UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Analýza vlivu hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodnost

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Eva Fišerová, Ph.D.**
Vypracoval(a): **Sylva Šmoldasová**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2018

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Sylva Šmoldasová

Název práce: Analýza vlivu hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodnost

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2018

Abstrakt: Práce se zabývá analýzou vlivu hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodnost. Pro analýzu je zvolena metoda logistické regrese, o které pojednává teoretická část. Jsou popsány principy výstavby logistického regresního modelu, metody odhadu regresních koeficientů, způsoby hodnocení kvality modelu a možnosti porovnání více modelů mezi sebou. V praktické části je vytvořen a analyzován model popisující vliv hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění. S cílem zpřesnit vytvořený model, jsou přidáním proměnných nalezeny další modely. Na závěr jsou všechny vytvořené modely porovnány. Analýza je provedena za pomoci statistického softwaru R, verze 3.3.1.

Klíčová slova: Antimülleriánský hormon, Folikulostimulační hormon, logistická regrese, metoda maximální věrohodnosti, Newtonova-Raphosonova metoda, test poměrem věrohodností, křížová validace, testy dobré shody, ROC křivka, software R

Počet stran: 83

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Sylva Šmoldasová

Title: Analysis of the effect of antimüllerian hormone levels on the probability of getting pregnant in women treated for infertility.

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2018

Abstract: The thesis deals with the analysis of the effect of antimüllerian hormone levels on the probability of getting pregnant in women treated for infertility. The logistic regression method, which is discussed in the theoretical part, is chosen for the analysis. The principles of the construction of the logistic regression model, methods of estimation of regression coefficients, methods of evaluation of the quality of the model, and techniques for comparison of multiple models are described. In the practical part, a model describing the effect of antimüllerian hormone levels on the probability of getting pregnant is created and analyzed. To improve created model, other models were created by adding variables. Finally, all created models are compared. The analysis is performed using the statistical software R, version 3.1.1.

Key words: Antimüllerian hormone, Follicle–stimulating hormone, logistic regression, maximum likelihood method, Newton-Raphson method, likelihood-ratio test, cross-validation, goodness of fit tests, ROC curve, Software R

Number of pages: 83

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Děkuji vedoucí mé bakalářské práce, paní doc. RNDr. Evě Fišerové, Ph.D., za odborné vedení, ochotu, cenné rady a věnovaný čas. Dále bych ráda poděkovala své rodině a přátelům za podporu při studiu.

Úvod

Jak už název napovídá, cílem této práce je zkoumat vliv hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodnost. Tento hormon ovlivňuje zrání vajíček ve vaječníku a jeho hladina souvisí právě s jejich počtem. Můžeme tedy říci, že je jakýmsi ukazatelem ovariální rezervy u žen. Analýzu budeme provádět na reálných datech metodou logistiké regrese, kterou si nejprve podrobně představíme.

Logistická regrese neboli také logistický regresní model či model logistické regrese je regresní model navržený v roce 1958 Davidem Coxem [6] jako alternativa k metodě nejmenších čtverců. V současnosti se tato metoda hojně využívá v nejrůznějších oborech od biologie a medicíny, přes bankovnictví až například po kriminologii. Cílem logistické regrese je najít model, který popíše vztah mezi nezávislými vysvětlujícími proměnnými a závislou vysvětlovanou proměnnou, která je diskrétního typu. Tato závislá proměnná představuje nastoupení či nenastoupení nějakého jevu či stavu. Například zda žena otěhotní či neotěhotní, zda onemocníme nějakou konkrétní chorobou či nikoli, v bankovnictví zda je daná osoba schopna splácet úvěr nebo ve školství, zda je žák schopen složit zkoušky.

Závislá proměnná tedy může nabývat dvou a více hodnot a podle jejího typu rozlišujeme následující druhy logistických regresí[15]:

- Binární logistická regrese – vysvětlovaná proměnná je alternativní (binární, dichotomická), nabývá tedy pouze dvou možných hodnot, například nastoupení či nenastoupení nějakého jevu, žena či muž, přežil či nepřežil.

- Ordinální logistická regrese – jak název napovídá, vysvětlovaná proměnná je ordinálního typu, tedy nabývá více než dvou hodnot, které můžeme seřadit, například známky studentů u zkoušky nebo sportovní výkony při hodnocení nízký výkon, podprůměr, průměr, nadprůměr a vysoký výkon.

- (multi)nominální logistická regrese – v tomto případě je vysvětlovaná proměnná opět kategoriální a nabývá nejméně tří různých úrovní, přičemž mezi nimi neexistuje pořadí, pouze odlišnost, například krevní skupina, barva, atd.

Zatímco na závislou proměnnou máme výše popsanou podmínku kategorizace, vysvětlující proměnná může být jak kategoriálního typu, tzv. faktor, tak i typu spojitého, tzv. prediktor. Příkladem tak může být vliv věku, pohlaví, rodinných dispozic, kouření, BMI, sedavého způsobu života a hladiny cholesterolu na výskyt infarktu.

V této práci si představíme binární logistickou regresi, neboť chceme zkoumat pravděpodobnost otěhotnění či neotěhotnění ženy, tedy úspěchu a neúspěchu.

V první kapitole si představíme závislou proměnnou vstupující do logistické regrese, zavedeme některá potřebná značení a popíšeme si princip výstavby logistického modelu. Druhá kapitola nás seznámí s metodami a postupy pro odhad regresních koeficientů. Ve třetí kapitole se budeme zabývat metodami pro výběr vhodného modelu. V kapitole čtvrté si představíme hodnotící metody, které nám poskytnou informace o kvalitě námi vytvořeného modelu. Nakonec, v páté kapitole provedeme praktickou analýzu vlivu hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění žen léčících se na neplodnost.

Kapitola 1

Úvod do logistické regrese

V této kapitole si popíšeme, jaký tvar má vysvětlovaná náhodná veličina v logistické regresi, zavedeme si některá značení a základní pojmy. V druhé části této kapitoly si popíšeme princip výstavby logistického modelu. Vycházíme přitom ze znalostí pravděpodobnosti a matematické statistiky na úrovni vysoké školy, např. [12].

1.1. Základní poznatky

Logistická regrese se snaží predikovat pravděpodobnost nastoupení resp. nenastoupení nějakého jevu či stavu na základě nezávislých náhodných veličin. Tyto vysvětlující nezávislé proměnné si označme jako $X_1, \ldots, X_m$ a vysvětlovanou závislou proměnnou jako Y. Realizace náhodné veličiny Y nás informuje o tom, zda jev nastoupil či nikoli. V případě nastoupení jevu budeme mluvit o úspěchu, naopak v případě nenastoupení jevu o neúspěchu. Předpokládáme tedy, že náhodná veličina Y má alternativní rozdělení s parametrem π, $\pi \in (0, 1)$, značíme $Y \sim Alt(\pi)$. Náhodná veličina Y nabývá hodnoty 1 (značí úspěch) nebo 0 (značí neúspěch) s následujícími pravděpodobnostmi:

$$P(Y = 0) = 1 - \pi, \qquad P(Y = 1) = \pi. \tag{1.1}$$

Jinak řečeno, pravděpodobnost neúspěchu je rovna $1 - \pi$ a pravděpodobnost úspěchu je rovna π. Platí

$$P(Y = 1) = 1 - P(Y = 0). \tag{1.2}$$

Můžeme psát

$$P(Y=y)=\begin{cases}\pi, & y=1,\\ 1-\pi, & y=0,\\ 0, & jinak.\end{cases} \tag{1.3}$$

Střední hodnota a rozptyl takovéto náhodné veličiny je:

$$E(Y)=\pi, \qquad var(Y)=\pi(1-\pi). \tag{1.4}$$

Pro úplnost můžeme uvést pravděpodobnostní funkci náhodné veličiny Y:

$$P(Y=y)=\pi^y(1-\pi)^{1-y}, \qquad y\in\{0,1\}. \tag{1.5}$$

Někdy nás může zajímat informace, kolikrát je vyšší pravděpodobnost úspěchu oproti neúspěchu. Odpověď na tuto otázku nám dává *šance* (anglicky *odds*), kterou definujeme jako podíl pravděpodobností nastoupení jevu ku nenastoupení jevu. V této práci budeme šanci značit řeckým písmenem ω:

$$\omega(1)=\frac{P(Y=1)}{P(Y=0)}=\frac{P(Y=1)}{1-P(Y=1)}. \tag{1.6}$$

Dalším pojem, který je třeba zavést je *logit*, který získáme logaritmováním šance, tedy:

$$logit\left(P\left(Y=1\right)\right)=\ln(\omega(1))=\ln\left(\frac{P(Y=1)}{P(Y=0)}\right). \tag{1.7}$$

Je dobré povšimnout si faktu, že z veličiny Y, která nabývala pouze dvou hodnot $\{0,1\}$, jsme získali šanci, která nabývá hodnot v intervalu $(0,\infty)$ a nakonec logitovou funkci, která se realizuje na intervalu $(-\infty,\infty)$.

1.2. Tvorba regresního modelu

Je zřejmé, že při tvorbě regresního modelu, kdy je závislá proměnná alternativní, nemůžeme postupovat stejně jako v případě, kdy je vysvětlovaná proměnná spojitého typu

(tedy jako u klasické lineární regrese). Jestliže bychom využili model

$$E(Y|(X_1,\ldots,X_m)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_m X_m,$$

pak by pro i-té pozorování platilo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_m X_{mi} + \epsilon_i, i = 1,\ldots,n$, kde $\beta_0,\ldots\beta_m$ jsou neznámé parametry, $X_1,\ldots,X_m$ jsou vysvětlující proměnné a ϵ_i náhodná chyba při i-tém pozorování. Výsledky, které bychom dostali tímto postupem by byly zjevně chybné a nedávaly by žádný smysl, neboť levá strana této rovnosti by nabývala pouze hodnot 0 nebo 1, zatímco na straně pravé budou hodnoty zajisté jiné. Je tedy třeba najít způsob jak převést binární proměnnou na spojitou, která může nabývat libovolných hodnot.

Jak již ale bylo řečeno v úvodu této kapitoly, logistická regrese se snaží predikovat *pravděpodobnost* toho, že náhodná veličina Y nabude nějaké hodnoty. Díky rovnosti (1.2) si můžeme sami zvolit, zda budeme modelovat pravděpodobnost nastoupení či nenastoupení jevu. V této práci budeme modelovat pravděpodobnost, že jev nastoupí, tedy, že se závislá veličina Y realizuje hodnotou 1. Dostáváme:

$$P(Y=1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_m X_m.$$

Levá strana rovnice nyní představuje pravděpodobnost a nabývá pouze hodnot z intervalu (0,1), zatímco na pravé straně rovnice máme spojité a kategoriální proměnné a neznámé parametry β_j, pro které platí $-\infty < \beta_j < \infty$. Při odhadech neznámých parametrů $\beta_0,\ldots\beta_m$ se můžeme dostat do situace, kdy predikované hodnoty pravděpodobnosti pro nějakou konkrétní realizaci $x_{1i},\ldots,x_{mi}$ náhodných veličin $X_1,\ldots,X_m$ budou ležet mimo interval (0,1).

Řešení tohoto problému jsme si nastínili už dříve, neboť *šance* též modeluje pravděpodobnost $P(Y=1)$. Nicméně šance nám také nestačí, neboť její hodnoty leží v intervalu $(0,\infty)$, zatímco *logit*, tedy hodnoty logaritmované šance se realizují na intervalu $(-\infty,\infty)$. Logit je příkladem spojovací funkce, neboť určuje vztah mezi závislou proměnnou a regresní funkcí. Platí tedy rovnost

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_m X_m. \tag{1.8}$$

Na tomto místě je vhodné poznamenat, že parametr β_0 je roven velikosti logitu pro referenční kategorie (nulové hodnoty) všech vysvětlujících proměnných. Pro $\beta_0 > 0$ je šance, že $Y = 1$ větší jak jedna, tedy $\pi > 0.5$. Naopak pro $\beta_0 < 0$ je tato šance menší jak jedna ($\pi < 0.5$) a nakonec pro $\beta_0 = 0$ je tato šance vyrovaná, neboli $\pi = 0.5$

Před dalšími úpravami si zavedeme značení $\boldsymbol{\beta} = (\beta_0, \dots \beta_m)'$ a $\mathbf{X} = (1, X_1, \dots, X_m)'$, které nám usnadní zápisy. Nyní si uvedenou rovnici (1.8) upravíme, abychom získali vztah pro pravděpodobnost $P(Y = 1)$. V prvním kroku se budeme chtít zbavit přirozeného logaritmu, proto celou rovnici odlogaritmujeme a dostaneme:

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

ve druhém kroku se zbavíme zlomku na levé straně vynásobením celé rovnosti jmenovatelem, tedy

$$P(Y = 1) = [1 - P(Y = 1)]\, e^{\boldsymbol{\beta}'\mathbf{X}}.$$

Dále roznásobíme závorku na pravé straně a člen obsahující $P(Y = 1)$ převedeme na stranu levou

$$P(Y = 1) + P(Y = 1)e^{\boldsymbol{\beta}'\mathbf{X}} = e^{\boldsymbol{\beta}'\mathbf{X}}.$$

Nakonec na levé straně vytkneme $P(Y = 1)$ a celou rovnici vydělíme výrazem $(1 + e^{\boldsymbol{\beta}'\mathbf{X}})$, dostaneme tak

$$P(Y = 1) = \frac{e^{\boldsymbol{\beta}'\mathbf{X}}}{1 + e^{\boldsymbol{\beta}'\mathbf{X}}}. \tag{1.9}$$

Analogicky lze určit $P(Y = 0)$ využitím vztahu $1 - P(Y = 1)$:

$$P(Y = 0) = 1 - \frac{e^{\boldsymbol{\beta}'\mathbf{X}}}{1 + e^{\boldsymbol{\beta}'\mathbf{X}}} = \frac{1}{1 + e^{\boldsymbol{\beta}'\mathbf{X}}}. \tag{1.10}$$

Označme jako $\mathbf{x} = (1, x_1, \dots, x_m)'$ realizace náhodného vektoru $\mathbf{X}$. Zjevně pravděpodobnosti (1.9) a (1.10) budou pro různé realizace $\mathbf{x}$ nabývat různých hodnot. Proto je třeba tyto pravděpodobnosti uvažovat jako podmíněné, tj.

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}}}, \qquad P(Y = 0|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + e^{\boldsymbol{\beta}'\mathbf{x}}}. \tag{1.11}$$

Můžeme si povšimnout, že na vysvětlující náhodné veličiny nejsou kladeny žádné podmínky, tím se logistická regrese stává velmi praktickou.

Zbývá zavést pojem *poměr šancí* (anglicky *odds ratio*), obvykle jej označujeme OR. Umožnuje nám srovnání dvou různých skupin pomocí šancí. Jeho interpretace se liší podle typu proměnné. V případě spojité proměnné $\mathrm{OR}(X_j)$ říká, kolikrát se zvětší šance, aby náhodná veličina Y nabyla hodnoty 1, jestliže se hodnota spojité vysvětlující proměnné X_j zvýší o jednotku a zbylé vysvětlující náhodné veličiny (celkem m -1 proměnných) zůstanou nezměněné, což zachycuje následující vztah:

$$\mathrm{OR}(X_j) = \frac{\omega(P(Y=1|X_1=x_1,\ldots,X_{j-1}=x_{j-1},X_j=x_j+1,X_{j+1}=x_{j+1},\ldots,X_m=x_m))}{\omega(P(Y=1|X_1=x_1,\ldots,X_{j-1}=x_{j-1},X_j=x_j,X_{j+1}=x_{j+1},\ldots,X_m=x_m))}. \quad (1.12)$$

Tento vztah si upravme. Uvažujme nejprve pouze čitatel, do kterého dosaďme z (1.6):

$$\frac{P(Y=1|X_1=x_1,\ldots,X_{j-1}=x_{j-1},X_j=x_j+1,X_{j+1}=x_{j+1},\ldots,X_m=x_m)}{1-P(Y=1|X_1=x_1,\ldots,X_{j-1}=x_{j-1},X_j=x_j+1,X_{j+1}=x_{j+1},\ldots,X_m=x_m)}. \quad (1.13)$$

Následně dosaďme z (1.11) vztahy pro dané pravděpodobnosti:

$$\frac{\frac{e^{\beta_0+\beta_1x_1+\ldots+\beta_{j-1}x_{j-1}+\beta_j\left(x_j+1\right)+\beta_{j+1}x_{j+1}+\ldots+\beta_mx_m}}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_{j-1}x_{j-1}+\beta_j\left(x_j+1\right)+\beta_{j+1}x_{j+1}+\ldots+\beta_mx_m}}}{1-\frac{e^{\beta_0+\beta_1x_1+\ldots+\beta_{j-1}x_{j-1}+\beta_j\left(x_j+1\right)+\beta_{j+1}x_{j+1}+\ldots+\beta_mx_m}}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_{j-1}x_{j-1}+\beta_j\left(x_j+1\right)+\beta_{j+1}x_{j+1}+\ldots+\beta_mx_m}}}. \quad (1.14)$$

Toto lze upravit a zkrátit na výraz:

$$e^{\beta_0+\beta_1x_1+\ldots+\beta_{j-1}x_{j-1}+\beta_j(x_j+1)+\beta_{j+1}x_{j+1}+\ldots+\beta_mx_m}. \quad (1.15)$$

Výraz ve jmenovateli OR si vyjádříme analogickým způsobem a dostaneme:

$$e^{\beta_0+\beta_1x_1+\ldots+\beta_{j-1}x_{j-1}+\beta_jx_j+\beta_{j+1}x_{j+1}+\ldots+\beta_mx_m}. \quad (1.16)$$

Podle pravidel pro počítání s mocninami můžeme přejít ke krácení, přičemž nám zůstane jedinný člen v čitateli a to e^{β_j}. Dostali jsme vztah:

$$OR(X_j) = e^{\beta_j}, j = 1,\ldots,m. \quad (1.17)$$

Pro kategoriální proměnnou je interpretace OR odlišná. V tomto případě potřebujeme tzv. *referenční (výchozí) kategorii*. Ostatní kategorie dané proměnné poté srovnáváme právě s referenční. V případě, že má uvažovaná proměnná pouze dvě kategorie, postupujeme tak, že jednu z nich zvolíme jako referenční. Příslušný regresní koeficient β poté interpretujeme jako změnu šance na úspěch oproti výchozí kategorii.

