu je energetickým zdrojem pouze glukóza a to energeticky velmi nevýhodným způsobem,

- **RQ** - respirační kvocient; je to poměr vydýchaného oxidu uhličitého a přijatého kyslíku; protože různé živiny potřebují na svou oxidaci odlišné množství kyslíku a z jejich spalování vzniká také odlišné množství oxidu uhličitého, dá se s pomocí RQ určit momentální podíl bílkovin, cukrů a tuků na tělesné produkci energie; měl by být větší než 1,0.

Všechny výše jmenované sledované velčiny jsou spojité. Jejich základní sumární charakteristiky jsou uvedeny v tabulkách 10 a 11.

Promenna	cetnost	celkove	min	prumer	median	max	smer_odchylka
ANP	45	6970,3	136,1	154,8956	154,8	180,8	1,630182352
BMI	48	1176,35	21,02	24,50729	24,25	31,3	0,30518259
RQ	45	50,72	1,01	1,127111	1,13	1,23	0,008309798
TF_klid	48	2402	36	50,04167	49	69	1,068138806
TF_max	45	8236	168	183,0222	182	201	1,296260702
VK(l)	45	283,25	4,52	6,294444	6,18	8,69	0,130724692
VO2_max	45	2522,422	42,8	56,05383	57,3	64,2	0,783566619
hmotnost	48	4568,5	78	95,17708	97	121,5	1,607399886
proc_tuku	48	524,3	4,1	10,92292	11	20	0,516313992
vek	48	1226,54	16,8	25,55292	25,55	35,6	0,676185324
vyska	48	9455	177	196,9792	198	216	1,30956979

Tabulka 10: Sumární charakteristiky sledovaných proměnných

promenna	hodnota	pocet	procenta
post	K	23	47,91667
post	P	16	33,33333
post	R	9	18,75

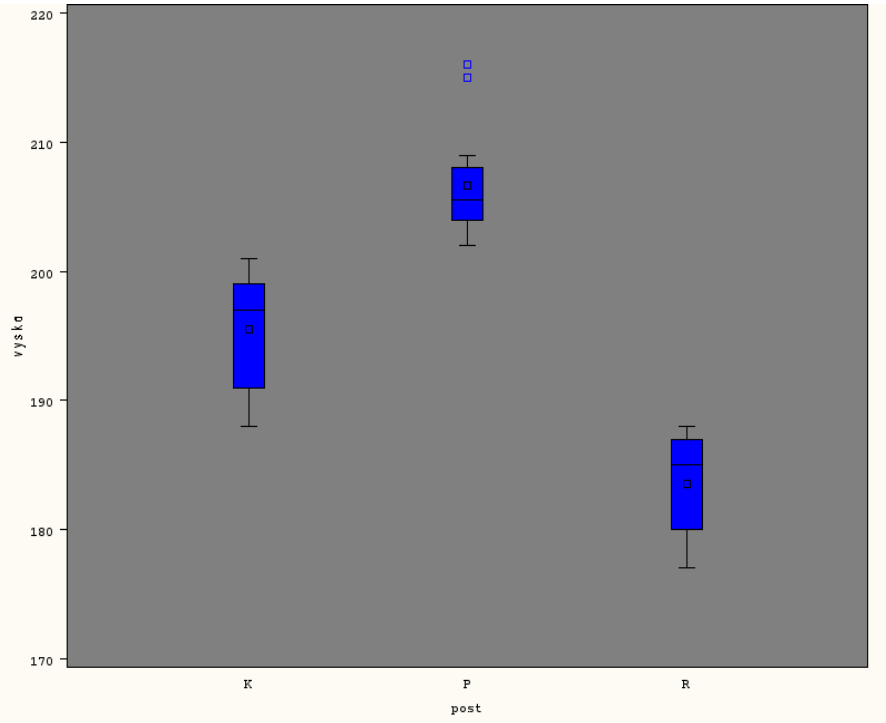
Tabulka 11: Četnosti jednotlivých postů

Zdálo se mi však, že vysvětlujících proměnných je příliš mnoho a tak jsem se snažil jejich počet zredukovat. Proto jsem si zjistil rozložení jednotlivých proměnných a vykreslil si nějakou grafickou podobu potenciální závislosti 2 proměnných, kterou bych mohl využít pro redukci.

U všech vstupních proměnných lze zjistit, pomocí úlohy Distribution Analysis, jejich důležité kvantily (dolní kvantil, medián, horní kvantil) a následně vykreslit srovnávací boxploty pro jednotlivé posty, které nám ukazují jejich rozložení. Pro názornost uvedu jen některé z nich.

Post	25% Q1	Medián	75% Q3
Křídlo	191	197	199
Pivot	204	205,5	208
Rozehrávač	180	185	187

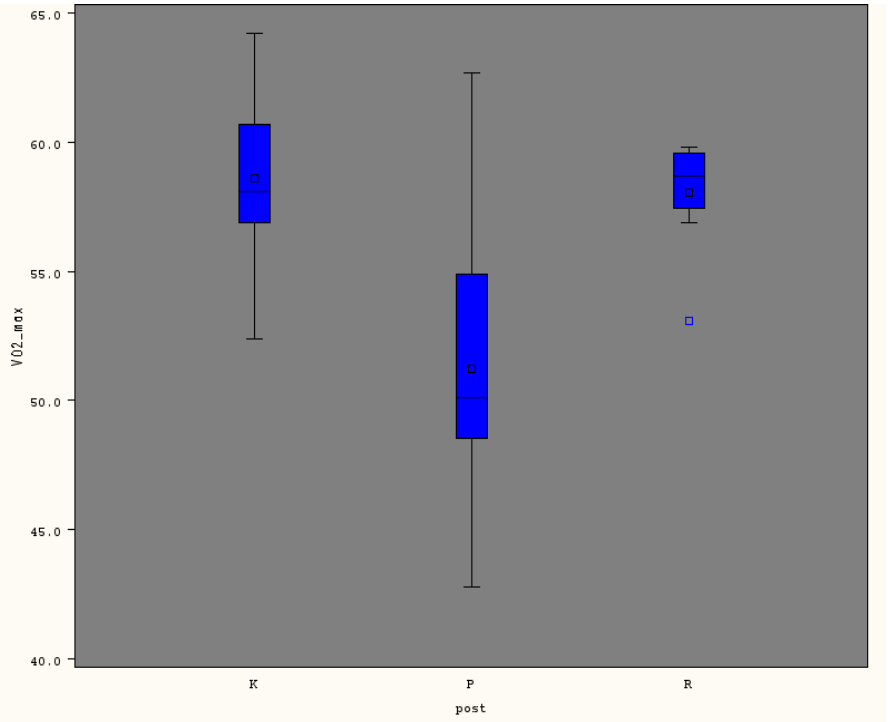
Tabulka 12: Důležité kvantily u parametru výška



Obrázek 8: Boxplot vstupní proměnné výška

Post	25% Q1	Medián	75% Q3
Křídlo	56,9	58,1	60,7
Pivot	48,5	50,1	54,9
Rozehrávač	57,5	58,7	59,6

Tabulka 13: Důležité kvantily u parametru VO2 max



Obrázek 9: Boxplot vstupní proměnné VO2 max

Dále můžeme také testovat závislost mezi jednotlivými sledovanými charakteristikami sportovců. Pokud bychom nějakou závislost mezi dvěmi charakteristikami odhadli, je zřejmě zbytečné uvažovat v modelu obě, ale stačí uvažovat pouze jednu.

Nejprve jsem se nad příkladem zamyslel a snažil jsem se sám určit dvojice charakteristik, které by mohli mít mezi sebou nějakou závislost. Mohlo by se jednat o BMI v kombinaci s tělesnou výškou, tělesnou hmotností, příp. % tuku. Tyto 2 parametry se používají pro výpočet BMI. Spolu mohou souviset rovněž výška a VK(l), protože větší hráči by měli mít pravděpodobně i větší plíce. Dále VO2 max a hmotnost, protože čím je hráč těžší, tím je méně pohyblivější a proto jeho kondiční připravenost není na tak vysoké úrovni.

Pomocí úlohy Scatter Plot Matrix jsem vykreslil maticový graf, ze kterého lze na první pohled poznat, mezi kterými dvojicemi by mohla existovat nějaká závislost. Pokud má graf vzhled odpovídající nahodilému rozložení bodů, pak závislost mezi proměnnými pravděpodobně neexistuje. Pokud jsou však body uspořádány v nějakém pravidelném shluku, např. vzdáleně připomínají přímku,

můžeme říci, že u těchto dvou proměnných pravděpodobně existuje závislost.

Obrázek 10: Maticový graf pro sledované proměnné

Z obrázku 10 můžeme vyčíst, že pravděpodobně existuje nějaká závislost mezi dvojicemi výška a hmotnost, výška a VK(l), hmotnost a BMI, BMI a % tuku, hmotnost a VO2 max, TF max a ANP.

Pokud tedy pravděpodobně existuje závislost mezi dvojicí proměnných, můžeme uvažovat o tom, že jednu z nich z modelu vyloučíme. Proto jsem redukoval počet vstupních proměnných a v modelu jsem nechal pouze tyto proměnné: věk, výška, VO2 max, TF klid a ANP.

Následně jsem v programu SAS EG prostřednictvím úlohy Logistic Regression vyvolal proceduru LOGISTIC. Na výstupu se ale objevilo varování, že maximálně věrohodné odhady nemusí existovat. Důvodem tohoto varování byla detekce tzv. kvazi-kompletní separace dat, o které jsem pojednal v kapitole 5.4.