V případě, že má uvažovaná nezávislá proměnná více než dvě kategorie ($r > 2$), zavedeme $r-1$ nezávislých indikátorových proměnných. Tyto proměnné odpovídají jednotlivým

kategoriím původní proměnné a nabývají hodnot 0 nebo 1 podle toho, zda dané pozorování do příslušné kategorie patří či nikoli. Vynechaná kategorie je referenční. Celkem tak získáme $r - 1$ regresních koeficientů, které interpretujeme jako změnu šance na úspěch oproti referenční kategorii.

Nyní víme, jak interpretovat parametry β_j, $j = 1, \ldots, m$. Zbývá vysvětlit interpretaci pro parametr β_0. Parametr β_0 je roven velikosti logitu pro nulové hodnoty všech vysvětlujících proměnných. Jestliže je tedy tento parametr roven nule, znamená to, že šance na úspěch i neúspěch jsou vyrovnané ($\pi = 0.5$). Pro kladné hodnoty parametru β_0 je šance na úspěch oproti neúspěchu větší ($\pi > 0.5$) a naopak pro záporné hodnoty tohoto parametru je tato šance menší ($\pi < 0.5$).

Kapitola 2

Odhad koeficientů logistického modelu

Pro konstrukci odhadů koeficientů se v logistické regresi využívá *metody maximální věrohodnosti* [1, 2, 23], která vyústí v soustavu nelineárních rovnic. Výsledné odhady koeficientů získáme použitím vhodného iteračního algoritmu, často využívaná je *Newtonova-Raphsonova metoda* [1, 2, 8].

2.1. Metoda maximální věrohodnosti

Metoda maximální věrohodnosti (anglicky *maximum-likelihood estimation*, zkráceně MLE) je spojena s pojmem *funkce věrohodnosti* (anglicky *likelihood function*). Cílem MLE je najít takový odhad parametru, resp. vektoru parametrů $\boldsymbol{\theta}$, pro nějž bude *funkce věrohodnosti* maximální.

Abychom mohli funkci věrohodnosti zavést, je nutné znát sdružené rozdělení náhodného vektoru $\mathbf{Y}$. Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, kde Y_i, $i = 1, \ldots, n$, jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti. V případě spojitého rozdělení s hustotou $f(y|\boldsymbol{\theta})$ a v případě diskrétního rozdělení s pravděpodobnostní funkcí $p(y|\boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \in \Omega$ je vektor neznámých parametrů.

Předpokládejme diskrétní rozdělení. Sdružená pravděpodobnostní funkce náhodného vektoru $\mathbf{Y}$ má tvar

$$p(y_1, \ldots, y_n|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}), \tag{2.1}$$

přičemž $\boldsymbol{\theta}$ zde stojí v roli parametrů a $\mathbf{y} = (y_1, \ldots, y_n)'$ uvažujeme jako proměnné. Věrohodnostní funkci definujeme stejným předpisem jako tuto sdruženou hustotu, ale s tím

rozdílem, že $\boldsymbol{\theta}$ bude nyní představovat proměnnou, $y_1, \ldots, y_n$ parametry a budeme ji značit $L(\boldsymbol{\theta}|\mathbf{y})$. Funkce věrohodnosti má tedy tvar

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}). \tag{2.2}$$

MLE můžeme interpretovat jako metodu, která hledá takový odhad $\hat{\boldsymbol{\theta}}$, který maximalizuje pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení. Odhad $\hat{\boldsymbol{\theta}}$ nazveme *maximálně věrohodný* právě tehdy, když hodnota věrohodnostní funkce bude pro tento odhad větší nebo rovna hodnotě věrohodnostní funkce pro jiné odhady $\tilde{\boldsymbol{\theta}}$ při fixním $\mathbf{y} = (y_1, \ldots, y_n)'$.

V některých případech může být maximalizování funkce věrohodnosti v tomto tvaru je příliš komplikované. Zjednodušit jej můžeme logaritmováním, díky kterému přejdeme od násobení ke sčítání. Získáme tak tzv. *logaritmickou funkci věrohodnosti*, píšeme

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \ln \left(\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) \right) = \sum_{i=1}^{n} \ln p(y_i|\boldsymbol{\theta}). \tag{2.3}$$

Maximum logaritmické věrohodnostní funkce hledáme pomocí parciálních derivací této funkce podle všech proměnných, které srovnáme s nulou. Nakonec je třeba ověřit, že takto nalezený odhad opravdu maximalizuje věrohodnostní funkci. Předpokládejme tedy, že logaritmická věrohodnostní funkce má parciální derivace alespoň druhého řádu na množině Ω. Nejprve pomocí prvních parciálních derivací vytvoříme tzv. *systém věrohodnostních rovnic*

$$\frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{y})}{\partial \theta_i} = 0, \ i = 1, \ldots, m. \tag{2.4}$$

Protože řešení tohoto systému rovnic nelze najít obecně v algebraickém tvaru, je nutné použít numerické metody. Jednou z nejpoužívanějších numerických metod v logistické regresi je Newtonova-Raphsonova metoda, kterou si ukážeme později. Prozatím předpokládejme, že jsme toto řešení nalezli a označme si jej $\hat{\boldsymbol{\theta}}$.

Zbývá pomocí Hessovy matice druhých parciálních derivací logaritmické věrohodnostní funkce ověřit, že naše řešení opravdu maximalizuje logaritmickou věrohodnostní funkci. Aby toto platilo, musí být Hessova matice negativně definitní. Při ověřování využijeme faktu, že obecně matice $\mathbf{A}$ je pozitivně (semi)definitní právě tehdy, když je matice $\mathbf{-A}$

negativně (semi)definitní. Stačí nám tedy ověřit, že mínus Hessova matice je pozitivně definitní dle definice: $\mathbf{x} \neq 0 \Rightarrow \mathbf{x}'\mathbf{A}\mathbf{x} > 0$. Hessova matice je definovaná jako

$$\mathbf{H}(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial^2 \ln L(\boldsymbol{\theta}|\mathbf{y})}{\partial\theta_i \partial\theta_j} \right)_{i,j=1}^{m}. \tag{2.5}$$

2.1.1. Vlastnosti maximálně věrohodných odhadů

Předpokladem pro platnost následujících vlastností je dostatečně velký počet pozorování. Tento předpoklad nesmí být porušen, neboť vlastnosti ML odhadů jsou asymptotické a jeho porušením bychom mohli získat zavádějící výsledky.

Než si tyto vlastnosti přiblížíme, je třeba zavést tzv. *Fisherovu informační matici* [8], která představuje míru informace o parametru $\boldsymbol{\theta}$ obsaženou v jednom pozorování. Fisherovu informační matici značíme $\mathbf{J}(\boldsymbol{\theta})$ a její prvky získáme jako kovariance parciálních derivací logaritmu sdružené hustoty náhodného výběru $\mathbf{Y}$ podle θ_i kde $i = 1, \ldots, m$, tj.

$$J_{ij}(\boldsymbol{\theta}) = cov\left(\frac{\partial \ln f(\mathbf{Y}|\boldsymbol{\theta})}{\partial\theta_i}, \frac{\partial \ln f(\mathbf{Y}|\boldsymbol{\theta})}{\partial\theta_j} \right) = E\left(\frac{\partial \ln f(\mathbf{Y}|\boldsymbol{\theta})}{\partial\theta_i} \cdot \frac{\partial \ln f(\mathbf{Y}|\boldsymbol{\theta})}{\partial\theta_j} \right). \tag{2.6}$$

Prvky Fisherovy informační matice lze vyjádřit i v následujícím tvaru:

$$J_{ij}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln f(\mathbf{Y}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right). \tag{2.7}$$

Nyní si už můžeme představit vlastnosti odhladů nalezených metodou maximální věrohodnosti.

- Asymptotická normalita - pro dostatečně velký počet pozorování n má $\hat{\boldsymbol{\theta}}$ přibližně normální rozdělení, tj.

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \quad \dot{\sim} \quad \mathbf{N}(0, [\mathbf{J}(\boldsymbol{\theta})]^{-1}).$$

- (Slabá) konzistence - s rostoucím počtem pozorování konverguje podle pravděpodobnosti odhad parametru $\hat{\boldsymbol{\theta}}$ ke skutečné hodnotě $\boldsymbol{\theta}$, tj.

$$\lim_{n\to\infty} P\left(\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|^2 < \epsilon \right) = 1, \forall\epsilon > 0, \forall\boldsymbol{\theta} \in \Theta.$$

- Asymptotická eficience - neboli vydatnost znamená, že náš odhad má mezi dalšími konzistentními odhady nejmenší rozptyl.

2.1.2. MLE v logistické regresi

Začněme opět sdruženým rozdělením náhodného vektoru. Uvažujme náhodný výběr $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ z alternativního rozdělení s parametrem π a realizacemi $\mathbf{y} = (y_1, \ldots, y_n)'$. Pro pravděpodobnostní funkci náhodné veličiny Y_i dle vztahu (1.5) platí:

$$P(Y_i = y_i) = \pi^{y_i}(1-\pi)^{1-y_i}, \quad y_i \in \{0,1\}, i = 1, \ldots, n. \tag{2.8}$$

Sdruženou pravděpodobnostní funkci náhodného vektoru $\mathbf{Y}$ můžeme s využitím předpokladu nezávislosti náhodných veličin Y_i a vztahu (2.8) zapsat jako:

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}, \quad y_i \in \{0,1\}, i = 1, \ldots, n. \tag{2.9}$$

Dále víme, že $\pi = E(Y_i) = P(Y_i = 1)$ a $1 - \pi = P(Y_i = 0)$. Ze vztahu (1.11) si můžeme vyjádřit podmíněnou pravděpodobnost toho, že se náhodná veličina Y_i realizuje hodnotou 1 za podmínky, že se náhodný vektor X_i realizoval hodnotami x_i. Protože tato hodnota závisí na konkrétních realizacích náhodného vektoru X_i, označme si ji jako π_i:

$$\pi_i(\boldsymbol{\beta}) = P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}}. \tag{2.10}$$

Pravděpodobnost $P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i)$ získáme analogicky a označíme jako $1 - \pi_i(\boldsymbol{\beta})$. Dosazením těchto vztahů do sdružené pravděpodobnostní funkce dostaneme věrohodnostní funkci proměnných $\beta_0, \ldots, \beta_m$ s parametry $y_1, \ldots, y_n$. Potom věrohodnostní funkce je ve tvaru:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}} \right)^{1-y_i} = \prod_{i=1}^{n} \frac{(e^{\boldsymbol{\beta}'\mathbf{x}_i})^{y_i}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}}. \tag{2.11}$$

Logaritmická věrohodnostní funkce je následně ve tvaru

$$\ln L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\boldsymbol{\beta}'\mathbf{x}_i) - \sum_{i=1}^{n} \ln(1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}). \tag{2.12}$$

Systém věrohodnostních rovnic získáme z parciálních derivací této funkce podle všech proměnných, které srovnáme s nulou.

$$0 = \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{x_{ij} e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}}, \quad j = 0, 1, \ldots, m. \tag{2.13}$$

Tento systém lze dále upravit

$$\sum_{i=1}^{n}\left[y_i - \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right]x_{ij} = 0, \quad j = 0,\ldots,m. \tag{2.14}$$

Získali jsme nelineární soustavu $m+1$ věrohodnostních rovnic o $m+1$ neznámých. Řešením soustavy jsou odhady $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)'$, které, jak už bylo řečeno, budeme hledat iteračně. Tento postup si ukážeme později, nyní předpokládejme, že jsme řešení nalezli. Zbývá tedy sestavit Hessovu matici, která slouží k ověření, zda námi nalezené odhady opravdu maximalizují věrohodnostní funkci:

$$\mathbf{H}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial^2\beta_0} & \frac{\partial^2 lnL(\boldsymbol{\beta})}{\partial\beta_0\partial\beta_1} & \cdots & \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_0\partial\beta_m} \\ \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_0} & \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial^2\beta_1} & \cdots & \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 lnL(\boldsymbol{\beta})}{\partial\beta_m\partial\beta_0} & \cdots & \cdots & \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial^2\beta_m} \end{pmatrix}.$$

Prvky této matice získáme následujícím postupem:

$$\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k} = \frac{\partial}{\partial\beta_k}\left[\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n}\frac{x_{ij}e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right] = -\sum_{i=1}^{n} x_{ij}\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}x_{ik}\left(1+e^{\boldsymbol{\beta}'\mathbf{x}_i}\right) - e^{\boldsymbol{\beta}'\mathbf{x}_i}x_{ik}e^{\boldsymbol{\beta}'\mathbf{x}_i}}{\left(1+e^{\boldsymbol{\beta}'\mathbf{x}_i}\right)^2}.$$

Po vytknutí výrazu $x_{ik}e^{\boldsymbol{\beta}'\mathbf{x}_i}$ získáme:

$$\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k} = -\sum_{i=1}^{n} x_{ij}x_{ik}\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}\left(1+e^{\boldsymbol{\beta}'\mathbf{x}_i} - e^{\boldsymbol{\beta}'\mathbf{x}_i}\right)}{\left(1+e^{\boldsymbol{\beta}'\mathbf{x}_i}\right)^2} = -\sum_{i=1}^{n} x_{ij}x_{ik}\left(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)\left(\frac{1}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right).$$

Nyní je třeba ověřit, že je matice s těmito prvky skutečně negativně definitní. Jak už bylo řečeno výše, ověření provedeme pomocí mínus Hessovy matice, která musí být pozitivně definitní ($\mathbf{x} \neq \mathbf{0} \Rightarrow \mathbf{x}'\left[\text{-}\mathbf{H}\right]\mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{R}^m$, kde m je počet sloupců matice $\mathbf{H}$). Prvky mínus Hessovy matice můžeme zapsat jako:

$$-H_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik}\underbrace{\left(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)\left(\frac{1}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)}_{pravděpodobnost>0}, \quad j,k = 0,\ldots,m.$$

S využitím (2.10) a dosazením $\hat{\boldsymbol{\beta}}$ za $\boldsymbol{\beta}$ si tento vztah vyjádříme maticově jako:

$$-H_{jk} = -\mathbf{x}_j' \mathbf{U} \mathbf{x}_k, \quad j,k = 0,\ldots,m, \tag{2.15}$$

kde $\mathbf{U} = diag\left(\pi_1(\hat{\boldsymbol{\beta}})(1-\pi_1(\hat{\boldsymbol{\beta}})),\ldots,\pi_n(\hat{\boldsymbol{\beta}})(1-\pi_n(\hat{\boldsymbol{\beta}}))\right)$ je diagonální matice $n \times n$ s prvky $\pi_i(\hat{\boldsymbol{\beta}})(1-\pi_i(\hat{\boldsymbol{\beta}}))$, kde $i = 1,\ldots,n$ na digonále a $\mathbf{x}_j, \mathbf{x}_k$ jsou sloupce regresní matice $\mathbf{X}$ (matice plánu). Maticový zápis mínus Hessovy matice je ve tvaru:

$$\textbf{-H} = \mathbf{X}'\mathbf{U}\mathbf{X}.$$

Odtud vidíme, že bude platit podmínka pro pozitivní definitnost mínus Hessovy matice. Tedy:

$$-\mathbf{x}' \left[\mathbf{H}\right] \mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{R}^m, \mathbf{x} \neq \mathbf{0}.$$

Je-li tedy mínus Hessova matice pozitivně definitní, můžeme říci, že Hessova matice je negativně definitní. Jinak řečeno, věrohodnostní funkce pro logistickou regresi má negativně definitní Hessián.

Hessovu matici lze kromě ověření, že námi nalezený odhad vektoru parametrů maximalizuje věrohodnostní funkci, využít i pro odhad varianční matice maximálně věrohodných odhadů $\hat{\boldsymbol{\beta}}$. Jak jsme si uvedli, jednou z vlastností MLE odhadů je asymptotická normalita, přičemž varianční matice je ve tvaru $[\mathbf{J}(\boldsymbol{\beta})]^{-1}$. Fisherova informační matice je definovaná jako:

$$\mathbf{J}(\boldsymbol{\beta}) = -E\mathbf{H}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right), i,j = 1,\ldots,m. \tag{2.16}$$

V praxi ovšem skutečnou hodnotu neznáme a hledáme pouze její odhad $\mathbf{J}(\hat{\boldsymbol{\beta}})$. Postupujeme tak, že nejprve nalezneme mínus matici druhých parciálních derivací logaritmické věrohodnostní funkce podle proměnných, tedy mínus Hessovu matici, kterou vyčíslíme pro $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Prvky Fisherovy informační matice jsou dány jako:

$$J_{jk}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} x_{ij}x_{ik}\left(\frac{e^{\hat{\boldsymbol{\beta}}'\mathbf{x}_i}}{1+e^{\hat{\boldsymbol{\beta}}'\mathbf{x}_i}}\right)\left(\frac{1}{1+e^{\hat{\boldsymbol{\beta}}'\mathbf{x}_i}}\right), j,k = 0,\ldots,m. \tag{2.17}$$

2.2. Iterační metody výpočtu koeficientů logistického modelu

V této podkapitole si představíme některé iterační metody sloužící k výpočtu koeficientů modelu logistické regrese ze systému věrohodnostních rovnic, které jsme si dříve vy-

jádřili ve tvaru (2.12). Využití iteračních algoritmů je nezbytné, neboť tyto rovnice nejsou lineární vzhledem k hledaným parametrům. Ze známých algoritmů můžeme jmenovat například Newtonovu-Raphosonovu metodu, Böhningovu metodu, iterativní vážení nebo modifikované iterativní vážení.[16] V této práci si blíže představíme Newtonovu-Raphosonovu metodu a iterační váženou metodu nejmenších čtverců.

2.2.1. Newtonova-Raphosonova metoda

Nejprve si popišme, jak tato metoda funguje a jak je definována. Mějme tedy obecně soustavu m nelineárních rovnic o m neznámých:

$$f_1(x_1, \ldots, x_m) = 0$$

$$\vdots$$

$$f_m(x_1, \ldots, x_m) = 0.$$

Tuto soustavu můžeme zapsat vektorově jako:

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \mathbb{R}^m, \mathbf{0} = (0, \ldots, 0)' \in \mathbb{R}^m. \tag{2.18}$$

Řešením tohoto systému nelineárních rovnic označme vektor $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)'$, bude tedy platit $\mathbf{F}(\boldsymbol{\xi}) = \mathbf{0}$.