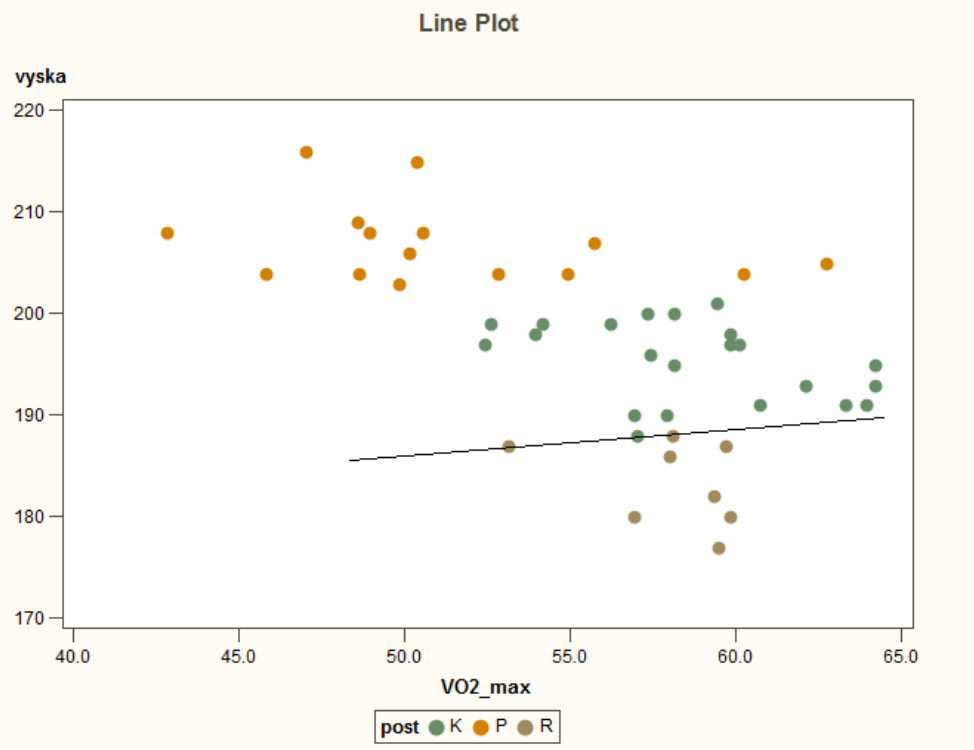
Model Convergence Status
Quasi-complete separation of data points detected.

Warning: The maximum likelihood estimate may not exist.

Obrázek 11: Detekce kvazi-kompletně separovaných dat na vstupu

Procedura LOGISTIC pokračuje ve výpočtech navzdory varování. Výsledky jsou však založeny na poslední iteraci výpočtu před jeho zastavením. Platnost modelu je diskutabilní, proto ani výstup procedury nebudu zobrazovat.

Protože by bylo složité vykreslit graf, kde jsou zohledněny všechny proměnné, vykreslil jsem, pomocí úlohy Line Plot, pro ilustraci pouze 2D graf závislosti výšky na VO2 max. V něm je dobře partná tzv. kvazi-kompletní separace vstupních dat, viz obrázek 12.



Obrázek 12: Kvazi-kompletně separovaná data výška a VO2 max

Tento problém dokáže částečně vyřešit Firthova penalizační metoda, viz kapitola 5.3.3. Tato metoda ale pomůže najít maximálně věrohodné odhady parametrů pouze u standardní logistické regrese, tedy u regrese s binární výstupní

proměnnou. V našem příkladě pouze tehdy, pokud budeme rozhodovat, zda se hráč hodí na konkrétní pozici (např. post pivota).

Dále se proto budu věnovat už jen alternativní úloze standardní logisitcké regrese, tj. situaci, kdy trenér bude hledat hráče na konkrétní post, např. na post pivota. Úlohu popíšu právě pro tento post. Pro ostatní dva posty by byl postup tentýž.

9.2.1 Výběr hráče na konkrétní post

Závislá (výstupní) proměnná post pivota má binární charakter - hodnota P znamená, že se hodí na post pivota, hodnota 0 znamená, že se nehodí na post pivota. Vysvětlující proměnné mají spojitý charakter. Budeme tedy uvažovat základní model bez interakce v následujícím tvaru:

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5)}}$$

$$logitP(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

α ..... absolutní člen (intercept)

β_1 ..... neznámý regresní parametr odpovídající výšce

β_2 ..... neznámý regresní parametr odpovídající věku

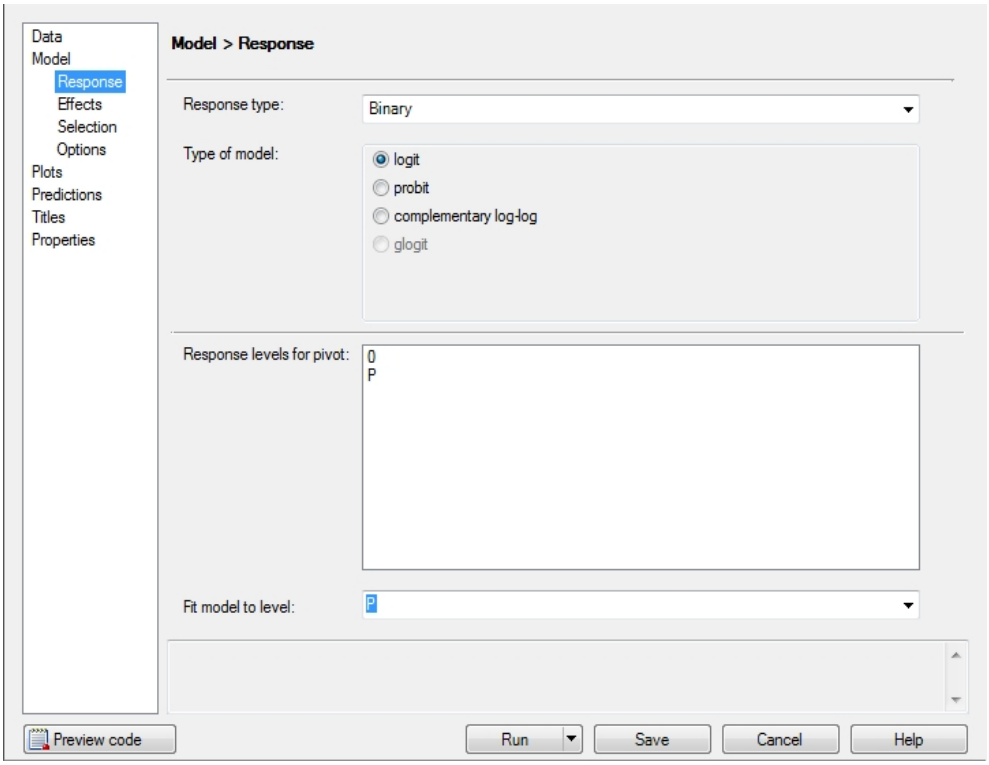
β_3 ..... neznámý regresní parametr odpovídající proměnné VO2 max

β_4 ..... neznámý regresní parametr odpovídající proměnné TF klid

β_5 ..... neznámý regresní parametr odpovídající proměnné ANP

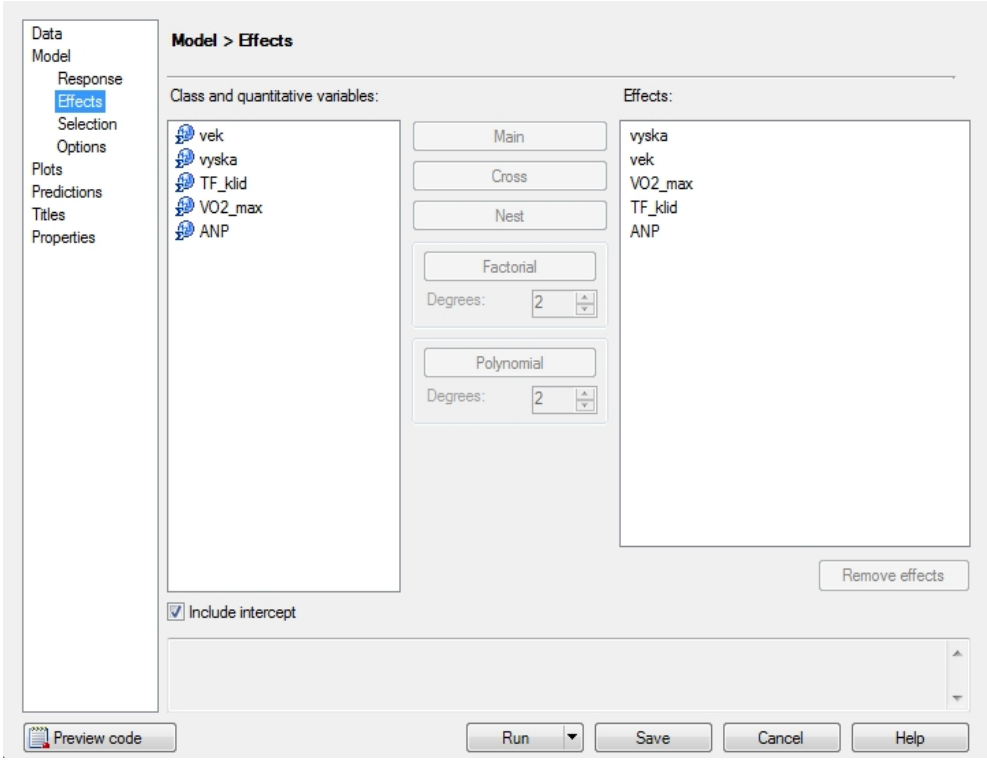
V datech pak vytvoříme, pomocí nástroje Query Builder, novou proměnnou nabývající hodnoty P pro pivoty a 0 pro hráče na ostatních postech.