Před samotnou definicí Newtonovy-Raphsonovy metody je třeba definovat *Jacobiovu matici* funkce $\mathbf{F}$, kterou označujeme $\mathbf{J}_\mathbf{F}(\mathbf{x})$ a je ve tvaru:

$$\mathbf{J}_\mathbf{F}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_m} \end{pmatrix}.$$

Nechť je matice $\mathbf{J}_\mathbf{F}(\mathbf{x})$ regulární ($\det(\mathbf{J}_\mathbf{F}(\mathbf{x})) \neq 0$ a existuje inverzní matice $\mathbf{J}_\mathbf{F}^{-1}(\mathbf{x})$). A dále nechť má matice $\mathbf{J}_\mathbf{F}(\mathbf{x})$ spojité prvky v okolí bodu $\boldsymbol{\xi}$ (tento předpoklad je důležitý pro existenci jednotlivých iterací). Využitím Taylorova rozvoje funkce $\mathbf{F}(\mathbf{x})$ v okolí bodu $\mathbf{x}^{(0)}$ (kde $\mathbf{x}^{(0)}$ je počáteční aproximace) dostaneme:

$$\mathbf{F}(\mathbf{x}) \approx \mathbf{F}(\mathbf{x}^0) + \mathbf{J}_\mathbf{F}(\mathbf{x}^0)(\mathbf{x} - \mathbf{x}^0) \overbrace{=}^{\mathbf{F}(\mathbf{x})=\mathbf{0}} \mathbf{0} \Rightarrow \mathbf{x} \approx \mathbf{x}^0 - \mathbf{J}_\mathbf{F}^{-1}(\mathbf{x}^0)\mathbf{F}(\mathbf{x}^0). \tag{2.19}$$

Potom metodu

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \mathbf{J}_{\mathbf{F}}^{-1}(\mathbf{x}^k)\mathbf{F}(\mathbf{x}^k), \quad k = 0, 1, 2, \ldots, \tag{2.20}$$

kde $\mathbf{x}^k$ je k-tá iterace maximálně věrohodného odhadu ξ (index k udává pořadí iterace), nazýváme Newtonovou-Raphsonovou metodou pro řešení nelineárního systému rovnic $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. O podmínkách konvergence posloupnosti $\left\{\mathbf{x}^k\right\}_{k=0}^{\infty}$ k řešení $\boldsymbol{\xi}$ pojednává následující věta.

Věta 2.1 *Nechť $\boldsymbol{\xi}$ je kořenem rovnice $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$. Nechť $\boldsymbol{J_F}(\boldsymbol{x})$ je regulární matice se spojitými prvky v okolí $O(\boldsymbol{\xi})$ bodu $\boldsymbol{\xi}$, přičemž*

$$\left\|\boldsymbol{J}_{\boldsymbol{F}}^{-1}(\boldsymbol{x})\right\|_{\infty} \leq K, \quad K = konst.,$$

pro všechna $\boldsymbol{x}$ z tohoto okolí. Nechť funkce f_i,$i = 1, \ldots, k$, mají spojité druhé parciální derivace v okolí $O(\boldsymbol{\xi})$ bodu $\boldsymbol{\xi}$.

Posloupnost $\left\{\boldsymbol{x}^k\right\}_{k=0}^{\infty}$ určená určená Newtonovou-Raphosonovou metodou konverguje ke kořenu $\boldsymbol{\xi}$ za předpokladu, že počáteční aproximace $\boldsymbol{x}^0$ leží dostatečně blízko bodu $\boldsymbol{\xi}$.[11]

Výrazem $\left\|\mathbf{J}_{\mathbf{F}}^{-1}(\mathbf{x})\right\|_{\infty}$ chápeme řádkovou normu matice $\mathbf{J}_{\mathbf{F}}^{-1}(\mathbf{x})$, kterou získáme tak, že po řádcích sečteme absolutní hodnoty prvků matice a z nich vybereme maximum.

Nyní můžeme tyto znalosti aplikovat na náš systém věrohodnostních rovnic, který máme ve tvaru (2.14). Označme si tedy úlohou jako:

$$\mathbf{V}(\boldsymbol{\beta}) = \mathbf{0}. \tag{2.21}$$

Jacobiova matice $\mathbf{J}_{\mathbf{V}}(\boldsymbol{\beta})$ funkce $\mathbf{V}$ je ve tvaru:

$$\mathbf{J}_{\mathbf{V}}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial V_0(\boldsymbol{\beta})}{\partial \beta_0} & \cdots & \frac{\partial V_0(\boldsymbol{\beta})}{\partial \beta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial V_m(\boldsymbol{\beta})}{\partial \beta_0} & \cdots & \frac{\partial V_m(\boldsymbol{\beta})}{\partial \beta_m} \end{pmatrix} = \mathbf{V}'(\boldsymbol{\beta}).$$

V našem případě je Jacobiho matice $\mathbf{J}_{\mathbf{V}}(\boldsymbol{\beta})$ rovna Hessově matici $\mathbf{H}(\boldsymbol{\beta})$ a také mínus Fisherově informační matici $\mathbf{J}(\boldsymbol{\beta})$, tedy:

$$\mathbf{J}_{\mathbf{V}}(\boldsymbol{\beta}) = \mathbf{H}(\boldsymbol{\beta}) = -\mathbf{J}(\boldsymbol{\beta}) \tag{2.22}$$

Dále je třeba ověřit podmínku regularity matice $\mathbf{J_V}(\boldsymbol{\beta})$ z věty 2.1. Protože je Fisherova informační matice $\mathbf{J}(\boldsymbol{\beta})$ pozitivně definitní, je matice $\mathbf{J_V}(\boldsymbol{\beta})$ negativně definitní. Podle Sylvestrova kritéria platí pro determinanty negativně definitní matice následující pravidlo: znaménka determinantů $D_0, \ldots, D_m$ se střídají počínaje záporným, tedy $D_0 < 0$. Podmínku regularity Jacobiovy matice ($\det(\mathbf{J_V}(\boldsymbol{\beta})) \neq 0$) máme splněnou.

Poslední podmínkou konvergence posloupnosti $\left\{\boldsymbol{\beta}^k\right\}_{k=0}^{\infty}$ k řešení $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta_m})'$ je volba počáteční aproximace $\boldsymbol{\beta}^{(0)}$, která musí být dostatečně blízko řešení $\hat{\boldsymbol{\beta}}$. Počáteční aproximaci $\boldsymbol{\beta}^{(0)}$ lze uvažovat jako vektor počátečních aproximací β_k^0 pro jednotlivé rovnice $V_k, k = 0, \ldots, m$. Tyto počáteční aproximace lze odhadnout například pomocí dalších iteračních metod viz [7].

Nechť je $\boldsymbol{\beta}^0$ počáteční aproximace a $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_0)'$ je řešením systému nelineárních rovnic $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{0}$. Potom jsou splněny předpoklady věty 2.1 a iterační rovnice pro nalezení odhadu parametrů $\boldsymbol{\beta}$ pomocí Newtonovy-Raphosonovy metody je ve tvaru:

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \mathbf{J}_{\mathbf{V}}^{-1}(\boldsymbol{\beta}^k)\mathbf{V}(\boldsymbol{\beta}^k). \tag{2.23}$$

Tento iterační algoritmus poté konverguje k řešení $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)'$. Algoritmus probíhá dokud není splněna podmínka $\left\|\boldsymbol{\beta}^N - \boldsymbol{\beta}^{N-1}\right\| < \epsilon$, kde ϵ je předem určená malá konstanta (například $\epsilon = 10^{-8}$). Jinak řečeno, požadujeme, aby vektorová norma vektoru rozdílů dvou po sobě jdoucích iterací byla menší než zadaná (malá) konstanta. Poté prohlásíme poslední provedenou iteraci - obecně $\boldsymbol{\beta}^N$ (po N krocích) numerickým řešením systému nelineárních rovnic $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{0}$.

Další iterační metodou pro odhad regresních koeficientů u modelu logistické regrese je tzv. *Fisherova skórovací metoda*, která se v obecném případě od Newtonovy-Raphsonovy metody liší, nicméně v případě logistické regrese dávají obě metody totožné odhady. Iterační rovnice pro nalezení odhadů regresních koeficientů $\boldsymbol{\beta}$ pomocí Fisherovy skórovací metody je v následujícím tvaru:

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + \mathbf{J}^{-1}(\boldsymbol{\beta}^k)\mathbf{V}(\boldsymbol{\beta}^k). \tag{2.24}$$

Tato metoda tedy využívá k výpočtu odhadů regresních koeficientů Fisherovu informační matici $\mathbf{J}(\boldsymbol{\beta})$. V případě logistické regrese je však tato matice rovna mínus Hessově matici a tedy také mínus Jacobiho matici ($\mathbf{J}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{J_V}(\boldsymbol{\beta})$). Po dosazení mínus Hessovy

matice (resp. mínus Jacobiho matice) do rovnice (2.24) dostáváme Newtonovu-Raphsonovu metodu, kterou máme ve tvaru (2.23). [1]

Alternativou k Newtonově-Raphsonově metodě je *iterativní vážená metoda nejmenších čtverců*, kterou získáme modifikací právě výše posané metody. Abychom mohli provést potřebné úpravy, je třeba si vyjádřit Newtonovu-Raphsonovu metodu v maticovém tvaru. Objevuje se nám zde Fisherova informační matice $\mathbf{J}(\boldsymbol{\beta})$ (resp. její inverze) a systém věrohodnostních rovnic.

Prvky Fisherovy matice máme ve tvaru:

$$\mathbf{J}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij}x_{ik}\left(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)\left(\frac{1}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right), j,k = 0,\ldots,m.$$

Maticový zápis je s využitím matice $\mathbf{U}$ zavedenou v (2.15) následovný:

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{U}(\boldsymbol{\beta})\mathbf{X}, \tag{2.25}$$

Systém věrohodnostních rovnic jsme označili jako $\mathbf{V}$ a je ve tvaru:

$$\sum_{i=1}^{n}\left[y_i - \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right]x_{ij} = 0, \quad j = 0,\ldots,m.$$

Opět je zde člen $\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}} = \pi_i(\boldsymbol{\beta})$. Označme si jako $\boldsymbol{\pi}(\boldsymbol{\beta})$ n-rozměrný vektor podmíněných pravděpodobností $\pi_i(\boldsymbol{\beta})$. Maticový zápis systému věrohodnostních rovnic bude poté ve tvaru:

$$\mathbf{V}(\boldsymbol{\beta}) = \mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\beta})) = 0. \tag{2.26}$$

Nyní si můžeme vyjádřit Newtonovu-Raphsonovu metodu v maticovém tvaru:

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \left[\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\beta}^k)\right). \tag{2.27}$$

Následně provedeme několik úprav. Nejdříve člen $\boldsymbol{\beta}^k$ na pravé straně vynásobíme Fisherovou maticí i její inverzní maticí, poté vztah formálně upravíme a dostáváme:

$$\boldsymbol{\beta}^{k+1} = \left[\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{X}\boldsymbol{\beta}^k - \left[\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\beta}^k)\right) =$$

$$= \left[\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\left[\mathbf{X}\boldsymbol{\beta}^k + \mathbf{U}^{-1}(\boldsymbol{\beta}^k)\left(\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\beta}^k)\right)\right] =$$

$$= \left[\mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{X}\right]^{-1} \mathbf{X}'\mathbf{U}(\boldsymbol{\beta}^k)\mathbf{Z}(\boldsymbol{\beta}^k),$$

kde $\mathbf{Z}(\boldsymbol{\beta}^k) = \mathbf{X}\boldsymbol{\beta}^k + \mathbf{U}^{-1}(\boldsymbol{\beta}^k)\left(\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\beta}^k)\right)$. Tento tvar je podobný tvaru výpočetního vztahu pro odhad metodou vážených nejmenších čtverců u lineární regrese, odtud název metody výpočtu koeficientů.

Kapitola 3

Výběr vhodného modelu

Nyní již víme, jak vytvořit model a odhadnout koeficienty tohoto modelu. Pokud máme více vysvětlovaných proměnných, nastává otázka, které z nich do modelu zahrnout, a které vynechat. Právě této problematice se věnuje následující kapitola.

3.1. Test poměrem věrohodností

Cílem tvorby regresního modelu je vždy sestavit takový model, který bude dobře vysvětlovat pozorovaná data a zároveň bude obsahovat co nejméně proměnných.

U lineární regrese se pro posouzení kvality modelu využívá reziduálního součtu čtverců, u logistické regrese využíváme *věrohodnostní poměr*, neboli tzv. *devianci*. Než si testovou statistiku představíme, zavedeme si několik pojmů.

- Saturovaný model - neboli maximální model, obsahuje všechny proměnné, hodnotu jeho logaritmické věrohodnostní funkce, označme l_S
- Nulový model - naopak uvažuje pouze absolutní člen, tedy konstantu, hodnotu jeho logaritmické věrohodnostní funkce označme l_0
- Zkoumaný model - obsahuje námi uvažovaných p proměnných a konstantu, tedy celkem $p+1$ parametrů, označme l_M hodnotu jeho logaritmické věrohodnostní funkce

Deviance se zavádí jako dvojnásobek rozdílu mezi hodnotou logaritmu věrohodnostní funkce saturovaného modelu a hodnotou logaritmu věrohodnostní funkce námi uvažovaného modelu, značíme D [1, 2, 23], tedy

$$D_M = 2(l_S - l_M). \tag{3.1}$$

Interpretace deviancí je obdobná jako u reziduálního součtu čtverců v lineární regresi. Platí totiž, že čím je hodnota deviance větší, tím hůře model popisuje naše data. Saturovaný model má hodnotu deviance nula.

Problémem samotných deviancí je upřednostňování modelů s více parametry. Pokud bychom tedy srovnávali pouze deviance danných modelů, zjistili bychom, že nejlepší model je saturovaný, což ovšem z praktického hlediska není.

Deviance nyní využijeme pro zkonstruování *testu poměrem věrohodností*, který budeme značit ΔD. [2, 18] Uvažujme tedy dva modely popisující naše data. Model M_1 s p parametry a model M_2 s $p-1$ parametry, který vznikl vynecháním j-tého regresoru. Testem poměrem věrohodností nyní můžeme testovat hypotézu:

$$H_0 : \beta_j = 0 \;\text{ vs. }\; H_1 : \beta_j \neq 0.$$

Testujeme tedy významnost j-té vysvětlující proměnné v modelu. K testování použijeme testovou statistiku:

$$\Delta D_{M_1 M_2} = D_{M_2} - D_{M_1} = 2(l_{M_1} - l_{M_2}) \quad \dot{\sim} \quad \chi^2(1). \tag{3.2}$$

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, jestliže $\Delta D_{M_1 M_2} \geq \chi^2_{1-\alpha}(1)$, kde $\chi^2_{1-\alpha}(1)$ je $(1-\alpha)$-kvantil χ^2-rozdělení o 1 stupni volnosti. Počet stupňů volnosti 1 odpovídá rozdílu počtu parametrů jednotlivých modelů, tedy: $p-(p-1)=1$. Jestliže zamítneme nulovou hypotézu, znamená to, že j-tý regresor je statisticky významný a nelze přistoupit k jednoduššímu modelu pouze s $p-1$ parametry.

Obecně lze tuto metodu využít pro testování hypotéz o podmodelech. Uvažujme model M_1 s k_1 a model M_2 s k_2 parametry, kde $k_1 > k_2$. Nulová hypotéza je ve tvaru:

$$H_0 : \text{Model } M_2 \text{ je podmodelem modelu } M_1.$$

Testujeme pomocí statistiky (3.2), která má v tomto případě asymptoticky χ^2 rozdělení o k_1-k_2 stupních volnosti. Zamítnutí nulové hypotézy ($\Delta D_{M_1 M_2} \geq \chi^2_{1-\alpha}(k_1-k_2)$) znamená, že nelze přistoupit od složitějšího modelu M_1 k jednoduššímu modelu M_2, neboť ten data příliš zjednodušuje a hrozí ztráta informace.

Pro zjištění, zda jsme v našem zkoumaném modelu nevynechali nějakou významnou proměnnou, můžeme srovnat náš model se saturovaným. Testujeme hypotézu, zda je náš zkoumaný model podmodelem saturovaného. Zamítnutí nulové hypotézy znamená, že v našem modelu skutečně chybí nějaká statisticky významná proměnná. Na stranu druhou můžeme chtít zjistit, zda se náš model významně liší od pouhé konstanty, v tomto případě budeme srovnávat náš zkoumaný model s nulovým.

3.2. Koeficienty determinace

Koeficienty determinace jsou výledkem snahy rozšířit index determinace z lineární regrese též do logistické regrese, a to v návaznosti na značnou podobnost deviancí a reziduálního součtu čtverců. Tyto koeficienty můžeme interpretovat jako míry snížení neurčitosti v datech, kterého se povedlo dosáhnout nalezeným modelem.[2, 18]

- McFaddenův koeficient:

$$R^2_{MF} = 1 - \frac{l_M}{l_0}.$$

Připomeňme, že l_M je hodnota logaritmické věrohodnostní funkce pro zkoumaný model a l_0 pro model nulový. $R^2_{MF} \in \langle 0, 1\rangle$ a platí, že čím blíže je tento koeficient roven jedné, tím je model lepší. V praxi lze McFaddenův koeficient využít k porovnání dvou modelů, přičemž lepší model má vyšší hodnotu koeficientu. Nevýhodou této statistiky je nadhodnocování modelu s vyšším počtem pozorování. McFaddenův koeficient poté relativně snadno dosahuje hodnot blízkých 1 a model nadhodnocuje. Tento problém se snaží vyřešit následující statistika.

- Coxové-Snellův koeficient:

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}.$$

Povšimněme si, že na rozdíl od předchozího koeficientu, se zde objevují hodnoty L_M (hodnota věrohodnostní funkce pro zkoumaný model) a L_0 (hodnota věrohodnostní funkce pro nulový model). Ovšem i tato statistika má svoji nevýhodu. Nedosahuje totiž maximální hodnoty jedna, ale hodnoty $1 - [L_0]^{\frac{2}{n}}$, což je menší než jedna a výrazně znesnadňuje interpretaci. Toto maximum však závisí pouze na hodnotě π a to následujícím způsobem [3]:

$$\max = 1 - \left[\pi^{\pi}\left(1-\pi\right)^{(1-\pi)}\right]^2.$$

Můžeme si tedy snadno vypočítat, že pro $\pi = 0.5$ je maximální hodnota koeficientu R^2_{CS} 0.75. Dále například pro hodnoty $\pi = 0.9$, resp. $\pi = 0.1$ je tato hodnota 0.48. Představme si ještě upravený Coxové-Snellův koeficient, též nazývaný Nagelkerkův koeficient, jehož hodnoty leží v intervalu $\langle 0, 1\rangle$.