Odhady parametrů provedeme pomocí metody maximální věrohodnosti. Při využití programu SAS a procedury LOGISTIC je výpočet jednoduchý. Zvolíme typ výstupní proměnné, typ modelu, reps. typ linkové funkce (z důvodu interpretovatelnosti zvolíme implicitní logit). Zvolíme referenční kategorii jako P (pivot), viz obrázek 13.



Obrázek 13: Zadání typu výstupní proměnné a typu modelu v SASu

Poté musíme zadat všechny efekty - vysvětlující proměnné, které mají být v modelu zahrnuty, viz obrázek 14. A v podokně Options zaškrtneme u možnosti Model Fitting methods, z důvodu kvazi-kompletní separace vstupních dat, Firthovu penalizační metodu.



Obrázek 14: Zadání vstupních proměnných v SASu

Výstup v programu SAS EG se skládá z mnoha tabulek. V prvních části výstupu jsou zobrazeny tři tabulky udávající informace o modelu, počet pozorování a profil výstupní proměnné, viz obrázek 15.

Model Information	
Data Set	WORK.SORTTEMPTABLESORTED
Response Variable	pivot
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Likelihood Penalty	Firth's bias correction

Number of Observations Read	48
Number of Observations Used	45

Response Profile		
Ordered Value	pivot	Total Frequency
1	0	30
2	P	15

Probability modeled is pivot='P'.

Obrázek 15: První část výstupu v SASu

Zde můžeme vyčíst, že byla použita zmiňovaná Firthova penalizační metoda a že ze 48 původních pozorování bylo použito pouze 45, z důvodů chybějících údajů u 3 pozorování. To je způsobeno tím, že 3 hráči neabsolvovali ze zdravotních důvodů celé sportovní vyšetření a tedy hodnoty některých sledovaných charakteristik u nich chybí.

V druhé části výstupu jsou uvedeny údaje, zda je splněno konvergenční kritérium, jako výchozí je nastaveno konvergenční kritérium GCONV s hodnotou 1E-8. To ale můžeme změnit, viz kapitola 8.1.2. Dále je zde zobrazeno vyhodnocení modelu a následné testování globální nulové hypotézy, že celý vektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)'$ je roven nule, viz obázek 16.

Intercept-Only Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	27.205	6.839
SC	29.012	17.679
-2 Log L	25.205	-5.161

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	30.3668	5	<.0001
Score	30.0356	5	<.0001
Wald	11.4989	5	0.0423

Obrázek 16: Druhá část výstupu v SASu

Zde můžeme vidět, že konvergenční kritérium GCONV je splněno. Ve druhé tabulce jsou zobrazeny 3 kritéria vhodnosti modelu. Model, který ma nižší hodnotu AIC a SC je lepší. V našem případě je lepší plný model (se všemi kategoriemi). Ve třetí tabulce je testována globální nulová hypotéza. Ve sloupci Chi-Square je uvedena hodnota χ^2 testové statistiky, ve sloupci DF je uveden počet

stupňů volnosti a v posledním sloupci Pr > ChiSq je uvedena P-hodnota pro daný test.

Je patrné, že pro všechny tři testy zamítáme nulovou hypotézu ve prospěch alternativy pro hladinu testu 0,05, tedy vektor $\boldsymbol{\beta}$ je statisticky významný.

Ve třetí části výstupu jsou uvedeny maximálně věrohodné odhady parametrů, včetně Waldových testů parametrů, a také jejich intervaly spolehlivosti, viz obrázek 17.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.5855	21.4740	0.5268	0.4680
vyska	1	0.1327	0.0668	3.9433	0.0471
vek	1	-0.0801	0.1068	0.5632	0.4530
VO2_max	1	-0.1682	0.1153	2.1279	0.1446
TF_klid	1	-0.0302	0.0618	0.2384	0.6253
ANP	1	0.0117	0.0456	0.0655	0.7980

Parameter Estimates and Wald Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	-15.5855	-57.6738	26.5029
vyska	0.1327	0.00172	0.2636
vek	-0.0801	-0.2894	0.1291
VO2_max	-0.1682	-0.3942	0.0578
TF_klid	-0.0302	-0.1513	0.0910
ANP	0.0117	-0.0776	0.1010

Obrázek 17: Třetí část výstupu v SASu

Nejdůležitější je poslední sloupec v první tabulce, kde je zobrazena P-hodnota Waldova testu. Je-li hodnota větší než hladina testu (0,05), pak nulovou hypotézu $H_0 : \beta_i = 0$, $i = 1, 2, 3, 4, 5$ nelze zamítnout. Je-li naopak P-hodnota menší než hladina testu, pak nulovou hypotézu zamítáme ve prospěch alternativy $H_A : \beta_i \neq 0$, $i = 1, \ldots, 5$.

Z výsledků vyplývá, že parametry $\alpha, \beta_2, \beta_3, \beta_4, \beta_5$ jsou statisticky nevýznamné a nemůžeme zamítnout, že mohou nabývat hodnoty 0. Oproti tomu pro parametr β_1 (parametr výšky) zamítáme H_0 ve prospěch alternativy, což znamená, že jediný parametr β_1 je statisticky významný.

To je patrné i z intervalů spolehlivosti u jednotlivých parametrů. Pouze inter-

val speciálně pro parametr β_1 neobsahuje 0. Všechny ostatní parametry 0 obsahují.

Tento výsledek se dá interpretovat i tak, že zkušenější trenéři se rozhodují o hráči na post pivota ze všech našich sledovaných charakteristik pravděpodobně pouze podle tělesné výšky.

Ve čtvrté části výstupu jsou zobrazeny odhady poměrů šancí OR a jejich 95 % intervaly spolehlivosti, viz tabulka 14.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
vyska	1.142	1.002	1.302
vek	0.923	0.749	1.138
VO2_max	0.845	0.674	1.059
TF_klid	0.970	0.860	1.095
ANP	1.012	0.925	1.106

Tabulka 14: Čtvrtá část výstupu v SASu

Hodnota parametru $OR_1 = e^{\hat{\beta}_1} = 1,142$. To znamená, že pokud bude hráč o 1 cm vyšší, zvětší se šance jeho zařazení na post pivota, a to 1,142-krát (o 14,2 %). To je jednoduše odůvodnitelné tím, že větší hráči se lépe prosazují pod košem a lépe doskakují, a proto je trenéři na tento post dosazují.

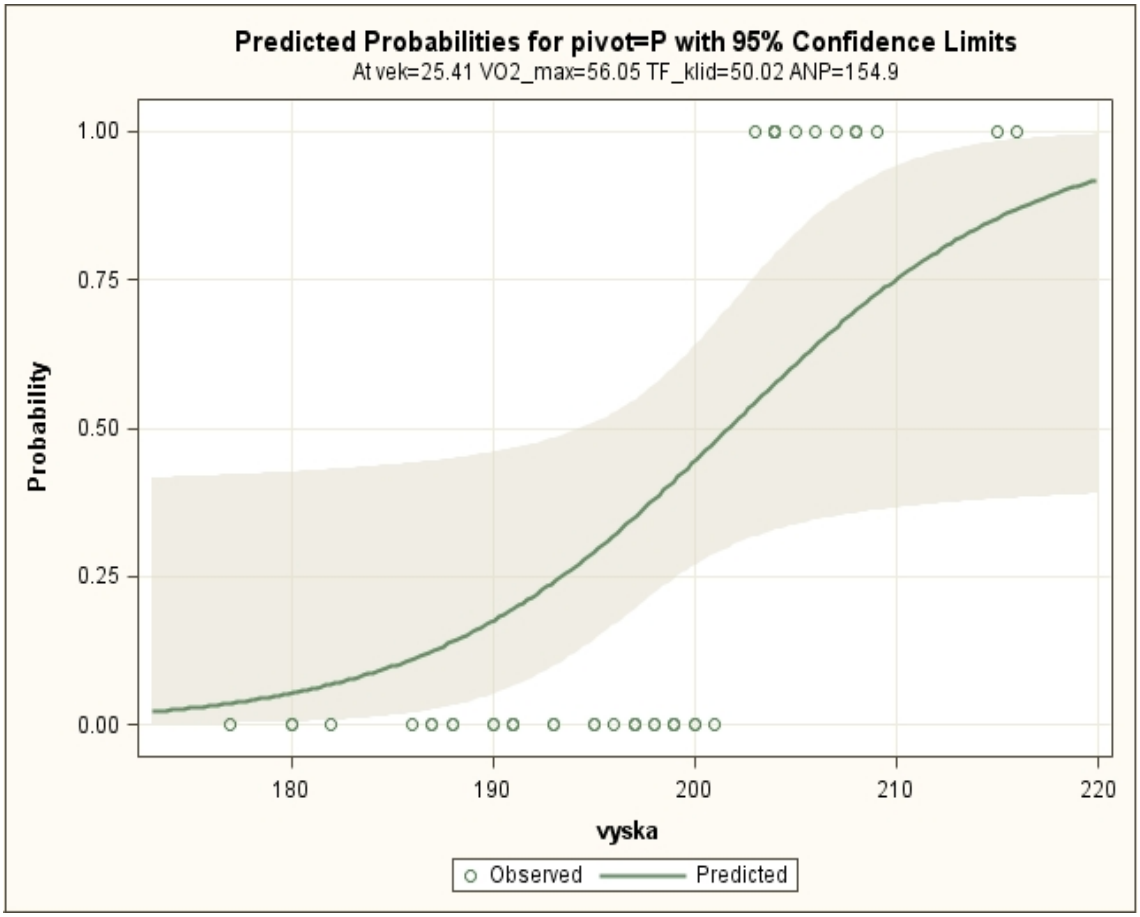
Naopak hodnota parametru $OR_3 = e^{\hat{\beta}_3} = 0,845$. To znamená, že pokud se naměřená hodnota VO2 max u nového hráče zvýší o jednotku, šance jeho zařazení na post pivota klesne, a to 0,845-krát (o 15,5 %). Tuto závislost můžeme odůvodnit tím, že trenér potřebuje více kondičně připravené hráče na jiných pozicích, než na pozici pivota.