- Nagelkerkův koeficient:

$$R_N^2 = \frac{1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}}{1 - (L_0)^{\frac{2}{n}}}.$$

Vidíme, že se jedná o původní koeficient R_{CS}^2 vydělený jeho maximální hodnotou. Společnou nevýhodou všech těchto koeficientů je zvýhodňování modelů s větším počtem parametrů a upřednostňování složitějších modelů, stejně jako u deviancí samotných. Je tedy třeba najít upravené statistiky, které budou nějakým způsobem penalizovat počet parametrů v modelu.

3.3. Informační kritéria

Informační kritéria jsou modifikací deviancí. Opět poskytují pouze možnost vzájemně porovnat modely, ale neposkytují informaci o kvalitě modelu samotného. V praxi předpokládáme, že při vytvoření jakéhokoli modelu dochází ke ztrátě informace obsažené v datovém souboru. Naším cílem je vybrat takový model, který tuto ztrátu minimalizuje. Čím nižší hodnotou informačního kritéria model má, tím je pro nás vhodnější. Nejčastěji užívaným je tzv. *Akaikeho informační kritérium*, definované jako

$$AIC = -2l_M + 2p, \tag{3.3}$$

kde p je počet parametrů v modelu. Označme AIC_{min} minimální hodnotu mezi dalšími hodnotami $AIC_1, \ldots, AIC_s$, kde s je počet uvažovaných (zkoumaných) modelů, potom můžeme vyjádřit pravděpodobnost, že i-tý model minimalizuje ztrátu informace jako

$$exp\left(\frac{AIC_{min} - AIC_i}{2}\right). \tag{3.4}$$

Dalším typem informačního kritéria je *Bayesovo informační kritérium*, které je dáno vztahem

$$BIC = -2l_M + p\ln(n), \tag{3.5}$$

kde n je počet pozorování. V porovnání s AIC Bayesovo kritérium silněji penalizuje počet odhadovaných parametrů. Tyto dvě informační kritéria rozhodně nejsou jedinná, nicméně pro účely této práce postačí.

3.4. Waldův test

Při rozhodování, které vysvětlující proměnné je vhodné zařadit do modelu, nám může pomoci i *Waldův test.*[1, 2] Stejně jako testem poměrem věrohodností testujeme hypotézu:

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0.$$

Waldův test je asymptotický, je tedy možné jej použít pouze pro dostatečně velký počet pozorování. Za platnosti nulového hypotézy má Waldova statistika asymptoticky normální normované rozdělení, tj.:

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{var}(\hat{\beta}_j)}} \quad \dot{\sim} \quad N(0,1), \tag{3.6}$$

kde $\sqrt{\widehat{var}(\hat{\beta}_j)}$ je odhad směrodatné chyby pro příslušný odhad parametru, který získáme z odhadnuté Fisherovy informační matice. Platí $\sqrt{\widehat{var}(\hat{\beta}_j)} = \sqrt{\left(J(\hat{\boldsymbol{\beta}})^{-1}\right)_{jj}}$. Jedná se tedy o analogii pro t-test u lineární regrese. Pokud je však absolutní hodnota odhadu regresního koeficientu velká a je též velká hodnota příslušného odhadu směrodatné chyby, pak je hodnota Waldovy statistiky malá a vede k selhání zamítnutí nulové hypotézy, v takovém případě se použití Waldovy statistiky nedoporučuje.[14]Poznamenejme, že lze použít i statistiku Z_j^2, která má přibližně χ^2 rozdělení s jedním stupněm volnosti.

Pomocí Waldova testu lze testovat též hypotézu $H_0 : \beta_j = 0$ pro $j = 0, \ldots, m$, tedy že všech $m + 1$ koeficientů je rovno nule, případně s vyloučením koeficientu β_0 hypotézu $H_0 : \beta_j = 0$ pro $j = 1, \ldots, m$. V tomto případě bude Waldova statistika v následujícím tvaru:

$$Z = \hat{\boldsymbol{\beta}}' \left[\widehat{var}(\hat{\boldsymbol{\beta}})\right]^{-1} \hat{\boldsymbol{\beta}} \quad \dot{\sim} \quad \chi^2(p), \tag{3.7}$$

kde $\left[\widehat{var}(\hat{\boldsymbol{\beta}})\right]^{-1}$ je odhad Fisherovy inverzní matice, tedy matice $[\mathbf{J}(\hat{\boldsymbol{\beta}})]^{-1}$ a p je počet testovaných parametrů β.

Dodejme, že pro velké rozsahy dává Waldova statistika i test poměrem věrohodností přibližně stejné výsledky. Naopak při malém rozsahu pozorování se tyto výsledky mohou lišit. Ve většině případů je doporučováno použít test poměrem věrohodností. Naopak velkou výhodou Waldova testu je výpočetní jednoduchost, stačí totiž spočítat odhady pouze pro jeden model.

Waldova statistika ve tvaru (3.6) nám umožňuje konstruovat přibližné intervaly spolehlivosti pro jednotlivé koeficienty modelu. Přibližný $100(1-\alpha)\%$ interval spolehlivosti pro β_j je ve tvaru

$$\left\langle \beta_j - u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_j)}; \beta_j + u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_j)} \right\rangle, \tag{3.8}$$

kde $u_{1-\frac{\alpha}{2}}$ je $1-\frac{\alpha}{2}$ kvantil standardizovaného normálního rozdělení. Někdy nás může zajímat též interval spolehlivosti pro poměr šancí. Ten vytvoříme díky vztahu $OR(X_j) = e^{\beta_j}, j = 1, \ldots, m$ z (3.8) užitím exponenciální funkce. Dostáváme tedy:

$$\left\langle \exp\left(\beta_j - u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_j)}\right); \exp\left(\beta_j + u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_j)}\right) \right\rangle. \tag{3.9}$$

Kapitola 4

Hodnocení modelu

V této části už předpokládejme, že jsme našli model. Zbývá nalézt odpověď na otázku, jak dobře tento model popisuje reálná data. Metod k nalezení odpovědi na tuto otázku je mnoho. My si v této kapitole představíme křížovou validaci, testy dobré shody a koeficienty hodnotící míru asociace mezi pozorovanými a očekávanými hodnotami. Z grafických nástrojů si představíme ROC křivku. Nejdříve se ale seznámíme s klasifikací, neboť s tímto pojmem souvisí všechny zmíněné metody.

4.1. Klasifikační tabulka

Předpokládáme, že jsme nalezli model M. Tento model na základě realizací $x_1 \ldots, x_m$ nezávislých proměnných $X_1 \ldots, X_m$ modeluje pravděpodobnost, že se závislá proměnná Y_i, $i = 1, \ldots, n$, realizuje hodnotou 1. Tato pravděpodobnost se nazývá *(logistické) skóre i-tého pozorování*. Značíme s_i a vztah pro jeho výpočet je následující:

$$s_i = \hat{\pi}_i(\hat{\boldsymbol{\beta}}) = \frac{e^{\hat{\boldsymbol{\beta}}'\mathbf{x}_i}}{1 + e^{\hat{\boldsymbol{\beta}}'\mathbf{x}_i}}, \quad i = 1, \ldots, n. \tag{4.1}$$

V ideálním případě by měla platit rovnost $y_i = s_i$, avšak tohoto ideálu nelze v praxi nikdy dosáhnout, neboť s_i značí pravděpodobnost a leží v intervalu $\langle 0, 1 \rangle$, zatímco realizace y_i nabývají hodnot $\{0, 1\}$. Snažíme se alespoň k tomuto ideálu co nejvíce přiblížit. Požadujeme tedy, aby pro většinu realizací $y_i = 1$ byly hodnoty s_i blízké 1, a naopak, aby pro většinu hodnot $y_i = 0$ byly hodnoty s_i blízké 0. Pro případ, kdy je $s_i \approx 0.5$ a není jasné, do které skupiny toto pozorování zařadit, se zavádí pojem *prahový bod*. Prahový bod se značí jako P_c (z anglického *cut-off point)*. Pokud bude hodnota $s_i < P_c$, pak toto pozorování budeme považovat za neúspěch, naopak, pokud $s_i > P_c$, budeme jej považovat

za úspěch. Jak takovou hodnotu zvolit si uvedeme později. Nyní předpokládejme, že jsme tento prahový bod stanovili.

Na základě vytvořeného modelu a stanoveného prahového bodu jsme schopni získat očekávané hodnoty pro daná pozorování. Z pozorovaných hodnot $y_1, \ldots, y_n$ a jim odpovídajících očekávaných hodnot $\hat{y}_1, \ldots, \hat{y}_n$ můžeme vytvořit tzv. *klasifikační tabulku* viz tabulka 4.1, která slouží k vyhodnocení úspěšnosti predikce modelu.

	0 (Skutečnost)	**1 (Skutečnost)**
0 (Predikce)	Skutečně negativní (TN)	Falešně negativní (FN)
1 (Predikce)	Falešně pozitivní (FP)	Skutečně pozitivní (TP)

Tabulka 4.1: Klasifikační tabulka

Zkratky použité v tabulce pochází z anglických překladů (TN - True Negative, FP - False Positive, FN - False Negative a TP - True Positive). Z této matice můžeme určit celkovou přesnost (správnost) našeho modelu jako relativní četnost správných predikcí, tedy:

$$\text{Přesnost} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{TN}}. \tag{4.2}$$

Z této tabulky můžeme vypočítat též tzv. *pozitivní a negativní prediktivní hodnoty.*

- *Prediktivní hodnota pozitivního testu* je pravděpodobnost, že je pozorování skutečně úspěšné, jestliže bylo modelem klasifikováno jako úspěšné. Výpočetní vztah je ve tvaru:

$$\text{Prediktivní hodnota pozitivního testu} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

- *Prediktivní hodnota negativního testu* je pravděpodobnost, že je pozorování skutečně neúspěšné, jestliže bylo modelem klasifikováno jako neúspěšné. Výpočetní vztah je ve tvaru:

$$\text{Prediktivní hodnota negativního testu} = \frac{\text{TN}}{\text{FN} + \text{TN}}.$$

V souvislosti s klasifikační maticí si ještě představme *McNemarův test* [19]. Nulová hypotéza přitom říká, že procento skutečně pozitivních je rovno procentu očekávaných pozitivních. Testová statistika je ve tvaru:

$$\text{T} = \frac{(\text{FP} - \text{FN})^2}{\text{FP} + \text{FN}} \quad \dot{\sim} \quad \chi^2(1). \tag{4.3}$$

Nulovou hypotézu zamítáme, jestliže $T \geq \chi^2_{1-\alpha}(1)$. Protože jde o asymtotický test, musí být splněna podmínka četnosti, která je ve tvaru: $\mathrm{FP} + \mathrm{FN} \geq 8$.

Statistika zohledňující kromě vyčíslení shody v procentech i pravděpodobnost náhodné shody se nazývá *Cohenovo kappa.*[19] Tuto statistiku značíme κ a je dána vztahem:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \tag{4.4}$$

Kde $p_o = \frac{\mathrm{TP}+\mathrm{TN}}{n}$ je pravděpodobnost shody pozorovaných a očekávaných hodnot a $p_e = \frac{\mathrm{TP}+\mathrm{FP}}{n} \cdot \frac{\mathrm{TP}+\mathrm{FN}}{n} + \frac{\mathrm{FP}+\mathrm{TN}}{n} \cdot \frac{\mathrm{FN}+\mathrm{TN}}{n}$ je odhad pravděpodobnosti náhodné shody a n je celkový počet pozorování. V případě, kdy je $\kappa = 0$ mluvíme o velmi nízké shodě pozorovaných a očekávaných hodnot a celkové procento shody je rovno pravděpodobnosti shody při náhodném rozhodování. Naopak pokud je $\kappa = 1$, mluvíme o dokonalé shodě a model vyhodnotil všechna pozorování správně.

4.2. Křížová validace

Jestliže chceme znát přesnost predikce pro nová data, máme několik možností jak postupovat. Jednou z možností je rozdělit původní datovou sadu na trénovací a testovací část, přičemž trénovací by měla být větší než testovací. Využitím první sady poté vytvoříme model a ten aplikujeme na sadu druhou. V této fázi můžeme vytvořit klasifikační matici a určit přesnost modelu.

Nevýhod u této metody je hned několik. Například, jak rozdělit data? Jak velké datové sady vytvořit? Popřípadě se může stát, že budeme mít dat málo. Toto elegantně řeší metoda *křížové validace*.

Křížová validace funguje obdobně, až na to, že rozdělí celý datový soubor na obecně $k \geq 2$ menších datových souborů, které jsou navzájem disjunktní. Poté vždy jeden z menších souborů považujeme za testovací a ostatní soubory za trénovací skupinu. Tento proces se opakuje a vždy je za testovací soubor vybrána jiná skupina dat. Výsledkem křížové validace je celková přesnost vypočítaná jako průměr ze všech k přesností získaných z jednotlivých opakování procesu. Obvykle se volí $k = 10$. Existuje ale i speciální případ křížové validace, a to tzv. *leave-one-out*, kdy je k rovno počtu pozorování a za testovací soubor se považuje vždy právě jedno pozorování. Tento způsob je však vhodný spíše pro menší datové soubory právě kvůli jeho náročnosti.

4.3. Testy dobré shody

Testy dobré shody využijeme k ověření shody mezi očekávanými a skutečnými hodnotami. Skutečné hodnoty jsou v našem případě realizace $y_1, \ldots, y_n$ nezávislé náhodné veličiny Y. Jako očekávané hodnoty vezměme modelem predikované pravděpodobnosti $\hat{\pi}_i(\hat{\boldsymbol{\beta}})$, tedy logistická skóre s_i. Testujeme tedy hypotézu:

$$H_0 : y_i = s_i \;\text{ vs. }\; H_1 : y_i \neq s_i, \quad i = 1, \ldots, n.$$

K ověření této hypotézy slouží následující testy [9]:

- Pearsonův χ^2 test

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - s_i)^2}{s_i(1 - s_i)} \quad \dot{\sim} \quad \chi^2(n - k - 1).$$

Nulovou hypotézu zamítáme jestliže $\chi^2 \geq \chi^2_{1-\alpha}(n-k-1)$, kde k je počet regresorů v modelu.

- Rozdílový test deviace

$$D = -2 \sum_{i=1}^{n} \left[s_i \ln\left(\frac{s_i}{1 - s_i}\right) + \ln(1 - s_i) \right] \quad \dot{\sim} \quad \chi^2(n - k - 1).$$

Nulovou hypotézu zamítáme jestliže $D \geq \chi^2_{1-\alpha}(n - k - 1)$. Hodnoty testovacích statistik nebudou ve většině případů shodné, avšak velké rozdíly mohou poukazovat na nevyhovující využití těchto metod.

- Hosmerovy-Lemeshowovy testy

Tyto testy jsou založeny na hodnotách $s_1, \ldots, s_n$ seřazených podle velikosti, které poté dělíme do skupin. Právě podle způsobu vytváření skupin rozlišujeme několik metod.

První metoda je založena na vytvoření celkem G skupin (obvykle se volí $G = 10$), kdy první skupina obsahuje $n_1 = \frac{n}{G}$ nejmenších hodnot, až postupně vytvoříme i poslední skupinu obsahující $n_g = \frac{n}{G}$ největších skóre. Očekává se, že případy značící neúspěch budou v prvních skupinách a naopak, případy značící úspěch v posledních. Spočítáme tedy pozorované četnosti úspěchů i neúspěchů v každé skupině g, $g = 1, 2, \ldots, G$:

$$o_{1g} = \sum_{i=1}^{n_g} y_i \qquad o_{0g} = \sum_{i=1}^{n_g} (1 - y_i)$$

a očekávané četnosti v každé skupině:

$$e_{1g} = \sum_{i=1}^{n_g} s_i \quad o_{0g} = \sum_{i=1}^{n_g} (1 - s_i),$$

kde n_g je počet pozorování ve skupině g.

Testovací statitika, kterou pro tento typ rozdělení označujeme C_g, je Pearsonovo χ^2 počítané z tabulky $(2 \times G)$ pozorovaných a očekávaných četností:

$$C_g = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{(o_{kg} - e_{kg})^2}{e_{kg}} \quad \dot{\sim} \quad \chi^2(G-2).$$

Asymptotické rozdělení je platné za podmínky $k < G$, kde k je počet parametrů v modelu. Nulovou hypotézu zamítáme, jestliže $C_g \geq \chi^2_{1-\alpha}(G-2)$.

Druhá metoda, kterou si zde představíme je založena na rozdělení skóre na celkem 10 skupin podle pevně daných prahových bodů (0,1;0,2;...;0,9). V tomto případě tedy není určen počet skóre v každé skupině. Víme ale, že v první skupině budou všechna skóre menší jak 0,1 a naopak v poslední skupině budou všechna větší jak 0,9. Testová statistika, označovaná pro tento typ rozdělení jako H_g, je ve stejném tvaru jako C_g. Toto byl pouze příklad nejpoužívanějších testů, existuje jich ale celá řada, například další Hosmerovy-Lemeshowovy testy, Brownův test, Stukelův test a další viz [9].

4.4. Kolmogorovův-Smirnovův test dobré shody

Narozdíl od testů dobré shody popsaných výše, je tato statistika založena na distribuční funkci skóre. Testujeme nulovou hypotézu, která tvrdí, že distribuční funkce pozitivních a negativních případů se shodují.[17] Mějme indikátorovou funkci ve tvaru:

$$\mathbf{I}(A) = \begin{cases} 1 & \text{jestliže A platí,} \\ 0 & \text{jinak.} \end{cases}$$

Jestliže si navíc označíme n_0 jako počet pozorování, kde $y_i = 0$ a n_1 jako počet pozorování, kde $y_i = 1$, pak jsou distribuční funkce pro skóre dány následovně:

$$F_0(x) = \frac{1}{n_0} \sum_{i=1}^{n} I(s_i < x, y_i = 0), x \in (0,1), \tag{4.5}$$

$$F_1(x) = \frac{1}{n_1} \sum_{i=1}^{n} I(s_i > x, y_i = 1), x \in (0, 1). \tag{4.6}$$

$F_0(x)$ představuje pravděpodobnost, že skóre s_i je menší než prahový bod x a že se zároveň závislá proměnná realizovala hodnotou 0, zatímco $F_1(x)$ představuje pravděpodobnost, že je hodnota skóre s_i větší nebo rovna prahovému bodu x a že se zároveň závislá proměnná realizovala hodnotou 1. Konečně testovací statistika je ve tvaru:

$$KS = \sup_{x \in (0;1)} |F_0(x) - F_1(x)| = \max_{x \in (0;1)} |F_0(x) - F_1(x)| .$$

Výslednou hodnotu statistiky KS porovnáváme s tabelovanou hodnotou $KS_{1-\alpha}(n_0, n_1)$, kterou získáme ze speciálních tabulek pro dvouvýběrový Kolmogorovův-Smirnovův test. Konkrétně pro hladinu $\alpha = 0.05$ je $KS_{1-0.05}(n_0, n_1) = 1.36\sqrt{\frac{n_0+n_1}{n_0 n_1}}$. Nulovou hypotézu o shodě distribučních funkcí $F_0(x)$ a $F_1(x)$ zamítáme na hladině významnosti α, jestliže $KS > KS_{1-\alpha}(n_0, n_1)$.