Takto podobně by se daly interpretovat i ostatní poměry šancí.

U konfidenčních intervalů můžeme vidět i souvislost s odhady parametrů. Pokud jsme pro parametry v předchozí části výstupu nezamítli tvrzení nulové hypotézy, tj. $H_0 : \beta_i = 0$, $i = 1, 2, 3, 4, 5$, pak konfidenční interval poměru šancí pro proměnné odpovídající těmto parametrům obsahuje hodnoty 1, tedy změna hodnoty proměnné nemá vliv na změnu šance zařazení hráče na post pivota.

Pouze u parametru β_1 (výška) konfidenční interval OR nepokrývá hodnotu 1 a lze o něm říci, že tedy má vliv na změnu šance zařazení hráče na post pivota.

Poslední částí výstupu jsou grafická znázornění jednotlivých závislostí. Vybral jsem znázornění závislosti odhadované pravděpodobnosti zařazení hráče na post pivota na jeho výšce, viz obrázek 18.

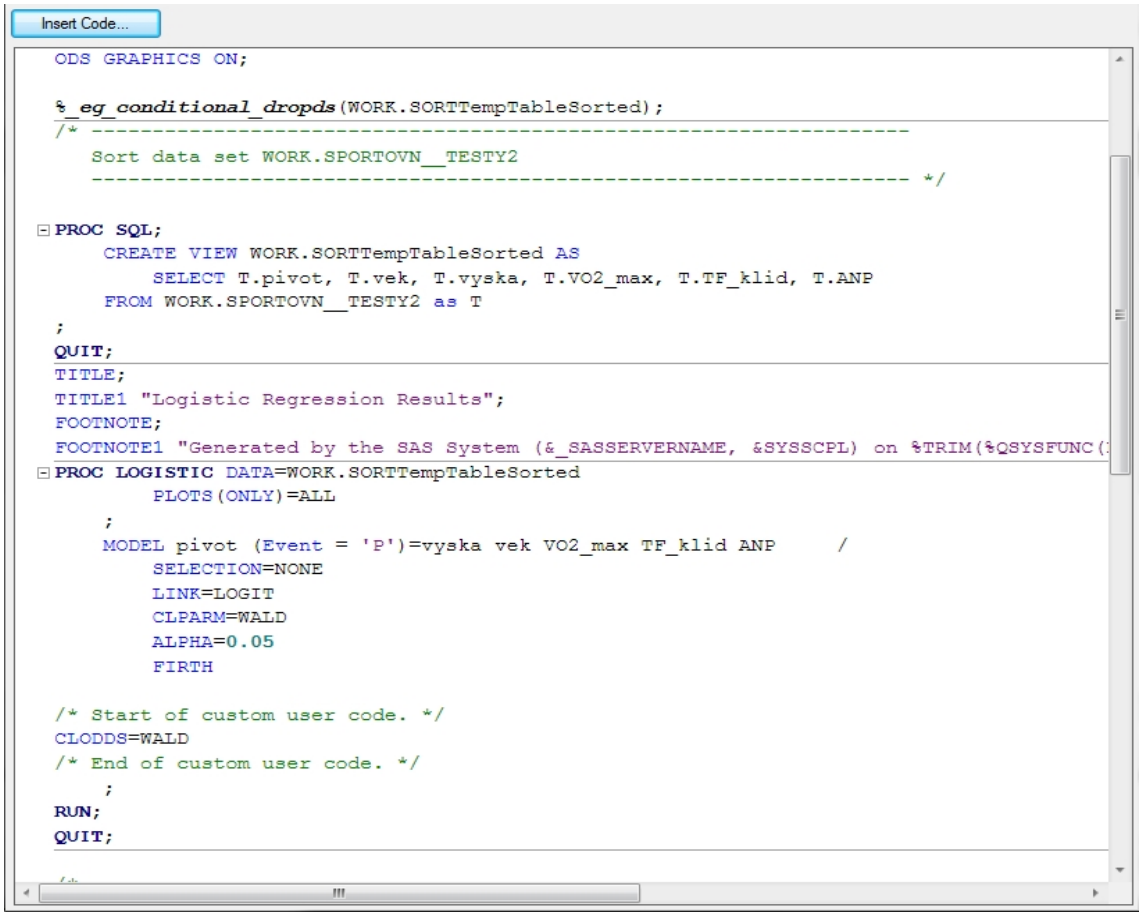


Obrázek 18: Graf závislosti pravděpodobnosti zařazení hráče na post pivota na jeho výšce

Chtěl jsem vykreslit i grafické znázornění poměrů šancí OR. Tuto možnost bohužel úloha Logistic Regression nenabízí a proto jsem ji musel sám připsat do zdrojového kódu procedury LOGISTIC.

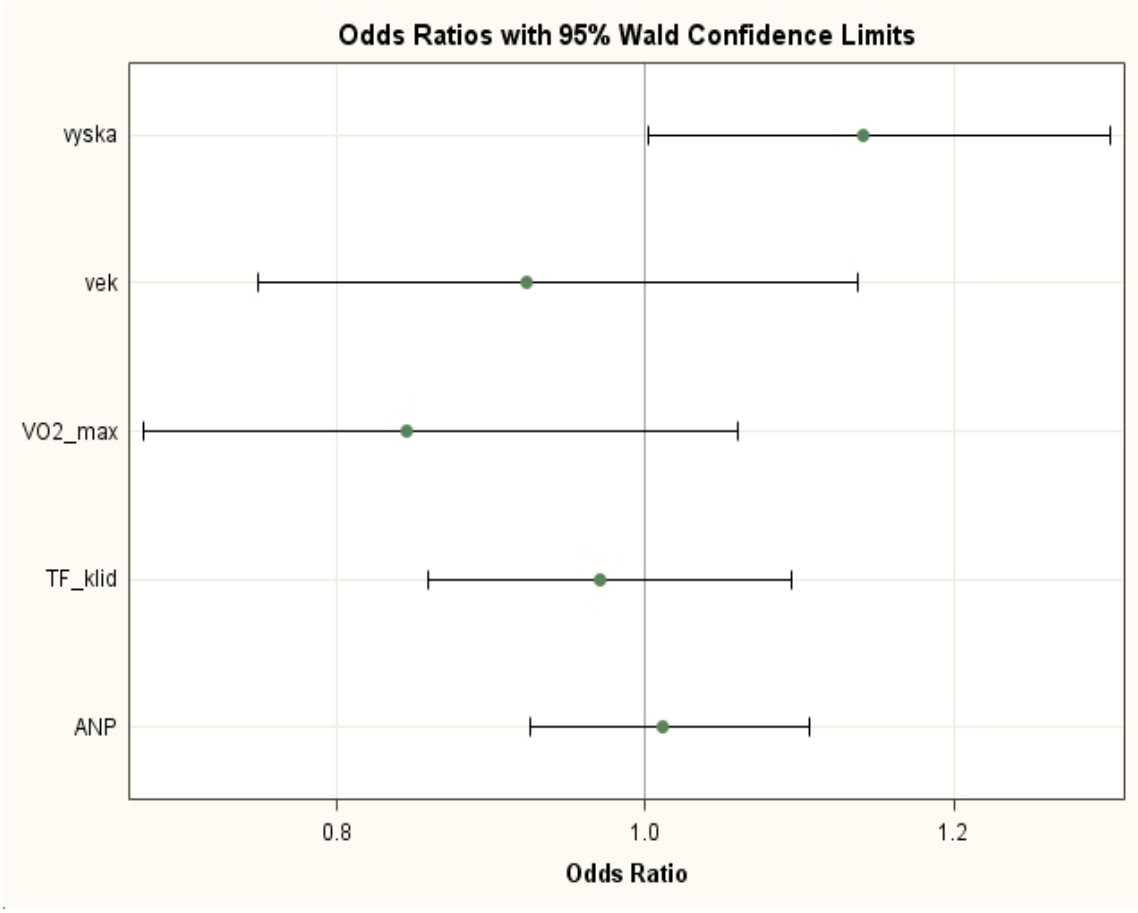
Do zdrojového kódu se dostaneme pomocí volby Preview code, při zádávání parametrů procedury, viz obrázek 11. Objeví se dialogové okno s kódem celém procedury tak, jak vypadá naprogramovaná v SAS jazyce. Poté stisknutím volby

Insert Code... program nabídne místa, kde je možné vložit vlastní příkaz či volbu, aniž bychom porušili celkový proces, viz obrázek 19.