4.5. ROC křivka

Nejčastěji používaným grafickým nástrojem pro hodnocení kvality logistického regresního modelu je *ROC křivka* (anglicky *Receiver Operating Characteristic*). Abychom pochopili, co nám takovýto graf říká, vysvětleme si nejprve dva pojmy - *senzitivitu* a *specificitu*. Oba tyto pojmy souvisí s hodnocením predikčních schopností modelu. K výpočtu těchto hodnot využijeme hodnoty z klasifikační tabulky.

- *Senzitivita* vyjadřuje úspěšnost, s jakou test správně klasifikuje úspěšné případy. Jedná se tedy o pravděpodobnost, že model klasifikuje pozorování jako úspěšné, jeli skutečně úspěšné. Nabývá hodnot od 0 do 1. Výpočetní vztah je ve tvaru:

$$\text{Senzitivita} = \frac{\text{TP}}{\text{TP} + \text{FN}} = P(\hat{y}_i = 1 | y_i = 1).$$

  Jestliže by byla tato hodnota rovna jedné, pak by model všechny skutečně pozitivní případy opravdu vyhodnotil jako pozitivní.

- *Specificita* vyjadřuje úspěšnost, s jakou test správně klasifikuje neúspěšné případy. Jinak řečeno, specificita je pravděpodobnost, že model klasifikuje pozorování jako

negativní, jeli skutečně negativní. Stejně jako senzitivita nabývá hodnot od 0 do 1. Výpočetní vztah je ve tvaru:

$$\text{Specificita} = \frac{\text{TN}}{\text{TN} + \text{FP}} = P(\hat{y}_i = 0 | y_i = 0).$$

Pokud by byla specificita rovna jedné, pak by model všechny skutečně negativní případy vyhodnotil jako negativní.

V ideálním případě chceme, aby se senzitivita i specificita testu rovnala 1. Hodnoty těchto pravěpodobností jsou ovlivněny volbou prahového bodu. Pokud například hodnotu prahového bodu snížíme, zvýší se počet pozorování predikovaných jako pozitivní (dojde ke zvýšení senzitivity), ale také zvýšíme četnost falešně pozitivních pozorování (dojde ke snížení specificity). Prahový bod volíme tak, aby byla splněna námi požadovaná podmínka, která může být v jednom z následujících tvarů:

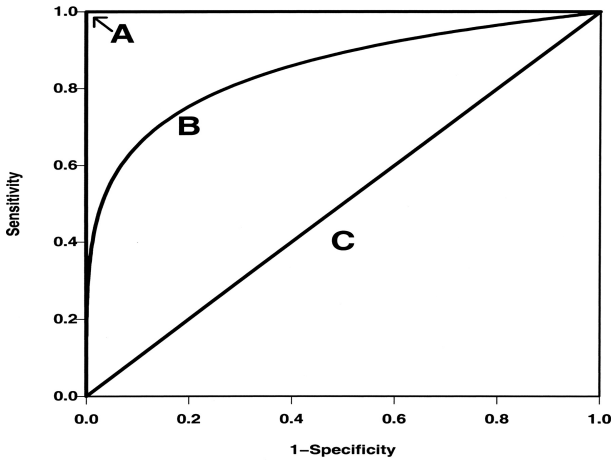
- Dosažení požadované hodnoty senzitivity modelu
- Dosažení požadované hodnoty specificity modelu
- Maximalizace součtu senzitivity a specificity testu

Nyní se vraťme k ROC křivce, kterou konstruujeme právě pomocí senzitivity a specificity. Jak už bylo řečeno, tyto hodnoty závisí na volbě prahového bodu, proto pro různé hodnoty P_c dostaneme různé hodnoty senizitivity a specificity, a tím také různě tvarované ROC křivky. ROC křivku definujeme jako množinu dvojic [4]:

$$\{(1 - F_0(x), F_1(x)) \, ; x = P_c \in (0, 1)\},$$

kde $F_0(x)$ a $F_1(x)$ jsou distribuční funkce z (4.5) a (4.6). Z toho, jak jsou tyto distribuční funkce zavedeny, můžeme nyní říct, že $F_0(x)$ pro prahový bod x představuje specificitu a $F_1(x)$ senzitivitu. ROC křivka je tedy graf závislosti senzitivity na 1-specificitě při daném prahovém bodu x.

Na obrázku 4.1 jsou ilustračně vyznačeny tvary tří křivek. Křivka A, která prochází body $[0; 0]$, $[0; 1]$ a $[1; 1]$ představuje křivku pro ideální model. Křivka B je křivka představující dobrý model a naopak křivka C - tedy úhlopříčka spojující body $[0; 0]$ a $[1; 1]$, představuje model náhodný. Model je tedy tím lepší, čím je křivka blíže ke křivce ideálního modelu.



Obrázek 4.1: Příklad ROC křivky [24]

S tímto grafem také souvisí pojem *AUC* z anglického *area under curve*, která číselně vyjadřuje plochu pod ROC křivkou. Hodnotu AUC můžeme interpretovat jako pravděpodobnost, že vybereme-li náhodné pozorování ze skupiny negativních pozorování, bude jeho očekávaná pravděpodobnost na úspěch menší, než u náhodně vybraného pozorování ze skupiny pozitivních pozorování. Z obrázku je zřejmé, že pro ideální model bude hodnota AUC = 1 a naopak pro model náhodný bude hodnota AUC = 0.5.

4.6. Míry asociace

Než si představíme samotné koeficienty hodnotící míru asociace mezi pozorovanými a očekávanými hodnotami, je třeba zavést několik pojmů a značení. Mějme tedy dvě pozorování. Pro první platí $y_i = 0$ a jeho skóre si označme s_0, pro druhé platí $y_i = 1$ a jeho skóre si označme s_1. Řekneme, že je pár pozorování:

- *Konkordantní* (ve shodě), pokud $s_1 > s_0$,
- *Diskordantní* (v neshodě), pokud $s_1 < s_0$,
- *Vázaný* (vyrovnaný), pokud $s_1 = s_0$ nebo $y_1 = y_0$.

Uvažujme rozsah souboru n, potom si označme jako S_k počet konkordantních párů v souboru, S_d počet diskordantních párů, $S_{v,s}$ počet vázaných párů v souboru s hodnotami skóre s_i, $S_{v,y}$ počet vázaných párů v souboru s pozorováními y_i a nakonec $S_{v,ys}$ počet

vázaných párů v obou souborech. Počet všech takových párů je pak roven

$$\frac{n(n-1)}{2} = S_k + S_d + S_{v,s} + S_{v,y} + S_{v,ys}.$$

Koeficienty založené na pořadové asociaci jsou ve tvaru:
Kendallovo τ_a:

$$\tau_a = \frac{S_k - S_d}{S_k + S_d + S_{v,s} + S_{v,y} + S_{v,ys}}, \tag{4.7}$$

Kendallovo τ_b:

$$\tau_b = \frac{S_k - S_d}{\sqrt{(S_k + S_d + S_{v,y})(S_k + S_d + S_{v,s})}}, \tag{4.8}$$

Somersovo D_s:

$$D_s = \frac{S_k - S_d}{S_k + S_d + S_{v,s}}, \tag{4.9}$$

Goodmanovo-Kruskalovo γ:

$$\gamma = \frac{S_k - S_d}{S_k + S_d}. \tag{4.10}$$

Hodnoty těchto koeficientů leží v intervalu $\langle -1, 1\rangle$. Jestliže jsou v souboru všechny páry konkordantní, koeficienty nabývají hodnoty 1. Jestliže jsou naopak v souboru všechny páry dikordantní, koeficienty nabývají hodnoty -1. Jestliže se budou hodnoty koeficientů pohybovat kolem 0, znamená to, že model špatně rozlišuje pozitivní a negativní případy.

Kendallovo τ_b se využívá při větším počtu nerozhodnutých párů. Somersovo D_s je modifikací τ_b a uvažuje pouze vazby mezi hodnotami skóre s_i. Goodmanova-Kruskalovo γ nebere v úvahu vázané páry, pokud by tedy v souboru nebyly diskordantní páry, bude jeho hodnota automaticky 1, přestože by byly přítomny páry vázané. [10]

Kapitola 5

Praktická část

Cílem práce je na základě reálného datového souboru analyzovat vliv hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodnost. K praktické analýze dat bude využit statistický software R. V případě využití jiných, než základních knihoven tohoto softwaru, bude uveden název odpovídajícího balíčku. V celé této práci budeme uvažovat hladinu významnosti $\alpha = 0.05$.

5.1. Popis datového souboru

Datový soubor obsahuje údaje o ženách léčených na neplodnost. Máme informace o celkem 157 ženách ve věku od 19 do 40 let. Kromě věku tato data obsahují údaje o hladině antimülleriánského hormonu, o hladině folikulostimulačního hormonu, o počtu jednotek tohoto hormonu dodaných v průběhu léčby, o počtu vajíček, a také o tom, který ze dvou lékařů měl danou ženu v péči. Dále máme informaci o úspěchu či neúspěchu léčby, tj. otěhotnění či neotěhotnění.

Antimülleriánský hormon (budeme používat zkratku AMH) je hormon, který je produkován ženskými granulózovými buňkami vaječníků, a který řídí dozrávání vajíček. Počet vajíček tak závisí právě na jeho hladině. Podle koncentrace tohoto hormonu v krvi lze poměrně přesně odhadnout, kolik vajíček ještě ženě zbývá. Proto může být tento hormon používán jako ukazatel toho, zda, popřípadě jak dlouho, je bezpečné odkládat těhotenství z hlediska nebezpečí snížení plodnosti, a jak velká je šance na přirozené otěhotnění.

Hladina hormonu AMH se mění v průběhu celého života ženy. Během menstruačního cyklu však hodnoty nijak zásadně nekolísají. K postupnému zvyšování hladiny dochází přibližně od 20. do 27. roku života, přičemž maximálních hodnot hladina hormonu dosahuje přibližně v 25. roku, kdy se pohybuje kolem 5 ng/ml. Zhruba od 27. roku začíná hladina hormonu AMH klesat až do období menopauzy. Ženy starší 40 let mívají hladinu hormonu

okolo 1 ng/ml a méně. Ideální hladina u ženy v reprodukčním věku je od 2 do 6.8 ng/ml. Z tohoto je patrné, že hlavním faktorem ovlivňujícím hladinu AMH je věk.[22]

V souvislosti s hladinou hormonu AMH je vhodné dodat několik slov o Syndromu polycystických ovárií (PCOS). Jedná se o jednu z nejčastěších endokrynních poruch u žen. Postihuje ženy v reprodukčním věku a bývá jednou z hlavních příčin neplodnosti. Mezi příznaky této nemoci patří zvýšená hladina hormonu AMH, která se může pohybovat od 2.41 ng/ml až do 17.1 ng/ml. [22]

Folikulostimulační hormon (budeme používat zkratku FSH), též folitropin je produkován buňkami hypofýzy a u žen podporuje růst vaječníkových váčků (folikulů), a také stimuluje tvorbu estrogenu. Hladina hormonu FSH se mění během menstruačního cyklu. Zvýšená hodnota FSH je též indikátorem pro blížící se přechod. Jelikož tento hormon příznivě ovlivňuje zrání vajíček, podává se léčebně před umělým oplodněním. Při zkoumání hladiny FSH je vhodné uvádět, ve které fázi se žena právě nacházela, neboť se tyto hodnoty mění. Při folikulární fázi se hodnoty u zdravé ženy mohou pohybovat mezi 3.85 – 8.78 IU/l, v preovulační fázi mezi 4.54-22.51 IU/l, luteální fázi mezi 1.79 – 5.12 IU/l a v menopauze mezi 16.74 – 113.59 IU/l. Hodnoty v našem datovém souboru se pohybují od 1.8 do 21.6, což může být zapříčiněno právě tím, že máme udáje od nemocných žen.[13]

5.2. Popisná statistika

V této podkapitole se blíže podíváme na jednotlivé proměnné. Vykreslíme si pro ně vybrané grafy a vypočítáme číselné charakteristiky. Nevynecháme ani graf znázorňující závislosti mezi veličinami.

Pro názornost je v tabulce 5.1 ukázka datového souboru, jedná se o prvních šest řádků. První sloupec obsahuje informace o věku žen, následují sloupce s hodnotami hladin hormonů FSH a AMH. Čtvrtý sloupec udává, kolik jednotek hormonů FSH bylo ženě při léčbě podáno. Pod zkratkou m2 se v pátém sloupci skrývá počet vajíček. V sloupci následujícím se pod označením asp (srdeční akce plodu) skrývá informace o výsledku léčby. Poslední sloupec nám dává informaci, ke kterému ze dvou lékařů žena docházela.

Nejdříve ze všeho je třeba zjistit, zda se v datovém souboru nevyskytují chybějící hodnoty. To provedeme příkazem `which(is.na(data))`, který nám vypíše čísla pozorování, u nichž chybí některá hodnota. V našem datovém souboru je celkem 21 chybějícíh hodnot, přičemž chybí 18 hodnot hladiny FSH a 3 údaje o tom, zda žena otěhotněla, či nikoli. Protože je počet chybějících údajů v porovnání s celkovým rozsahem souboru (celkem 157

vek	fsh	amh	jed	m2	asp	lekar
31	12.5	1.22	2250	4	1	1
35	21.6	0.40	3225	3	0	1
35	21.6	0.40	4050	2	0	1
35	21.6	0.40	4950	6	0	1
37	14.7	0.40	2625	2	0	1
37	14.7	0.40	3100	4	0	1

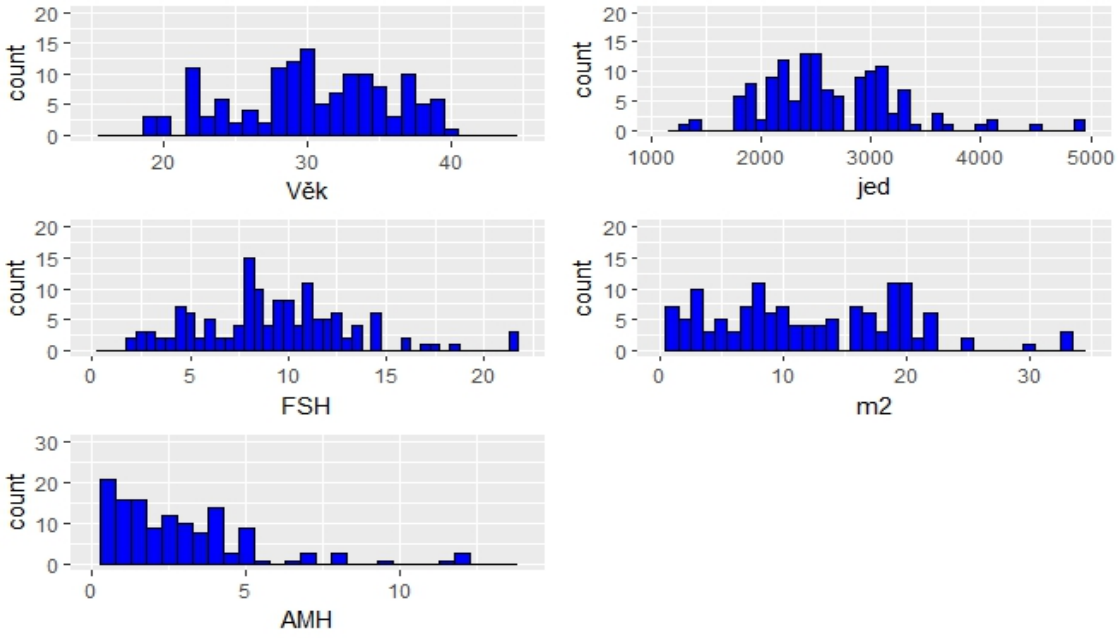
Tabulka 5.1: Ukázka dat

pozorování) zanedbatelný, vyřadíme jedince s chybějícími hodnotami z další analýzy.

Abychom získali prvotní představu o datovém souboru, je dobré vykreslit si některé grafy. My si nejprve vykreslíme histogramy, abychom zjistili, jak jsou proměnné rozloženy. Učiníme tak pro prvních pět proměnných. K jejich vytvoření využijeme knihovnu `ggplot2`. Na tomto místě uveďme příkaz pro vytvoření histogramu pro věk, příkazy k vytvoření ostatních histogramů se tvoří analogicky. Příkaz je ve tvaru:

```
>qplot(vek,geom="histogram", binwidth = 1, xlab = "Věk",
      fill=I("blue"), col=I("black"), xlim=c(15,45), ylim=c(0,20))
```

Jednotlivé histogramy následně spojíme do jednoho obrázku pomocí funkce `multiplot` [5].

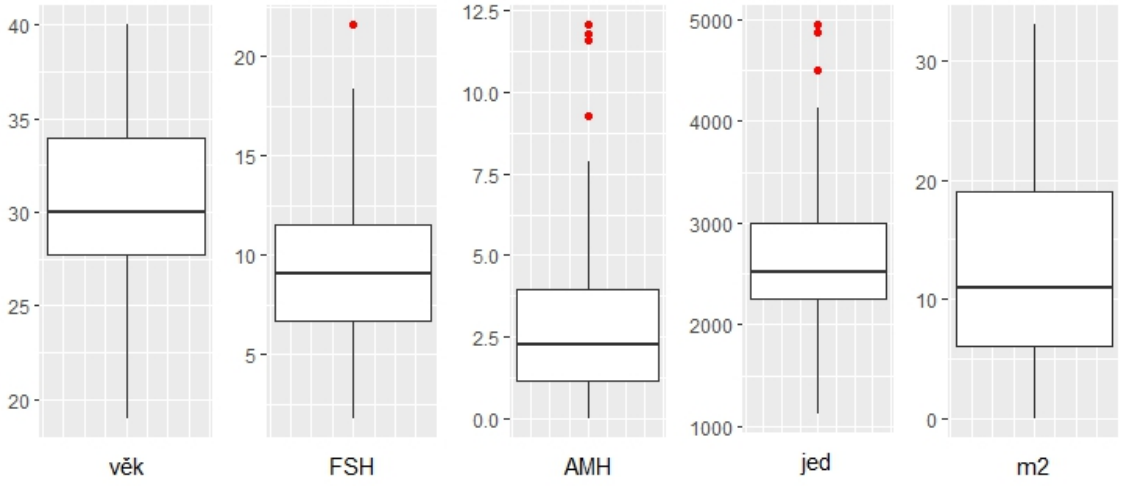


Obrázek 5.1: Histogramy

Z histogramů na obrázku 5.1 je vidět rozložení proměnných. Můžeme vyčíst četnosti pro jednotlivé hodnoty daných proměnných nebo jejich maximální a minimální hodnoty. Tím se dostáváme k charakteristikám polohy. Asi nejpoužívanějším grafickým nástrojem pro znázornění charakteristik polohy je krabicový graf (boxplot). Opět si je vykreslíme pro prvních pět proměnných za využití knihovny `ggplot2`. Příkaz pro vytvoření krabicového grafu pro proměnnou věk je ve tvaru:

```
>b1 <- ggplot(newdata,aes(y=vek,x=0))+geom_boxplot(outlier.colour = "red")
>b1= b1 + xlab("věk") +
  theme(axis.title.y = element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

Příkazy k vytvoření grafů pro zbývající proměnné se tvoří analogicky. Pro spojení jednotlivých grafů do jednoho obrázku opět využijeme funkce `multiplot`.



Obrázek 5.2: Krabicové grafy

Na obrázku 5.2 jsou červenými body vyznačeny odlehlé hodnoty, které můžeme pro jednotlivé proměnné přesně detekovat pomocí funkce `boxplot.stats("název proměnné")`. Zjištěné extrémní hodnoty zachycuje tabulka 5.2.

Po bližším prozkoumání hodnot jednotlivých proměnných není důvod předpokládat, že se jedná o chybné hodnoty. Hladina hormonu FSH může v některých případech dostahovat až hodnoty 113.59, hladina hormonu AMH u žen až 17.1 a detekované extrémní hodnoty přidávaných jednotek hormonů nejsou zásadně větší než jiné hodnoty, které jako extrémní nebyly označeny.

Proměnná	Extrémní hodnoty
FSH	21.6 21.6 21.6
AMH	11.78 11.78 9.27 11.56 12.05
jed	4950 4500 4875

Tabulka 5.2: Extrémní hodnoty

Když se ale ještě jednou vrátíme k obrázku 5.2, zjistíme, že z něj lze vyčíst přibližné hodnoty mediánu (50% kvantilu, značíme $\tilde{x}_{0,5}$), dolního (značíme $\tilde{x}_{0,25}$) i horního kvartilu (značíme $\tilde{x}_{0,75}$), ale i maximální nebo minimálního hodnoty. Přesné hodnoty můžeme zjistit příkazem `summary("název datového souboru")`. K těmto charakteristikám polohy patří také průměr (značíme $\bar{x}$) a modus (značíme $\hat{x}$), který nám udává nejčastěji se vyskytující hodnoty proměnné. Hodnotu průměru získáme pomocí stejného příkazu jako výše uvedené číselné charakteristiky polohy. Pro výpočet hodnoty modu není v základním balíčku softwaru R funkce, proto si ji musíme sami vytvořit. Jedna z možností, která vrátí nejen jeho hodnotu, ale i četnost s jakou se v souboru vyskytuje, je následující:

```
>Modus <- function(x){
  a = table(as.vector(x))
  return(a[which.max(a)])
}
```

Hodnoty všech zmíněných číselných charakteristik pro prvních pět proměnných shrnuje tabulka 5.3. Charakteristiky polohy nás informují o tom, jakých hodnot jednotlivé proměnné nabývají, kolem které hodnoty jednotlivé realizace kolísají nebo které hodnoty jsou nejčastější.

Proměnná	$\hat{x}$	$\bar{x}$	$\tilde{x}_{0,5}$	$\tilde{x}_{0,25}$	$\tilde{x}_{0,75}$	x_{min}	x_{max}
Věk	30.00	30.35	30.00	27.75	34.00	19.00	40.00
FSH	8.70	9.31	9.05	6.73	11.55	1.80	21.60
AMH	0.40	2.83	2.28	1.15	3.95	0.02	12.05
jed	2250	2644	2512	2250	3000	1125	4950
m2	20.00	11.96	11.00	6.00	19.00	0.00	33.00

Tabulka 5.3: Charakteristiky polohy

Pro lepší představu o datech používáme charakteristiky variability, které nám říkají, jak moc jsou data kolem charakteristiky polohy rozptýlena. Mezi tyto charakteristiky patří *varianční rozpětí*, *kvartilové rozpětí* a *průměrná odchylka*. Varianční rozpětí může být zkresleno odlehlými pozorováními, neboť se vypočte jako rozdíl maximální a minimální hodnoty,

kvartilové rozpětí se vypočítá jako rozdíl mezi horním a dolním kvartilem, je to tedy délka strany krabicového grafu. Průměrná odchylka jako aritmetický průměr rozdílů pozorovaných hodnot od mediánu. Hodnoty těchto chrakteristik shrnuje tabulka 5.4.

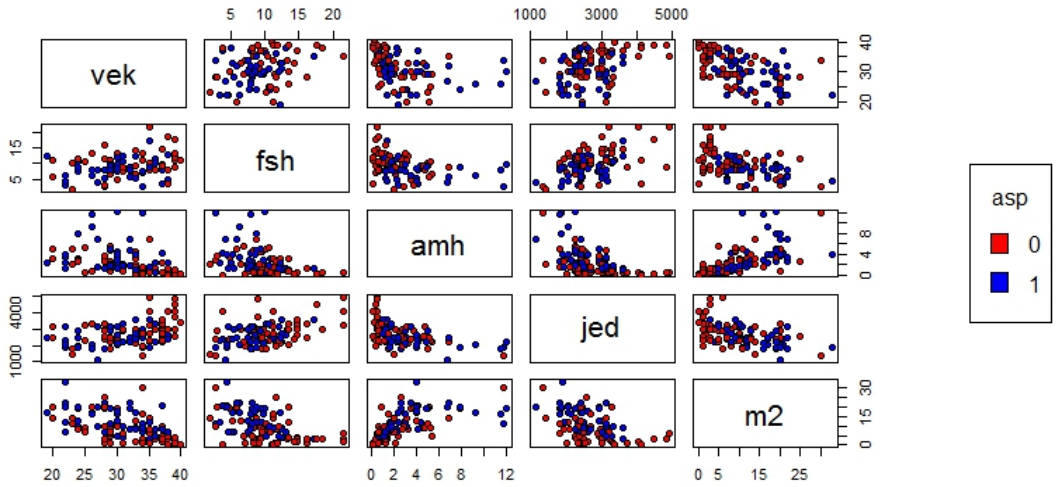
Proměnná	$R = x_{max} - x_{min}$	$R_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$	$d = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \tilde{x}_{0,5} \rvert$
Věk	21	6.25	4.375
FSH	19.8	4.825	3.015
AMH	12.03	2.8025	1.774
jed	3825	750	487.669
m2	33	13	6.463

Tabulka 5.4: Charakteristiky variability

Nyní se podívejme na vztahy mezi proměnnými. Opět si nejprve vykresleme graf. V našem případě zvolíme párový bodový graf, ve kterém navíc barevně odlišíme ženy, kterým se podařilo otěhotnět a ženy, kterým se to nepodařilo. Takovýto graf vytvoříme pomocí následující funkce:

```
>pairs(newdata[1:5], pch = 21, bg = c("red","blue")[unclass(newdata$asp)],
 oma=c(3,3,3,15))
>par(xpd = TRUE)
>legend("right", fill = c("red","blue"),
 legend = c( levels(newdata$asp)), title="asp")
```

Párový bodový graf je na obrázku 5.3. Červeně jsou označeny ženy, kterým se otěhotnět nepodařilo, zatímco modrá barva představuje ženy, které otěhotněly. Z grafu jsou vidět závislosti mezi veličinami. Přímá závislost je vidět mezi hladinou hormonu AMH a počtem vajíček (proměnná m2) a naopak nepřímá závislost je patrná mezi hladinou hormonu AMH a věkem. Tyto závislosti ostatně odpovídají tomu, co o hormonu AMH víme. Hormon AMH řídí zrání vajíček, ale s rostoucím věkem jeho hladina klesá. Tím se dostáváme k další závislosti. Jestliže s rostoucím věkem klesá hladina hormonu AMH, klesá pak i počet vajíček. Tato nepřímá závislost je z grafu též patrná. Co se týče věku a hladiny hormonu FSH, zde není viditelná závislost, což také odpovídá tomu, co víme o hormonu FSH, stejně tak mezi věkem a jednotkami dodaných hormonů. Co bychom ale možná nečekali je nepřímá závislost mezi jednotkami dodaných hormonů a počtem vajíček. Ta však může být pouze důsledkem dodání většího počtu jednotek hormonu ženám ve vyšším věku, jejichž hladina hormonu AMH a počet vajíček byl minimální. Tento jev lze pozorovat v podobě shluků červených bodů v grafu. Pokud se například podíváme na proměnnou AMH, zjistíme, že hladina tohoto hormonu je většinou nízká u žen, kterým se nepodařilo otěhotnět.



Obrázek 5.3: Párový bodový graf

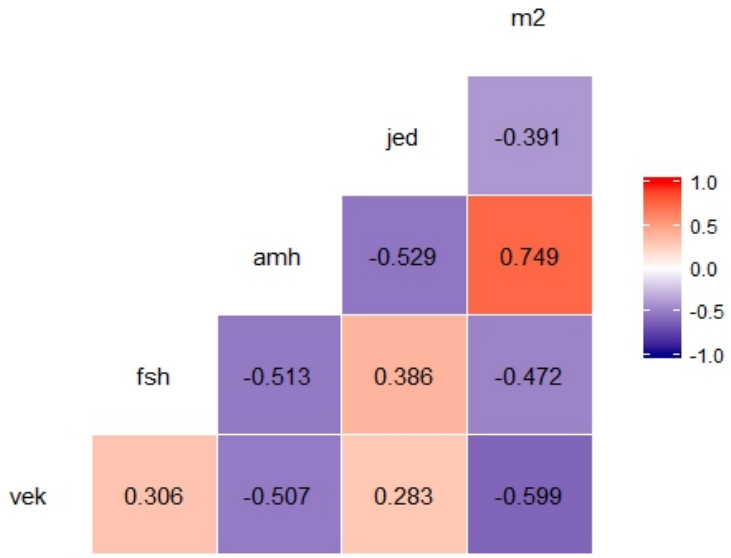
Směr a sílu těchto závislostí můžeme vyjádřit číselně pomocí Spearmanova korelačního koeficientu, který je neparametrický, pracuje s pořadími a je invariantní vůči odlehlým pozorováním. Zjištěné hodnoty máme na obrázku 5.4. Jedná se o tzv. heatmap, která zvýrazní nejsilnější závislosti a zároveň barevně rozliší její směr. Graf vytvoříme v knihovně `GGally` příkazem:

```
ggcorr(newdata[1:5], method = c("everything", "spearman"),
       label = TRUE, label_round = 3, low = "darkblue",
       mid = "white", high = "red")
```

Všechny závislosti se ukázaly významné, neboť odpovídající p-hodnoty byly menší než 0.05. P-hodnoty byly zjišťovány příkazem:

```
>cor.test("název 1. proměnné","název 2. proměnné", method="spearman")
```

Nejsilnější zjištěná závislost je přímá závislost mezi hladinou hormu AMH a počtem vajíček, o které jsme se zmínili již v souvislosti s párovým bodovým grafem. Další silným vztahem je nepřímá závislost mezi věkem a počtem vajíček, který jsme též zmiňovali. Třetím nejsilnějším vztahem je nepřímá závislost hladiny hormonu AMH a počtu dodaných jednotek hormonu FSH. Následují nepřímé závislosti mezi hladinou hormonu AMH a FSH a hladinou hormonu FSH a počtem vajíček. Jako nejslabší závislost byla detekována přímá závislost mezi věkem a jednotkami dodaných hormonů, následována přímou závislostí mezi věkem a hladinou hormonu FSH.

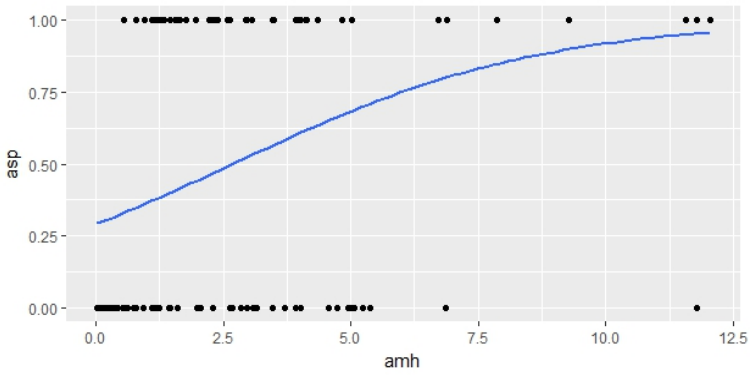


Obrázek 5.4: Spearmanovy korelační koeficienty

5.3. Analýza vlivu hladiny hormonu AMH na pravděpodobnost otěhotnění

V této podkapitole budeme modelovat vliv hladiny hormonu AMH na pravděpodobnost otěhotnění u žen léčených na neplodnost pomocí logistické regrese. Než začneme tvořit takovýto model, vykresleme si naši situaci pomocí grafu z knihovy `ggplot2`. Graf z obrázku 5.5 dostaneme zadáním funkce:

```
>ggplot(newdata,aes(amh,asp))+
geom_point()+
geom_smooth(method="glm",se=FALSE,method.args=list(family="binomial")).
```



Obrázek 5.5: Graf logistické regrese

Na ose x máme proměnnou AMH a na ose y proměnnou asp. Způsob jakým jsou body v grafu umístěny je typický právě pro logistickou regresi, neboť závislá proměnná asp nabývá pouze dvou hodnot. V grafu máme modře vyznačenou odhadnutou logistickou křivku odpovídající našim datům. Tuto křivku se pokusíme vyjádřit pomocí modelu logistické regrese.

Logistický regresní model v softwaru R nyní vytvoříme příkazem `glm` s agrumentem `family=binomial`. Informace o odhadnutém modelu zíkáme příkazem `summary("název modelu")`. Příkazy a výstup jsou ve tvaru:

```
>model1=glm(asp~amh,data=newdata,family=binomial)
>summary(model1)

Call:
glm(formula = asp ~ amh, family = binomial, data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4761  -0.9892  -0.2679   1.1389   1.4876

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.89083    0.30261  -2.944 0.003242 **
amh          0.33180    0.09681   3.427 0.000609 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 188.54  on 135  degrees of freedom
Residual deviance: 172.76  on 134  degrees of freedom
AIC: 176.76

Number of Fisher Scoring iterations: 4
```

Na základě velmi malé p-hodnoty vidíme, že vliv hladiny hormonu AMH na pravděpodobnost otěhotnění je statisticky významný. Model, který jsme takto vytvořili odpovídá následujícímu zápisu:

$$logit(P(Y=1)) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1.$$

Pokud dosadíme nalezené hodnoty regresních koeficientů, dostaneme:

$$logit(P(Y=1)) = \ln\left(\frac{P(asp=1)}{1-P(asp=1)}\right) = -0.89083 + 0.33180 * \text{AMH}.$$

Intervaly spolehlivosti pro nalezené koeficienty β_0, β_1 si můžeme vypočítat pomocí příkazu `confint("název modelu")`. Dostáváme:

```
                2.5 %     97.5 %
(Intercept) -1.5083201 -0.3169086
amh          0.1563611  0.5366142
```

Protože ani jeden z uvedených intervalů neobsahuje nulu, můžeme prohlásit, že jsou oba tyto parametry statisticky významné, což jsme ostatně viděli již z předchozího výstupu softwaru R.

Interpretovat nalezené hodnoty v tomto tvaru je však poněkud obtížné. Na konci první kapitoly jsme si ale odvodili vztah $OR(X_1) = e^{\beta_1}$. Ten nyní můžeme využít prostým aplikováním exponenciální funkce na nalezené koeficienty. Navíc lze tuto aplikaci použít i na intervaly spolehlivost pro regresní koeficienty, čímž získáme intervaly spolehlivosti pro poměr šancí.