Obrázek 19: Zdrojový kód procedury LOGISTIC

Na obrázku je zobrazen celý zdrojový kód procedury LOGISTIC s volbami, které jsem si navolil prostřednictvím SAS EG. Požadovaný graf poměrů šancí i s 95 % intervaly spolehlivosti jsem vykreslil volbou CLOODS=WALD v příkazu MODEL, viz obrázek 20.



Obrázek 20: Graf 95 % intevalů spolehlivosti OR

9.3 Kandidáti

Data pro svůj druhý praktický příklad jsem získal prostřednisctvím vlastního průzkumu formou dotazníku, viz příloha č. 2.

Dotazované osoby jsem žádal o vyplnění různých socio-ekonomických charakteristik, které jsem považoval za důležité při rozhodování o volbě prezidentského kandidáta. Byly to tyto charakteristiky:

- pohlaví,
- věk,
- dosažené vzdělání,
- pracovní poměr,

- měsíční příjem (hrubý),
- účast na posledních senátních volbách,
- případná účast v příme volbě prezidenta,
- politické preference.

Z důvodů malé četnosti některých kategorií odpovědí u sledovaných charakteristik, jsem se rozhodl některé kategorie sloučit. Pomocí volby Query Builder jsem vytvořil nové kategorie. Např. ve vysvětlující proměnné *pracovní poměr* jsem sloučil kategorie lidí, kteří jsou zaměstnáni či zaměstnávají v soukromém sektoru do kategorie *soukromý sektor*, podobně jsem sloučil i kategirie *nezaměstnaný*, *student* a *mateřská dovolená* do společné kategorie *sociální dávky*. Data jsou zobrazena v tabulce, viz příloha č. 3.

Protože ani jeden z respondentů by si nezvolil za prezidenta ČR Janu Bobošíkovou (kategorie C), můžeme zjednodušit náš model multinomické logistické regrese na model se 3 kategoriemi výstupní proměnné.

Všechny výše jmenované sledované veličiny, s výjimkou věku, jsou kategoriální (klasifikační). Veličina věk je spojitá. Četnostní tabulky sledovaných veličin jsou uvedeny v tabulkách níže.

Table of Pohlavi by Kandidat

		Kandidat			
		A	B	D	Total
Pohlavi					
M	Frequency	24	6	16	46
Z	Frequency	29	9	11	49
Total	Frequency	53	15	27	95

Table of PraceSektory by Kandidat

		Kandidat			
		A	B	D	Total
PraceSektory					
OSVC	Frequency	6	5	4	15
socialni	Frequency	14	3	5	22
soukromy	Frequency	21	2	11	34
verejny	Frequency	12	5	7	24
Total	Frequency	53	15	27	95

Obrázek 21: Četnostní tabulky část 1

Table of Prijem by Kandidat					
		Kandidat			
		A	B	D	Total
Prijem					
0-10	Frequency	11	3	5	19
10-20	Frequency	17	6	12	35
20-40	Frequency	17	4	8	29
40-60	Frequency	3	1	1	5
60+	Frequency	5	1	1	7
Total	Frequency	53	15	27	95

Table of VladniStrany by Kandidat					
		Kandidat			
		A	B	D	Total
VladniStrany					
jina	Frequency	10	3	1	14
opozice	Frequency	13	1	6	20
vlada	Frequency	28	10	17	55
Total	Frequency	51	14	24	89
Frequency Missing = 6					

Obrázek 22: Čestnostní tabulky část 2

Table of Vzdelani by Kandidat					
		Kandidat			
		A	B	D	Total
Vzdelani					
S	Frequency	24	8	12	44
V	Frequency	27	5	13	45
Z	Frequency	2	2	2	6
Total	Frequency	53	15	27	95

Table of Hlasoval by Kandidat					
		Kandidat			
		A	B	D	Total
Hlasoval					
0	Frequency	5	2	5	12
1	Frequency	48	13	22	83
Total	Frequency	53	15	27	95

Table of Pr_volba by Kandidat					
		Kandidat			
		A	B	D	Total
Pr_volba					
0	Frequency	4	3	3	10
1	Frequency	49	12	24	85
Total	Frequency	53	15	27	95

Obrázek 23: Čestnostní tabulky část 3

9.3.1 Model logistické regrese

Po zjednodušení máme model multinomické logistické regrese se 3 výstupními proměnnými (kategoriemi). Jako referenční kategorii jsem si zvolil Jana Fišera (kategorii A).

V multinomické logistické regresi se třemi výstupními kategoriemi musíme použít dvě logitové transformace modelu

$$\log\left[\frac{P(\text{kandidát B}|\mathbf{X})}{P(\text{kandidát A}|\mathbf{X})}\right] = \alpha_1 + \sum_{i=1}^{6} \beta_{1i} X_i,$$

$$\log\left[\frac{P(\text{kandidát D}|\mathbf{X})}{P(\text{kandidát A}|\mathbf{X})}\right] = \alpha_2 + \sum_{i=1}^{6} \beta_{2i} X_i,$$

kde i, $i = 1, \ldots, 6$, označuje počet vysvětlujících proměnných a g počet logitových transformací modelu (vždy o 1 menší, než je počet kategorií), $g = 1, 2$,

α_g ..... absolutní členy (intercepty),

β_{g1} ..... neznámý regresní parametr odpovídající věku,

β_{g2} ..... neznámý regresní parametr odpovídající vzdělání,

β_{g3} ..... neznámý regresní parametr odpovídající pohlaví,

β_{g4} ..... neznámý regresní parametr odpovídající příjmu,

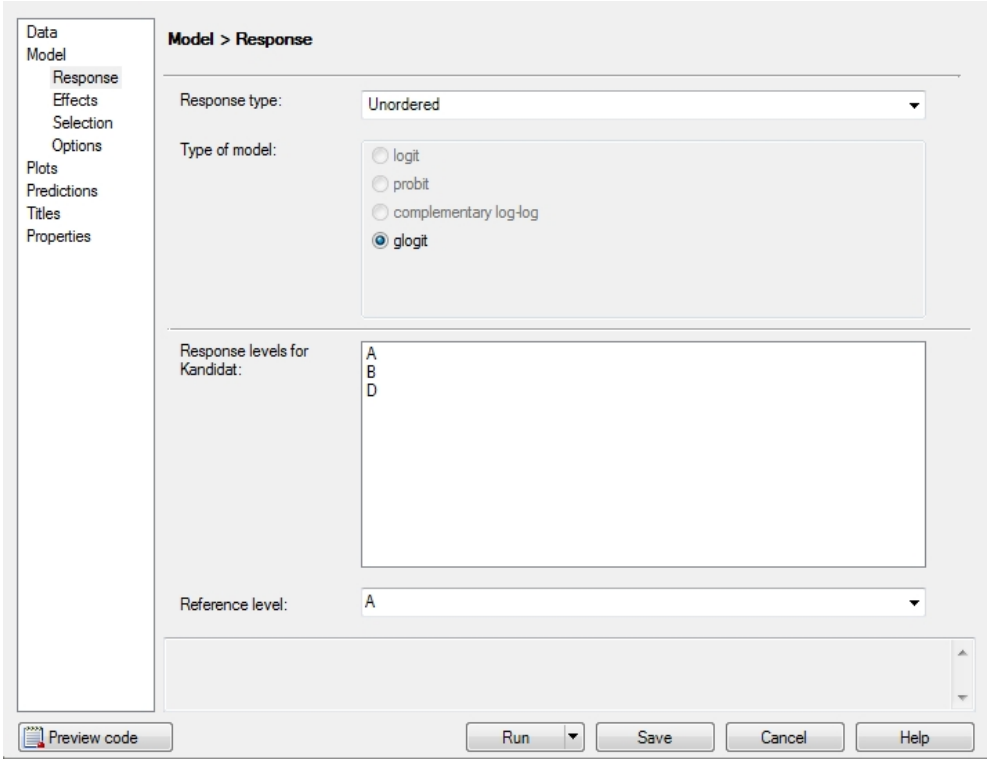
β_{g5} ..... neznámý regresní parametr odpovídající pracovnímu poměru,

β_{g6} ..... neznámý regresní parametr odpovídající politické preferenci.

Z modelu jsem záměrně vyloučil vysvětlující proměnné, zda-li respondent hlasoval v posledních senátních volbách a zda-li bude hlasovat v přímé volbě prezidenta, to z toho důvodu, že téměř všichni dotazovaní hlasovali v minulých volbách a půjdou hlasovat do přímé volby prezidenta, tyto proměnná nemají žádný význam.

Obdobně jako u předchozího praktického příkladu s basketbalisty vypočteme odhady parametrů metodou maximální věrohodnosti. Zadání úlohy Logistic Regression se liší pouze v prvním dialogovém okně, kde musíme nastavit typ linkové

funkce jako glogit a typ výstupní proměnné jako nominální, přirozeně neseřazené (unsorted), viz obrázek 24.



Obrázek 24: Zadání typu výstupní proměnné a typu modelu v SASu

Poté zadáme všechny efekty - vysvětlující proměnné, které mají být v modelu zahrnuty, stejně jako v předchozím příkladě.

Výstup lze opět rozdělit na několik částí. V první části jsou tři tabulky udávající informace o modelu, viz obrázek 25.