```
>exp(model1$coefficients)
(Intercept)         amh
  0.4103167   1.3934730

>exp(confint(model1))
                2.5 %    97.5 %
```

```
(Intercept) 0.2212814 0.7283973
amh         1.1692483 1.7102067
```

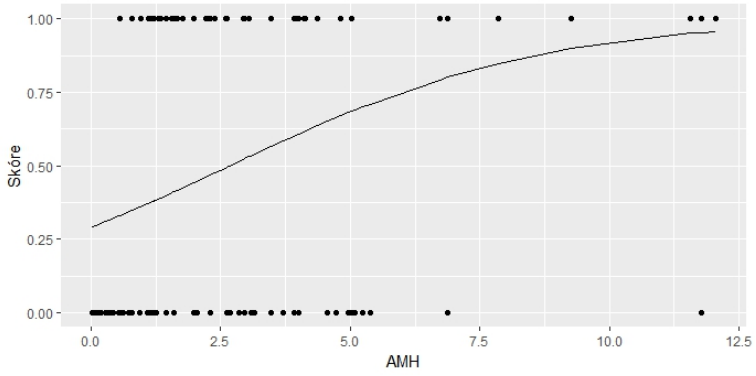
Teď již můžeme interpretovat hodnoty nalezených koeficientů. Koeficient β_1 nám říká, že pokud zvýšíme hodnotu hormonu AMH o jednotku, zvýší se šance na otěhotnění asi 1.39-krát. Koeficient β_0 udává šanci na otěhotnění, jestliže bude hodnota hormonu AMH nulová. Protože je tento koeficient menší než 1, je šance na neúspěch oproti úspěchu větší.

Pravděpodobnost otěhotnění pro různé hodnoty hladiny hormonu AMH dostaneme z rovnice:

$$P(asp = 1) = \frac{\exp\left(-0.89 + 0.33 * \text{AMH}\right)}{1 + \exp\left(-0.89 + 0.33 * \text{AMH}\right)}.$$

Abychom viděli, jak se tato pravděpodobnost mění pro různé hodnoty AMH, vykreslíme si graf závislosti modelem odhadnuté pravděpodobnosti otěhotnění na hodnotě hladiny AMH (obrázek 5.6). Pro úplnost do grafu vykreslíme i napozorované hodnoty asp. Dostaneme podobný graf jako je na obrázku 5.5, tentokrát bude ale logistická křivka odpovídat modelem odhadnuté pravděpodobnosti otěhotnění. Graf získáme opět v knihovně `ggplot2` příkazem:

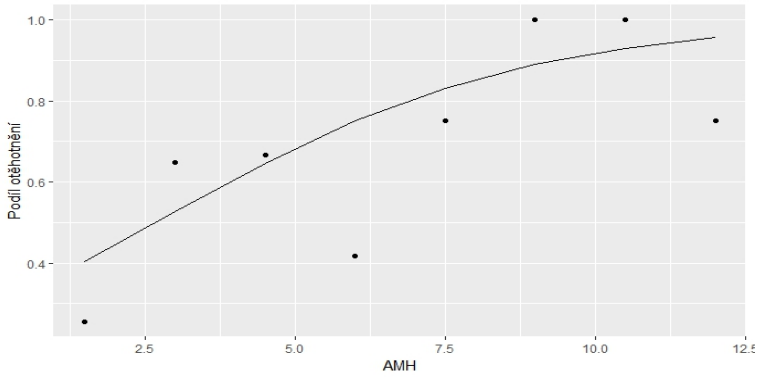
```
ggplot(newdata,aes(amh,asp))+
  geom_point(aes(y=asp))+
  geom_line(aes(x=amh,y=prob))+
  labs(y = "Skóre",x="AMH")
```



Obrázek 5.6: Modelem odhadnutá pravděpodobnost otěhotnění pro hodnoty AMH

Z obrázku 5.6 však není příliš zřetelné, zda takto vytvořený model dobře popisuje naše data. Vykresleme si proto ještě jeden graf (obrázek 5.7), ve kterém si rozdělme vysvětlující

proměnnou AMH na intervaly o délce 1.5 jednotek hladiny AMH, v každém z nich si vypočítejme podíl úspěšných pozorování a tyto hodnoty do grafu vykresleme namísto hodnot napozorovaných. Příkazy, kterými byl graf vytvořen jsou přiloženy v sekci přílohy.



Obrázek 5.7: Pravděpodobnost otěhotnění v závislosti na AMH

Nyní využijeme znalostí z kapitoly 4, abychom zjistili, jak dobrý náš model je. Nejprve provedeme křížovou validaci, kterou necháme proběhnout 10-krát s volbou prahového bodu $P_c = 0.5$. Za využití knihovny `caret` dostaneme pomocí příkazů přiložených v sekci Přílohy následující výstup:

```
>crossval
Generalized Linear Model

136 samples
  1 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 122, 123, 122, 122, 122, 123, ...
Resampling results:

  Accuracy   Kappa
  0.6263736  0.2444966
```

Z výstupu vidíme, že celkový rozsah uvažovaného datového souboru je 136 pozorování při jedné vysvětlující proměnné a dvou hodnotách vysvětlované proměnné. Křížová validace proběhla 10-krát s testovacími skupinami, jejichž rozsahy jsou uvedeny v `Summary of sample sizes`. V každém opakování tedy bylo uvažováno buď 122 nebo 123 pozorování

jako trénovací a 13 nebo 14 pozorování jako testovací skupina. Výsledkem takto provedené křížové validace je hodnota přesnosti 62.6% a hodnota Cohenova kappa 0.24. Je to však jen jeden z výsledků, neboť příkaz vždy vybere skupinu trénovacích a testovacích pozorování náhodně. Pro přesnější hodnoty těchto statistik můžeme využít funkci `confusionMatrix`, v jejímž výstupu najdeme mj. například klasifikační matici, celkovou přesnost modelu nebo odpovídající hodnoty senzitivity a specificity. Příkaz je ve tvaru:

```
>predikce=predict(model, type="response")>=0.5
>predikce <- ifelse(predikce=="TRUE",1,0)
>confusionMatrix(predikce,asp,positive="1")
```

Do příkazu `confusionMatrix` tedy dosazujeme modelem predikovaný výsledek léčby, pozorované hodnoty a je třeba určit, jak je v těchto hodnotách kódován úspěch. Prahový bod prozatím ponecháme $P_c = 0.5$. Podívejme se na výstup:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 48 31
         1 20 37

               Accuracy : 0.625
                 95% CI : (0.5379, 0.7065)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.002241

                  Kappa : 0.25
 Mcnemar's Test P-Value : 0.161429

            Sensitivity : 0.5441
            Specificity : 0.7059
         Pos Pred Value : 0.6491
         Neg Pred Value : 0.6076
             Prevalence : 0.5000
         Detection Rate : 0.2721
   Detection Prevalence : 0.4191
      Balanced Accuracy : 0.6250

       'Positive' Class : 1
```

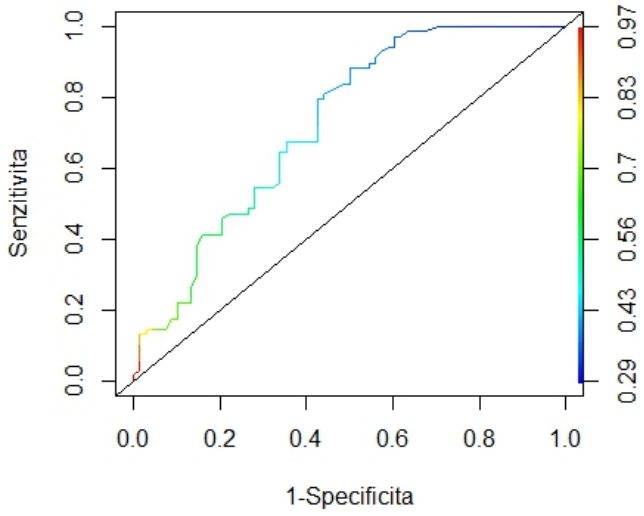
Nejdříve se podívejme na klasifikační matici. Vidíme, že celkem 48+37=85 pozorování bylo klasifikováno správně. Zatímco 31 pozorování vyšlo jako falešně negativní a 20 pozorování jako falešně pozitivní. Výsledná přesnost předpovědi je 62.5%, ve výstupu je i 95% interval spolehlivosti pro přesnost. `No Information Rate` (v našem případě 0.5) je míra správných predikcí, pokud bychom neuvažovali vysvětlující proměnné a pozorováním přiřazovali očekávané hodnoty zcela náhodně. `P-Value [Acc>NIR]` (rovna 0.002) říká, že náš uvažovaný model je lepší než náhodná predikce.

Následuje Cohenovo kappa, jehož hodnota 0.25 značí, že model data predikuje o jen něco lépe než náhodný proces. Dalším údajem je p-hodnota pro McNemarův test, který podobně jako Cohenovo kappa testuje shodu procenta skutečně úspěšných a očekávaných úspěšných případů. V tomto případě shodu nezamítáme. Pokud bychom využili vzorce z kapitoly 4.1, vyšla by nám testová statistika $T = 2.37$ a odpovídající kritická hodnota $\chi^2_{0.95}(1) = 3.84$. I touto cestou tedy shodu nezamítáme, neboť $T = 2.37 < 3.84$.

Následují hodnoty senzitivity a specificity. Připomeňme, že senzitivita vyjadřuje úspěšnost, s jakou model správně klasifikuje pozorování jako úspěšné, je-li skutečně úspěšné. Tento model vyhodnotil přibližně 54.4% skutečně úspěšných případů jako úspěšné. Naproti tomu specificita vyjadřuje úspěšnost, s jakou model správně klasifikuje pozorování jako neúspěšné, je-li skutečně neúspěšné. Náš model vyhodnotil přibližně 70.6% skutečně negativních jako negativní. Další vypočítanou hodnotou je prediktivní hodnota pozitivního testu, která říká, že přibližně 64.9% procent pozorování, která model označil jako úspěšná, jsou skutečně úspěšná. Následuje prediktivní hodnota negativního testu, která v tomto případě říká, že přibližně 60.8% pozorování vyhodnocených jako neúspěšných, je skutečně neúspěšných.

Prevalence obecně je výskyt sledovaného jevu v populaci. V našem případě se jedná o počet žen, kterým se podařilo otěhotnět (68) ku počtu všech žen (136) tedy $\frac{68}{136} = 0.5$. `Detection Rate` (0.27) je počet správně klasifikovaných úspěšných pozorování ku celkovému počtu pozorování. `Detection Prevalence` (0.42) je počet pozorování modelem klasifikovaných jako úspěšná ku počtu všech pozorování. Jedná se tedy o modelem vypočítanou míru výskytu jevu (otěhotnění) v populaci. `Balanced Accuracy` (62.5%) se vypočítá jako součet senzitivity a specificity vydělený dvěmi. V našem případě je tato hodnota rovna hodnotě pro přesnost, neboť máme v datovém souboru shodný počet úspěšných a neúspěšných pozorování (tzv. vyvážené třídění). Poslední hodnota z výstupu, označená jako `Positive Class`, nás informuje o tom, kterou skupinu pozorování software uvažoval jako pozitivní případy (v našem případě jsou pozitivní případy oznaeny jako 1).

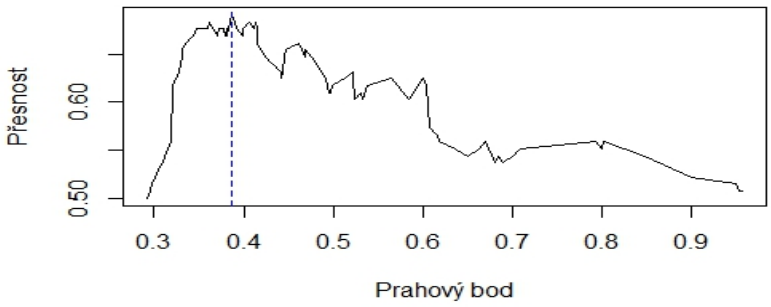
Jak už bylo řečeno, závislost senzitivity na 1-specificitě nám vykresluje ROC křivka, kterou si v sofwaru R můžeme vytvořit v knihovně `ROCR`. Dále si můžeme vypočítat odpovídající hodnotu AUC. Příkazy pro vytvoření grafu i výpočet hodnoty AUC jsou opět přiloženy v sekci Přílohy. Výsledný graf je na obrázku 5.8.



Obrázek 5.8: ROC křivka

Odpovídající hodnota AUC je přibližně 0.7234. Na obrázku 5.8 vidíme navíc i barevnou škálu, která představuje odpovídající prahové body. Pokud bychom zvolili prahový bod například 0.9, potom by byly hodnoty senzitivity i 1-specificity blízko 0. Naopak pokud bychom zvolili jako prahový bod hodnotu 0.3, byly by obě tyto hodnoty vysoké. Otázkou je, jak zjistit optimální hodnotu prahového bodu.

Jednou z možností je zvolit prahový bod tak, aby byla přesnost modelu maximální. Pro náš model je tato hodnota $P_c = 0.387$ s odpovídající přesností 0.6912. Hodnota senzitivity je přitom 0.8824 a specificita je rovna 0.5. Prediktivní hodnota pozitivního testu je rovna 63.83% a prediktivní hodnota negativního testu je rovna řibliže 80.95%. Celou situaci si můžeme znázornit jako graf závislosti přesnosti modelu na hodnotě prahového bodu. Příkazy pro výpočet těchto hodnot a tvorbu grafu jsou přiloženy v sekci Přílohy.



Obrázek 5.9: Graf znázorňující volbu prahového bodu

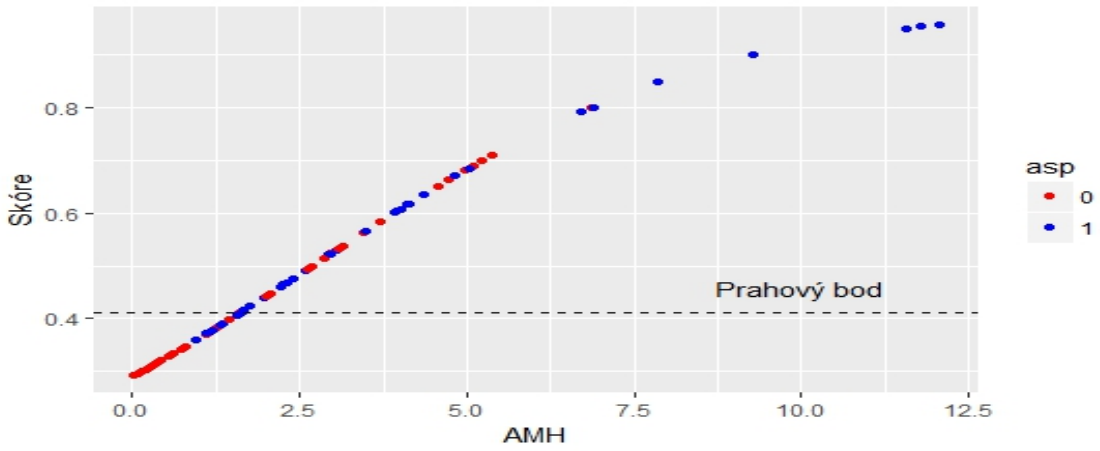
Jinou možností by byla volba prahového bodu tak, aby byly rovnoměrně maximalizovány hodnoty senzitivity a specifiticy. Příkaz pro zjištění jejich hodnot a odpovídajícího prahového bodu je přiložen v sekci Přílohy. [20] Výsledek pro náš model je následující:

```
sensitivity 0.7941176
specificity 0.5735294
cutoff      0.4125797
```

Pomocí přikazu `confusionMatrix` si můžeme opět vypočítat přesnost pro tento prahový bod (0.6765) i preditivní hodnotu pozitivního testu (64.63%) a prediktivní hodnotu negativního testu (72.22%). Zjištěné hodnoty lze interpretovat tak, že model správně klasifikoval 79.4% úspěšných případů a 57.4% neúspěšných případů. Dále jestliže bude skóre pro dané pozorování přiřazené modelem větší než 0.41, pak bude pravděpodobnost, že žena skutečně otěhotní rovna 64.63%, naopak pokud bude hodnota přiřazeného skóre menší než prahový bod, pak pravděpodobnost, že žena opravdu neotěhotní bude 72.22%. Cohenovo kappa je rovno 0.3529, což je lepší výsledek, než při prahovém bodu $P_c = 0.5$. P-hodnota McNemarova testu je však v tomto případě rovna 0.05002.

Na obrázku 5.10 je graf znázorňující odhadnuté pravděpodobnosti otěhotnění při daných hodnotách hladiny hormony AMH pro naše pozorování. Dále je zde vyznačena hodnota prahového bodu, který rovnoměrně maximalizuje hodnoty senzitivity a specifitity (tedy hodnota $P_c = 0.41$). Pozorování příslušné hodnotám pod touto čarou model vyhodnotil jako neúspěšné případy a naopak nad čarou jako úspěšné. Barevně jsou rozlišeny skutečné úspěšné a neúspěšné případy. Příkaz, kterým byl graf vytvořen je přiložen v sekci Přílohy.

Nakonec se podívejme na testy dobré shody, které nám ověřují, zda se významně neliší očekávané hodnoty (predikované modelem) a pozorované hodnoty. V tabulce 5.5 jsou



Obrázek 5.10: Odhady pro pravděpodobnost otěhotnění

hodnoty testových statistik a příslušné kritické hodnoty. Z výsledků plyne, že pouze na základě Pearsonova testu hypotézu o shodě pozorovaných a očekávaných hodnot nezamítáme. Příslušné příkazy pro zjištění těchto hodnot jsou přiloženy v sekci Přílohy.

Test	Hodnota test. statistiky	kritická hodnota	Závěr
Pearson	146.04	160.91	Nezamítáme
Roz. test deviance	172.76	160.91	Zamítáme
Hosmer-Lemeshow C_g	21.90	15.51	Zamítáme
Hosmer-Lemeshow H_g	23.48	15.51	Zamítáme

Tabulka 5.5: Testy dobré shody pro model1

Podívejme se nyní na Kolmogorovův-Smirnovův test dobré shody. Testujeme nulovou hypotézu, že se distribuční funkce úspěšných a neúspěšných případů pro skóre shodují. Hodnota testové statistiky je $\mathrm{KS} = 0.382$, odpovídající kritická hodnota je $\mathrm{KS}_{0.95}(68, 68) = 0.233$, kde 68 je počet úspěšných i neúspěšných pozorování. Hypotézu o shodě distribučních funkcí tedy zamítáme, neboť $\mathrm{KS} = 0.382 > 0.233$, což svědčí o dobré klasifikační schopnosti modelu. Příkaz pro výpočet hodnoty KS je přiložen v sekci Přílohy.

Podívejme se ještě na hodnoty koeficientů hodnotící míru asociace mezi pozorovanými a očekávanými hodnotami. Příkazem [21] (rovněž přiložen v sekci Přílohy) byly zjištěny hodnoty Kendallova τ_a, Somersova D_s, Goodmannova-Kruskalova γ, ale také procento konkordantních (72.12%), diskordantních (27.44%) a vázaných párů (0.43%) v souboru. Protože příkaz nevypočítá hodnotu Kendallova τ_b, vypočítáme ji dodatečně příkazem `cor.test(model$fitted.values, asp, method="kendall")`. Zjištěné hodnoty koeficientů jsou uvedeny v tabulce 5.6. Hodnoty jednotlivých koeficientů jsou kladné, což souhlasí

s tvrzením, že je většina párů konkordantních.

Koeficient	Hodnota
Kendallovo τ_a	0.225
Kendallovo τ_b	0.319
Somersovo D_s	0.449
Goodmanovo-Kruskalovo γ	0.447

Tabulka 5.6: Míry asociace modelu1

5.4. Analýza vlivu dalších proměnných na pravděpodobnost otěhotnění

V této části práce se zaměříme na hledání a analýzu dalších možných modelů, které by vylepšily stávající model.