Model Information	
Data Set	WORK.SORTTEMPTABLESORTED
Response Variable	Kandidat
Number of Response Levels	3
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	95
Number of Observations Used	89

Response Profile		
Ordered Value	Kandidat	Total Frequency
1	A	51
2	B	14
3	D	24

Logits modeled use Kandidat='A' as the reference category.

Obrázek 25: První část výstupu v SASu

Zde můžeme vyčíst, že byla použita Newton-Raphsonova metoda a že bylo z původních 95 pozorování použito pouze 89. To je způsobeno tím, že kolonka politické preference byla nepovinná a 6 respondentů využilo této možnosti a neudalo své politické preference.

V druhé části výstupu se nachází informace o parametrizaci vysvětlujících proměnných. Při zadávání dat jsem zvolil, aby vysvětlující kategoriální proměnné byly kódovány jako efekt, viz tabulka 15.

Class Level Information					
Class	Value	Design Variables			
Vzdelani	S	1	0		
	V	0	1		
	Z	-1	-1		
Pohlavi	M	1			
	Z	-1			
Prijem	0-10	1	0	0	0
	10-20	0	1	0	0
	20-40	0	0	1	0
	40-60	0	0	0	1
	60+	-1	-1	-1	-1
PraceSektory	OSVC	1	0	0	
	socialni	0	1	0	
	soukromy	0	0	1	
	verejny	-1	-1	-1	
VladniStrany	jina	1	0		
	opozice	0	1		
	vlada	-1	-1		

Tabulka 15: Parametrizace vysvětlujících kategoriálních proměnných

Ve třetí části jsou uvedeny údaje, zda je splňeno konvergenční kritérium. Dále je zde zobrazeno vyhodnocení modelu a následné testování globální nulové hypotézy, že celý vektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)'$ je roven nule, viz obrázek 26.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	175.491	200.480
SC	180.468	270.161
-2 Log L	171.491	144.480

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	27.0113	26	0.4087
Score	24.3704	26	0.5548
Wald	19.5558	26	0.8121

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Vek	2	1.6035	0.4485
Vzdelani	4	2.4362	0.6561
Pohlavi	2	2.1619	0.3393
Prijem	8	5.3205	0.7228
PraceSektory	6	8.3526	0.2134
VladniStrany	4	6.4989	0.1649

Obrázek 26: Třetí část výstupu v SASu

Zde můžeme vidět, že konvergenční kritérium GCONV bylo opět splněno. Tabulka testování globální nulové hypotézy nám říká, že vektor $\boldsymbol{\beta}$ je statisticky nevýznamný, tedy, že nelze zamítnout nulovou hypotézu $H_0 : \boldsymbol{\beta} = 0$ na hladině významnosti 0,05. Jinými slovy lze říci, že ani jedna z vysvětlujících proměnných nemá u respondentů vliv na volbu některého z kandidátů. To můžeme vidět i v poslední tabulce Type 3 Analysis of Effects (viz sloupec Pr > ChiSq).

Ve čtvrté části výstupu máme maximálně věrohodné odhady parametrů, včetně výseldků Waldových testů, viz tabulka 16.

Analysis of Maximum Likelihood Estimates							
Parameter		Kandidat	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		B	1	0.6389	1.9010	0.1130	0.7368
Intercept		D	1	-2.0285	1.2491	2.6374	0.1044
Vek		B	1	-0.0539	0.0452	1.4233	0.2329
Vek		D	1	0.00479	0.0249	0.0371	0.8473
Vzdelani	S	B	1	-0.3178	0.7304	0.1893	0.6635
Vzdelani	S	D	1	-0.1528	0.5137	0.0885	0.7661
Vzdelani	V	B	1	-1.0695	0.7736	1.9114	0.1668
Vzdelani	V	D	1	0.1460	0.5479	0.0710	0.7899
Pohlavi	M	B	1	-0.4802	0.4643	1.0698	0.3010
Pohlavi	M	D	1	0.2449	0.2975	0.6773	0.4105
Prijem	0-10	B	1	0.2332	1.4441	0.0261	0.8717
Prijem	0-10	D	1	1.4143	1.0129	1.9498	0.1626
Prijem	10-20	B	1	-0.5355	0.9012	0.3531	0.5524
Prijem	10-20	D	1	0.8047	0.6036	1.7771	0.1825
Prijem	20-40	B	1	-0.0428	0.7345	0.0034	0.9535
Prijem	20-40	D	1	-0.0941	0.5746	0.0268	0.8700
Prijem	40-60	B	1	0.8525	1.2179	0.4900	0.4839
Prijem	40-60	D	1	-0.6637	1.0237	0.4204	0.5167
PraceSektory	OSVC	B	1	1.3388	0.7495	3.1908	0.0741
PraceSektory	OSVC	D	1	0.2645	0.6702	0.1558	0.6931
PraceSektory	socialni	B	1	-1.3553	1.2992	1.0883	0.2969
PraceSektory	socialni	D	1	-1.1648	0.8603	1.8335	0.1757
PraceSektory	soukromy	B	1	-0.9498	0.7889	1.4496	0.2286
PraceSektory	soukromy	D	1	0.4384	0.4870	0.8103	0.3680
VladniStrany	jina	B	1	-0.0193	0.6945	0.0008	0.9779
VladniStrany	jina	D	1	-1.3274	0.7710	2.9641	0.0851
VladniStrany	opozice	B	1	-1.0351	0.7956	1.6928	0.1932
VladniStrany	opozice	D	1	0.3603	0.5549	0.4216	0.5162

Tabulka 16: Čtvrtá část výstupu v SASu

Stejně jako u předchozího příkladu i tady je nejdůležitější poslední sloupec, kde je zobrazena P-hodnota Waldova testu. Je-li hodnota větší než hladina významnosti testu (0,05), pak příslušnou nulovou hypotézu $H_0 : \beta_{gi} = 0$ nelze zamítnout. Je-li naopak P-hodnota menší než hladina testu, pak nulovou hypotézu zamítáme ve prospěch alternativy $H_A : \beta_{gi} \neq 0$.

Ani pro jeden parametr nemůžeme zamítnout původní tvrzení ve prospěch alternativy, tedy všechny parametry jsou statisticky nevýznamné a nemůžeme zamítnout, že mohou nabývat hodnoty 0.

Tento výsledek můžeme chápat i tak, že prezident by měl mít své voliče napříč

celým spektrem občanů České republiky, a tak nelze přesně určit skupinu občanů, kteří by volili daného kandidáta. Chyba může být ovšem i na mé straně, kdy jsem při tvorbě dotazníku vybral špatné kandidáty, a tak jsem zkreslil výsledky logistického modelu.

V poslední části výstupu jsou zobrazeny poměry šancí OR a jejich 95 % intervaly spolehlivosti, viz tabulka 17.

Odds Ratio Estimates				
Effect	Kandidat	Point Estimate	95% Wald Confidence Limits	
Vek	B	0.947	0.867	1.035
Vek	D	1.005	0.957	1.055
Vzdelani S vs Z	B	0.182	0.004	7.390
Vzdelani S vs Z	D	0.852	0.064	11.349
Vzdelani V vs Z	B	0.086	0.002	3.850
Vzdelani V vs Z	D	1.149	0.080	16.567
Pohlavi M vs Z	B	0.383	0.062	2.362
Pohlavi M vs Z	D	1.632	0.508	5.239
Prijem 0-10 vs 60+	B	2.097	0.022	199.884
Prijem 0-10 vs 60+	D	17.736	0.625	503.713
Prijem 10-20 vs 60+	B	0.972	0.031	30.334
Prijem 10-20 vs 60+	D	9.640	0.693	134.065
Prijem 20-40 vs 60+	B	1.591	0.075	33.884
Prijem 20-40 vs 60+	D	3.924	0.355	43.363
Prijem 40-60 vs 60+	B	3.895	0.096	157.447
Prijem 40-60 vs 60+	D	2.220	0.091	53.876
PraceSektory OSVC vs verejny	B	1.451	0.170	12.372
PraceSektory OSVC vs verejny	D	0.821	0.113	5.971
PraceSektory socialni vs verejny	B	0.098	0.003	3.646
PraceSektory socialni vs verejny	D	0.197	0.017	2.330
PraceSektory soukromy vs verejny	B	0.147	0.016	1.314
PraceSektory soukromy vs verejny	D	0.977	0.220	4.335
VladniStrany jina vs vlada	B	0.342	0.046	2.516
VladniStrany jina vs vlada	D	0.101	0.010	1.015
VladniStrany opozice vs vlada	B	0.124	0.011	1.353
VladniStrany opozice vs vlada	D	0.545	0.131	2.268

Tabulka 17: Pátá část výstupu v SASu

Interpretace parametrů OR je obdobná jako u příkladu s basketbalisty.

Závěr

Cílem této práce bylo seznámit čtenáře s logistickou regresní analýzou, zejména pak s multinomickou logistickou regresí. Ta jim následně může pomoci při studiu složitější literatury v cizím jazyce.

Praktické příklady ukazují konkrétní výpočty v programu SAS, respektive v jeho modulu SAS EG. Příklady vidím jako důležité prostředky pro názornost celé práce.

V prvním příkladu jsem aplikoval vztahy a vzorce logistické regrese v oblasti sportu. Konkrétně jsem chtěl pomoci začínajícímu trenérovi basketbalového týmu, který nemá zatím moc zkušeností, vybrat tu správnou pozici pro svého hráče na základě několika sledovaných charakteristik. Při výstupu se třemi kategoriemi (rozehrávač, křídlo a pivot) jsem ovšem narazil na problém separace vstupních dat. To mohlo být způsobeno malým počtem získaných dat. Proto jsem původní úlohu multinomické logistické regrese zjednodušil na úlohu klasické logistické regrese, kdy trenér rozhoduje, zda se nově příchozí hráč hodí na danou pozici či ne. Dospěl jsem k závěru, že zkušenější trenéři se rozhodují o hráči na určitý post ze všech sledovaných charakteristik pravděpodobně pouze podle tělesné výšky. To můžu potvrdit i z vlastních zkušeností, protože basketbal hraju profesionálně už 5 let. Pokud měříte více jak 2 metry, tak nemáte skoro žádnou šanci si zahrát na pozici rozehrávače.

V druhém příkladu jsem se věnoval zjišťování šancí zvolení jednotlivých kandidátů na post prezidenta ČR u voličů dle různých socio-ekonomických charakteristik. Pro lepší interpretovatelnost jsem zvolil pouze 4 kandidáty (Jan Fišer, Karel Schwarzenberg, Jana Bobošíková a Jan Švejnar). Došel jsem k závěru, že ani jedna ze sledovaných charakteristik není statisticky významná. To můžeme chápat tak, že prezident by měl mít své preference napříč celým spektrem občanů České republiky a tak nelze přesně určit skupinu občanů, kteří by volili daného kandidáta. Chyba může být i ve špatném výběru kandidátů, ale v době, kdy jsem dotazník tvořil, byli pouze tito 4 kandidáti rozhodnuti, že se zúčastní prezidentských voleb.

Od začátku jsem chtěl svoje praktické příklady počítat ve statistickém softwaru SAS EG. Postupem času jsem zjistil, že to nebyla nejšťastnější volba. Program SAS EG se během práce nespočetně-krát zasekl a celé postupy bylo třeba dělat znovu. Možná to bylo způsobeno špatnou verzí softwaru. Nicméně moje zkušenosti s tímto programem nejsou v tomto ohledu pozitivní. Na druhou stranu nabízí SAS EG poměrně intuitivní a bohaté prostředí pro analýzu modelů jak klasické (binární) logistické regrese, tak multinomické či ordinální logistické regrese, popř. podmíněné či exaktní logistické regrese. Možnosti využití úlohy Logistic Regression, stejně jako procedury LOGISTIC, jsou tedy širší, než jsem popsal. Metody ordinální, podmíněné nebo exaktní logistické regrese však přenechám pro zpracování do podoby diplomové práce dalším kolegům.

Věřím, že znalosti, dovednosti a zkušenosti získané při psaní této diplomové práce využiji v praxi či případném dalším studiu.

Příloha 1

ID	vek	vyska	hmotnost	BMI	proc_tuku	VO2_max	VK(l)	TF_klid	TF_max	ANP	RQ	post
1	18,3	197	85	22,0	4,8	60,1	5,95	54	195	163	1,10	K
2	21,7	190	86	23,7	8,7	57,9	5,93	54	194	161	1,17	K
3	32,6	204	95	23,1	11,0	52,8	5,59	57	180	149	1,13	P
4	29,0	216	122	26,0	14,9	47,0	8,35	48	168	136	1,18	P
5	18,9	205	89	21,2	5,7	62,7	5,88	57	182	154	1,16	P
6	23,4	206	105	24,7	12,9	50,1	8,20	44	194	156	1,15	P
7	23,3	180	99	30,6	11,0	56,9	4,52	62	182	154	1,05	R
8	28,6	202	101	24,7	6,0			55				P
9	29,4	198	96	24,5	12,0			41				K
10	34,6	197	100	25,8	12,3	52,4	6,77	48	174	143	1,12	K
11	26,7	195	99	26,1	9,8	58,1	6,41	57	176	148	1,11	K
12	26,0	190	84	23,2	6,6	56,9	5,99	45	172	142	1,12	K
13	17,8	195	84	21,9	4,1	64,2	5,36	49	184	155	1,09	K
14	19,3	193	99	26,6	12,0	64,2	6,46	52	187	158	1,11	K
15	22,7	191	85	23,2	9,8	63,9	6,11	44	192	160	1,19	K
16	33,6	204	97	23,2	12,5	54,9	6,88	36	180	145	1,18	P
17	19,9	204	87	21,3	7,2	60,2	6,38	43	178	147	1,21	P
18	24,3	180	101	31,3	11,0	59,8	5,12	46	182	151	1,09	R
19	25,3	185	82	23,8	8,9			55				R
20	27,3	208	117	26,9	17,6	48,9	5,90	51	178	145	1,06	P
21	16,8	191	78	21,3	12,5	63,3	6,00	52	194	163	1,17	K
22	18,3	198	82	21,0	6,0	53,9	5,31	64	185	155	1,15	K
23	30,4	200	96	24,1	11,8	57,3	6,56	38	170	139	1,05	K
24	35,6	199	100	25,3	12,9	56,2	7,15	41	172	141	1,10	K
25	27,7	196	99	25,7	9,8	57,4	6,32	49	172	143	1,23	K
26	23,2	177	79	25,3	11,2	59,5	5,20	61	187	170	1,19	R
27	22,8	209	113	25,9	13,2	48,5	6,59	50	182	166	1,05	P
28	25,8	199	98	24,7	16,7	52,6	5,89	45	177	162	1,16	K
29	27,4	215	105	22,8	18,9	50,3	6,80	57	173	158	1,16	P
30	25,1	204	106	25,5	11,2	48,6	6,18	44	177	162	1,19	P
31	27,8	188	81	22,8	6,3	58,0	5,48	49	181	165	1,19	R
32	23,6	182	80	24,1	10,2	59,3	5,04	45	201	181	1,17	R
33	22,1	201	98	24,3	14,8	59,4	7,37	56	188	170	1,18	K
34	27,5	203	107	26,0	13,7	49,8	6,28	54	195	176	1,18	P
35	22,7	187	78	22,4	9,1	59,7	6,05	50	195	176	1,19	R
36	32,1	199	97	24,4	12,9	54,1	7,04	54	186	169	1,19	K
37	19,3	193	99	26,6	10,0	62,1	6,84	47	194	160	1,03	K
38	20,3	198	94	24,0	8,8	59,8	6,42	69	193	164	1,15	K
39	23,7	191	87	23,8	10,8	60,7	6,48	51	193	160	1,11	K
40	29,0	187	92	26,3	16,1	53,1	6,18	46	177	144	1,13	R
41	28,3	208	121	28,0	20,0	42,8	6,02	65	178	146	1,08	P
42	27,2	197	93	24,0	9,4	59,8	6,19	47	173	144	1,07	K
43	32,9	208	103	23,8	10,3	50,5	7,63	41	178	142	1,05	P
44	26,3	207	103	24,0	12,3	55,7	8,69	47	184	151	1,09	P
45	31,1	200	97	24,2	8,8	58,1	7,23	39	173	142	1,05	K
46	24,5	188	79	22,4	4,8	57,0	5,03	40	189	153	1,06	K
47	24,8	204	107	25,7	10,5	45,8	5,98	57	197	159	1,01	P
48	27,5	186	84	24,3	12,5	58,0	5,50	46	174	145	1,12	R

Obrázek 27: Tabulka sledovaných charakteristik u basketbalistů

Příloha 2

PRŮZKUM

Ve své diplomové práci se mimo jiné zabývám průzkumem preferencí kandidátů v příštích volbách na post prezidenta České republiky. S ohledem na schválení přímé volby prezidenta, bych se Vás chtěl zeptat na jednu otázku. **Jakého kandidáta na post prezidenta byste případně volili??**

Prosím o vyplnění všech dotazovaných políček. Data budou zpracována anonymně, tedy Vaše jméno nebude nikde uvedeno. Předem děkuji za Vaši spolupráci.

KANDIDÁTI:

A	B	C	D
Jan Fišer	Karel Schwarzenberg	Jana Bobošíková	Jan Švejnar

DOTAZOVANÝ/Á:

1. *Pohlaví:* MUŽ ŽENA
2. *Věk:* .....
3. *Dosažené vzdělání:* základní středoškolské vysokoškolské
4. *Pracovní poměr:* zaměstnanec ve veřejném sektoru OSVČ/živnostník
   zaměstnanec v soukr. sektoru student
   zaměstnavatel (soukr. sektor) nezaměstnaný/á
   mateřská/rodičovská dovolená
5. *Měsíční příjem (hrubý):* méně než 10tis 10-20tis 20-40tis
   40-60tis 60tis a více
6. *Hlasoval/a jste v posledních volbách:* ANO NE
7. *Budete hlasovat v přímé volbě prezidenta:* ANO NE
8. *Politické preference (nepovinná položka, vyplňte v případě kladné odpovědi na ot. č. 6):*
   ODS ČSSD VV KSČM
   TOP09 KDU-ČSL jiná strana

Obrázek 28: Dotazník na prezidentské kandidáty

Příloha 3

ID	Pohlavi	Vek	Vzdelani	Prijem	PraceSektory	VladniStrany	Hlasoval	Pr_volba	Kandidat
1	M	29	V	40-60	soukromy	vlada	1	1	A
2	M	56	S	10-20	soukromy	vlada	1	1	D
3	M	59	S	20-40	verejny	opozice	1	1	A
4	Z	53	S	10-20	soukromy	opozice	1	1	A
5	M	56	V	20-40	soukromy	vlada	0	1	D
6	Z	39	V	20-40	soukromy	vlada	1	1	A
7	Z	37	S	0-10	socialni	vlada	1	1	D
8	M	28	V	10-20	soukromy	opozice	1	1	A
9	Z	36	S	0-10	socialni	vlada	1	1	A
10	Z	23	S	0-10	socialni	opozice	1	1	A
11	M	48	Z	10-20	OSVC	opozice	1	1	D
12	Z	25	S	10-20	OSVC	jina	1	1	B
13	M	23	S	10-20	verejny	opozice	1	1	D
14	M	39	S	10-20	verejny	opozice	1	1	D
15	Z	42	S	10-20	verejny		0	0	D
16	M	55	V	10-20	verejny	opozice	1	1	D
17	Z	39	S	10-20	verejny	vlada	1	1	A
18	M	28	V	10-20	verejny	jina	1	1	A
19	Z	44	V	20-40	verejny	opozice	1	1	D
20	Z	26	V	10-20	verejny	opozice	1	1	A
21	M	22	S	0-10	socialni	vlada	1	1	B
22	Z	26	V	10-20	verejny	jina	1	1	A
23	M	24	S	20-40	OSVC	vlada	1	1	A
24	M	45	S	20-40	soukromy	vlada	1	1	B
25	M	32	S	20-40	soukromy	jina	1	1	A
26	M	40	V	40-60	OSVC	vlada	1	1	B
27	M	49	Z	10-20	OSVC	vlada	1	1	B
28	Z	25	V	10-20	verejny	jina	1	1	A
29	Z	30	V	10-20	socialni	vlada	1	1	D
30	M	24	V	0-10	socialni	vlada	1	1	A
31	Z	26	S	10-20	verejny	opozice	1	1	A
32	Z	31	V	20-40	verejny	vlada	1	1	B
33	M	35	V	20-40	verejny	vlada	1	0	D
34	Z	36	S	10-20	soukromy	vlada	1	1	A
35	Z	29	V	0-10	socialni	opozice	1	1	A
36	M	26	V	20-40	soukromy	vlada	0	0	A
37	M	40	V	20-40	OSVC	vlada	1	0	D
38	M	19	S	10-20	OSVC	jina	0	1	A
39	M	20	S	0-10	socialni	vlada	1	1	D
40	M	35	S	10-20	OSVC	vlada	1	1	D
41	Z	29	V	0-10	socialni	vlada	1	1	A
42	M	24	S	20-40	OSVC	vlada	1	0	A
43	M	31	V	60+	OSVC	vlada	1	1	A
44	M	59	Z	10-20	verejny	vlada	1	1	A
45	M	24	S	60+	OSVC	vlada	1	1	B
46	Z	22	S	10-20	verejny	vlada	1	0	B
47	Z	53	S	10-20	verejny	opozice	1	1	A
48	Z	32	S	10-20	verejny	jina	1	1	A
49	Z	36	S	10-20	verejny	vlada	1	1	A
50	Z	46	V	20-40	verejny	vlada	1	1	D
51	M	22	S	10-20	socialni	vlada	1	1	A
52	Z	25	S	10-20	verejny	opozice	0	0	B
53	Z	58	V	20-40	OSVC	vlada	1	1	A
54	M	65	V	20-40	OSVC	vlada	1	1	B
55	M	24	V	20-40	socialni	vlada	1	1	A

56	Z	26	S	10-20	verejny	jina	1	1	B
57	Z	29	V	10-20	soukromy	vlada	1	1	D
58	Z	24	S	0-10	socialni	vlada	1	1	A
59	M	22	S	0-10	socialni		0	1	D
60	M	45	V	20-40	soukromy	vlada	1	0	A
61	M	52	Z	10-20	soukromy	vlada	1	1	D
62	Z	25	S	20-40	socialni	opozice	1	1	A
63	M	41	V	60+	soukromy	vlada	1	1	D
64	M	45	V	60+	soukromy	vlada	1	1	A
65	M	50	V	60+	soukromy	vlada	1	1	A
66	M	42	V	60+	soukromy	vlada	1	1	A
67	Z	40	V	40-60	soukromy	vlada	1	1	A
68	M	24	V	20-40	soukromy	vlada	1	1	A
69	Z	58	S	20-40	soukromy		0	1	A
70	Z	47	V	60+	soukromy	vlada	1	1	A
71	M	54	S	20-40	soukromy	vlada	1	1	A
72	Z	40	V	20-40	soukromy	jina	1	1	A
73	M	36	V	20-40	soukromy	vlada	1	1	A
74	Z	49	Z	10-20	soukromy	jina	1	1	A
75	Z	41	S	20-40	soukromy	vlada	1	1	D
76	M	39	S	40-60	soukromy	opozice	1	1	A
77	M	36	S	40-60	soukromy	vlada	1	1	D
78	Z	35	V	20-40	soukromy	vlada	1	1	D
79	Z	44	V	20-40	soukromy	jina	1	1	A
80	Z	32	V	10-20	soukromy	vlada	0	1	D
81	Z	30	V	20-40	verejny		0	0	A
82	M	30	V	20-40	OSVC		0	1	D
83	Z	40	V	20-40	verejny	vlada	1	1	B
84	M	60	V	0-10	socialni	vlada	1	1	A
85	Z	20	S	0-10	socialni	jina	1	1	B
86	Z	26	V	10-20	soukromy	vlada	1	1	B
87	M	59	S	10-20	socialni	opozice	0	1	A
88	Z	32	S	0-10	socialni	vlada	1	1	A
89	M	48	S	10-20	soukromy	opozice	1	1	D
90	Z	40	S	0-10	soukromy	vlada	1	1	D
91	Z	26	V	0-10	socialni	jina	1	1	D
92	Z	36	S	0-10	socialni	opozice	1	1	A
93	Z	38	S	0-10	socialni	opozice	1	1	A
94	Z	19	Z	0-10	socialni		0	0	B
95	Z	59	V	0-10	OSVC	jina	1	1	A

Obrázek 29: Tabulka socio-ekonomických charakteristik respondentů

Příloha 4

Syntax: LOGISTIC Procedure

The following statements are available in PROC LOGISTIC:

PROC LOGISTIC <options> ;
BY variables ;
CLASS variable <(options)><variable <(options)>...></ options> ;
CONTRAST 'label' effect values<, effect values,...></ options> ;
EFFECT name = effect-type (variables </ options>) ;
EFFECTPLOT <*plot-type*<(plot-definition-options)>></ options> ;
ESTIMATE <'label'> estimate-specification </ options> ;
EXACT <'label'><**INTERCEPT**><effects></ options> ;
EXACTOPTIONS options ;
FREQ variable ;
LSMEANS <model-effects> </ options> ;
LSMESTIMATE model-effect lsmestimate-specification </ options> ;
<*label:*> MODEL events/trials=<effects></ options> ;
<*label:*> MODEL variable <(variable_options)>=<effects></ options> ;
ODDSRATIO <'label'> variable </ options> ;
OUTPUT <**OUT**=SAS-data-set><keyword=name <keyword=name...>></ option> ;
ROC <'label'> <specification> </ options> ;
ROCCONTRAST <'label'><contrast></ options> ;
SCORE <options> ;
SLICE model-effect </ options> ;
STORE <**OUT**=>*item-store-name* </ **LABEL**='*label*'> ;
STRATA effects </ options> ;
<*label:*> TEST equation1 <,equation2,...></ option> ;
UNITS independent1=list1 <independent2=list2 ...></ option> ;
WEIGHT variable </ option> ;

Obrázek 30: Syntaxe procedury LOGISTIC

Literatura

[1] GILL, Jeff. Generalized linear models: a unified approach. Thousand Oaks, Calif.: Sage Publications, Inc., c2001, 101 s. Sage university papers series, no. 134. ISBN 07-619-2055-2.

[2] KLEINBAUM, David G. and Mitchel KLEIN. Logistic Regression: A Self-Learning Text. Third Edition. Atlanta: Springer, 2010. ISBN 978-1-4419-1741-6.

[3] HOSMER, David W. and Stanley LEMESHOW. Apllied Logistic Regression. Second Edition. Canada: John Wiley & Sons, INC., 2000. ISBN 0-471-35632-8.

[4] ANDĚL, J. Matematická statistika. SNTL Praha, 1985.

[5] KUNDEROVÁ, P. Základy pravděpodobnosti a matematické statistiky. Olomouc: Vydavatelství UP Olomouc, 2004.

[6] ALBERT, A. and J. A. ANDERSON. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. Biometrika, Vol. 71. 1984, s. 1-10. Dostupné z: http://www.jstor.org/stable/2336390

[7] YING, So. A Tutorial on Logistic Regression. SUGI Proceedings. SAS Institute Inc., Cary, NC, 1995.

[8] SEHNALOVÁ, Michala. Logistická regrese. Olomouc, 2009. Diplomová práce. UP Olomouc. Vedoucí práce Prof.RNDr. Ing. Lubomír Kubáček, DrSc.

[9] Sportvital [online]. 2010 [cit. 2012-02-21]. Dostupné z: http://www.sportvital.cz/rejstrik/

[10] Citace.com [online]. 2004 [cit. 2012-03-15]. Dostupné z: http://www.citace.com/

[11] The LOGISTIC Procedure: Syntax. SAS Costumer Support Knowledge Base and Community [online]. 2012 [cit. 2012-03-18]. Dostupné z: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect003.htm