Podívejme se, jak pravděpodobnost otěhotnění ovlivňuje volba lékaře. Víme, že ženy měly na výběr ze dvou lékařů. Na obrázku 5.11 je graf znázorňující volbu lékaře a věk žen, barevně je přitom rozlišen výsledek léčby. Graf byl vytvořen v knihovně `ggplot2` příkazem:

```
>ggplot(newdata, aes(x=lekar, y=vek, color=asp))+
  geom_point()
```

Obrázek 5.11: Věk žen a volba lékaře

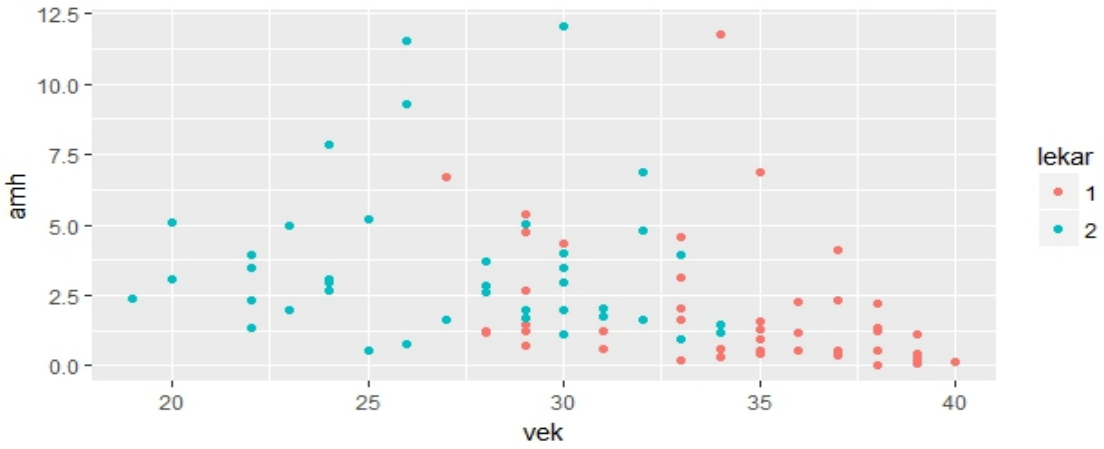
Vidíme, že lékaře číslo 1 si vybíraly o proti lékaři číslo 2 spíše starší ženy. Vidíme také, že většině z nich se otěhotnět nepodařilo, zatímco ženám navštěvujícím lékaře číslo 2 se to většinou podařilo. Příkazem `table(asp,lekar)` si můžeme tyto počty vyčíslit.

```
> table(asp,lekar)
   lekar
asp  1  2
  0 43 25
  1 13 55
```

Vidíme, že přibližně 77% žen navštěvujících lékaře číslo 1 neotěhotnělo. Naproti tomu u lékaře číslo 2 neotěhotnělo pouze 31% žen.

Podívejme se ještě na graf závislosti hladiny hormonu AMH a věku. Barevně rozlišme volbu lékaře. Graf vytvoříme v knihovně `ggplot2` příkazem:

```
>ggplot(newdata, aes(x=vek, y=amh, color=lekar))+
  geom_point()
```



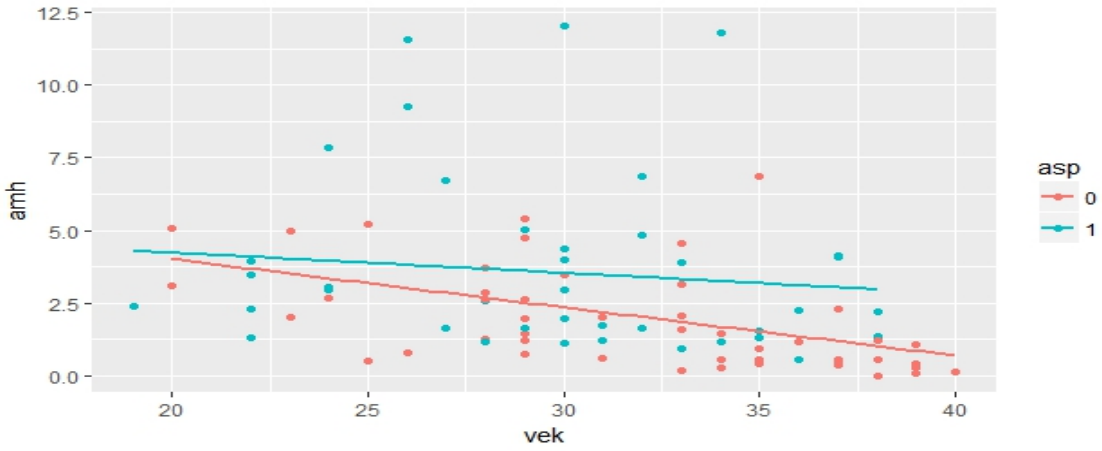
Obrázek 5.12: Graf závislosti hladiny AMH a věku podle volby lékaře

Z obrázku 5.12 je vidět, že lékaře číslo 1 navštěvovaly především mladší ženy, jejichž hladiny hormonu AMH byly vyšší, zatímco lékař číslo 2 měl ve své péči spíše starší ženy, jejichž hladiny hormonu AMH byly nižší, až nulové. Rozdíl v úspěšnosti léčby u obou doktorů tedy může být dán věkovou strukturou pacientek. Z tohoto důvodu nebudeme v dalších modelech uvažovat vliv lékaře jako takového, ale zkusíme analyzovat vliv věku.

5.4.1. Analýza vlivu hladiny AMH a věku na pravděpodobnost otěhotnění

Nejprve si vykresleme graf závislosti hladiny hormonu AMH a věku, tentokrát si ale jednotlivá pozorování rozlišíme barevně podle výsledku léčby. Graf vytvoříme v knihovně `ggplot2` příkazem:

```
>ggplot(newdata, aes(x=vek, y=amh, color=asp))+
      geom_point() +  geom_smooth(method = "lm",se=FALSE).
```



Obrázek 5.13: Graf závislosti AMH a věku podle výsledku léčby

Z obrázku 5.13 vidíme, že hladiny hormonů obou skupin žen se v mladším věku zásadně neliší, tento rozdíl se ale s postupem věku zvyšuje. Ženy, které ve vyšším věku otěhotněly měly vyšší hladinu hormonu AMH než ženy ve stejném věku, kterým se to nepodařilo.

Nejlepší nalezený model analyzující vliv těchto dvou proměnných na pravděpodobnost otěhotnění je model ve tvaru:

$$logit(P(Y = 1)) = \beta_0 + \beta_1 * \text{VĚK} + \beta_2 * \text{AMH} * \text{VĚK}.$$

Proměnnou AMH v tomto případě záměrně do modelu nezahrnujeme, neboť odpovídající regresní koeficient poté vychází nevýznamný. V softwaru tento model dostaneme příkazem:

```
> model2=glm(asp~amh:vek+vek,data=newdata,family=binomial)
> summary(model2)
Call:
glm(formula = asp ~ amh:vek + vek, family = binomial, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.18059  -1.03169  -0.08937   1.03920   1.65946

Coefficients:
```

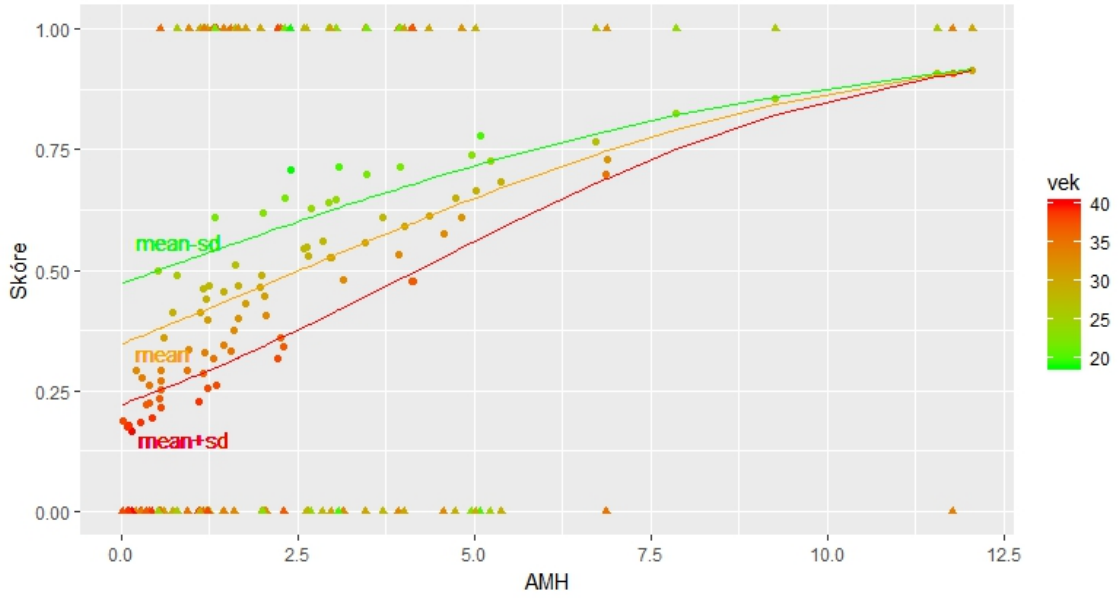
```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.486200   1.173752   2.118  0.03416 *
vek         -0.103880   0.036570  -2.841  0.00450 **
amh:vek      0.008304   0.003133   2.650  0.00805 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.54  on 135  degrees of freedom
Residual deviance: 167.25  on 133  degrees of freedom
AIC: 173.25

Number of Fisher Scoring iterations: 4
```

Vidíme, že všechny regresní koeficienty jsou statisticky významné. Hodnota Akaikeho informačního kritéria AIC je 173.25 (připomeňme, že hodnota AIC pro první model byla 176.76).

Protože je interpretace regresních koeficientů tohoto modelu díky interakci dvou spojitých proměnných značně komplikovanější, podíváme se namísto toho na grafické vyjádření celého modelu (obrázky 5.14 a 5.15). Vykreslíme si graf závislosti modelem predikované pravděpodobnosti otěhotnění na věku a AMH. Pravděpodobnost otěhotnění získáme z rovnice:

$$P(asp = 1) = \frac{\exp\left(2.49 - 0.10 * \text{VĚK} + 0.01 * \text{VĚK} * \text{AMH}\right)}{1 + \exp\left(2.49 - 0.10 * \text{VĚK} + 0.01 * \text{VĚK} * \text{AMH}\right)}.$$

Protože uvedená pravděpodobnost závisí na dvou spojitých proměnných, vykreslíme si nejprve na osu x proměnnou AMH, na osu y modelem přiřazené skóre a jednotlivé body v grafu barevně rozlišíme podle věku, přičemž vykreslíme body nejen pro modelem odhadnutá skóre ale i napozorované hodnoty (obrázek 5.14). Abychom viděli, jak se modelem predikovaná pravděpodobnost otěhotnění při stejných hodnotách AMH mění napříč různými věky, zvolíme si tři pevné hodnoty věku (průměr a průměr $\pm$ směrodatná odchylka). Pro tyto hodnoty věku poté do grafu vykreslíme křivky znázorňující závislost modelem predikované pravděpodobnosti otěhotnění na hladině AMH při pevně danném věku. Příkazy, kterými byl graf vytvořen jsou přiloženy v sekci Přílohy.



Obrázek 5.14: Graf závislosti skóre a hladiny AMH s ohledem na věk

Z obrázku 5.14 vidíme, že při stejné hladině hormonu AMH mají vyšší očekávanou pravděpodobnost otěhotnění mladší ženy (zelená křivka) a naopak pro starší ženy (červená křivka) je tato pravděpodobnost nižší. Se zvyšující se hladinou hormonu AMH se však tyto rozdíly zmenšují. Podívejme se nyní na oranžovou křivku, která odpovídá ženám ve věku 30 let (průměrný věk). Pravděpodobnost otěhotnění odpovídá zápisu:

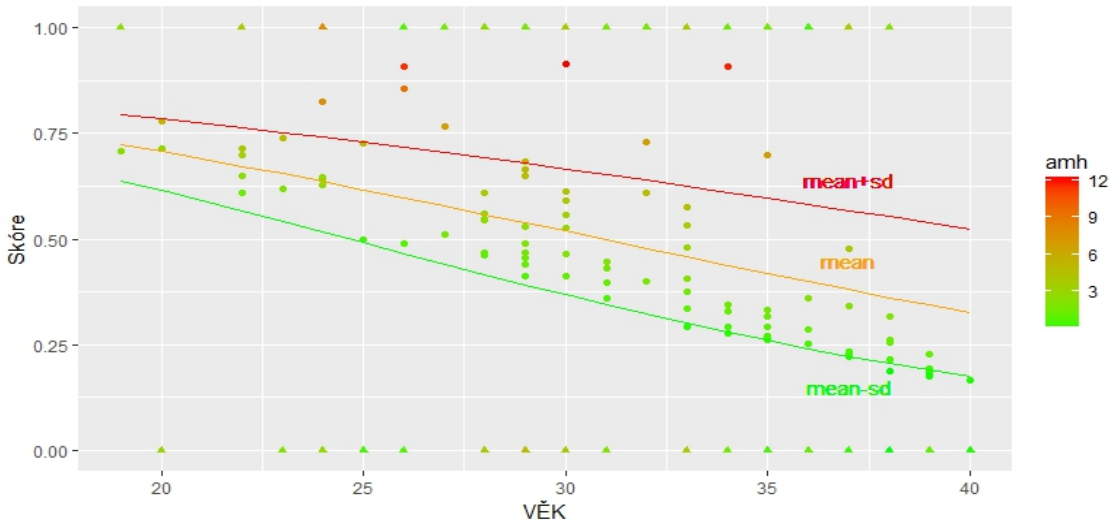
$$P(asp = 1|\text{VĚK} = 30) = \frac{\exp\left(2.49 - 0.10 * 30 + 0.01 * 30 * \text{AMH}\right)}{1 + \exp\left(2.49 - 0.10 * 30 + 0.01 * 30 * \text{AMH}\right)}.$$

Stejným způsobem bychom si mohli vyjádřit pravděpodobnost otěhotnění pro kterýkoli věk. Zaměřme se však nyní na interpretaci regresního koeficientu $\beta_2 = 0.01$, který odpovídá interakci proměnných. Při dosazení věku 30 let dostáváme:

$$e^{0.01*30} = 1.35.$$

Znamená to tedy, že jednotkové zvýšení hladiny hormonu AMH u 30-ti leté ženy znamená 1.35-krát zvýšení šance na otěhotnění.

Analogickým způsobem můžeme zjistit, jak se mění modelem odhadovaná pravděpodobnost otěhotnění pro stejné věky při různých hladinách hormonu AMH. Na obrázku 5.15 máme odpovídající graf, kde na ose y je modelem přiřazené skóre, na ose x je proměnná věk a body jsou barevně rozlišeny podle hladiny hormonu AMH. Do grafu jsou dále vykresleny i body odpovídající napozorovaným hodnotám.



Obrázek 5.15: Graf závislosti skóre a věku s ohledem na hladinu AMH

Opět jsou zde vykresleny tři křivky, které odpovídají modelem odhadovaným pravděpodobnostem otěhotnění pro různé hodnoty věku při pevně daných hodnotách hladiny AMH (průměr ± směrodatná odchylka). V tomto případě mají nižší odhadovanou pravděpodobnost otěhotnění při stejných hodnotách věku ženy s nižší hladinou hormonu AMH (zelená křivka) a naopak ženy s vyšší hladinou hormonu AMH mají tuto pravděpodobnost vyšší.

Podívejme, jak je vyjádřena pravděpodobnost otěhotnění při různých hodnotách věku pro průměrnou hodnotu hladiny AMH (2.83 ng/ml), kterou v grafu 5.15 představuje oranžová křivka.

$$P(asp = 1|\text{AMH} = 2.83) = \frac{\exp\left(2.49 - 0.10 * \text{VĚK} + 0.01 * \text{AMH} * 2.83\right)}{1 + \exp\left(2.49 - 0.10 * \text{VĚK} + 0.01 * \text{AMH} * 2.83\right)}.$$

Odtud si můžeme vypočítat, jak se pro ženu změní šance na otěhotnění za předpokladu zvýšení věku o jeden rok při průměrné hladině hormonu AMH. Dostáváme:

$$\text{OR}(\text{věk}) = e^{\beta_1 + \beta_2 * 2.83} = e^{-0.10 + 0.01 * 2.83} = 0.92.$$

Zvýšením věku ženy o jeden rok se tedy za předpokladu průměrné hodnoty hladiny AMH pravděpodobnost otěhotnění sníží 0.92-krát.

Analogickým postupem jako u modelu1 si můžeme vypočítat statistiky hodnotící kvalitu nalezeného modelu. Hodnota AUC je v tomto případě přibližně 0.715. Před výpočtem

dalších hodnotících statistik však nejprve nalezněme prahový bod. Prahový bod maximalizující přesnost tohoto modelu je $P_c = 0.316$. Pomocí příkazu `confusionMatrix` zjistíme, že odpovídající přenost je 0.6765. Hodnota senzitivity je 0.9706 a specificity 0.3824. Prediktivní hodnota pozitivního testu je potom 0.6111 a prediktivní hodnota negativního testu je 0.9286. Cohenovo kappa je rovno 0.3529 (stejně jako u modelu1). McNemarův test však nulovou hypotézu shody procenta skutečně pozitivních a procenta očekávaných pozitivních zamítá.

Podívejme se, jaké výsledky nám dávají testy dobré shody (tabulka 5.7), kde oproti modelu1 pozorujeme zlepšení, neboť shodu pozorovaných a očekávaných hodnot zamítáme pouze rozdílovým testem deviance.

Test	Hodnota test. statistiky	kritická hodnota	Závěr
Pearson	136.02	159.81	Nezamítáme
Roz. test deviance	167.25	159.81	Zamítáme
Hosmer-Lemeshow C_g	14.93	15.51	Nezamítáme
Hosmer-Lemeshow H_g	12.79	15.51	Nezamítáme

Tabulka 5.7: Testy dobré shody pro model2

Hodnota testové statistiky pro Kolmogorovův-Smirnovův test je KS = 0.353, kritická hodnota je opět 0.233. Stejně jako u modelu1 tedy zamítáme hypotézu o shodě distribučních funkcí skóre úspěšných a neúspěšných případů.

V tomto případě máme 71.30% konkordantních, 28.33% diskorantních a pouze 0.37% vázaných párů. Hodnoty koeficientů hodnotících míry asociace mezi pozorovanými a očekávanými hodnotami jsou uvedeny v tabulce 5.8. Tyto hodnoty jsou nižší než u modelu1, nedošlo však k nijak zásadnímu poklesu.

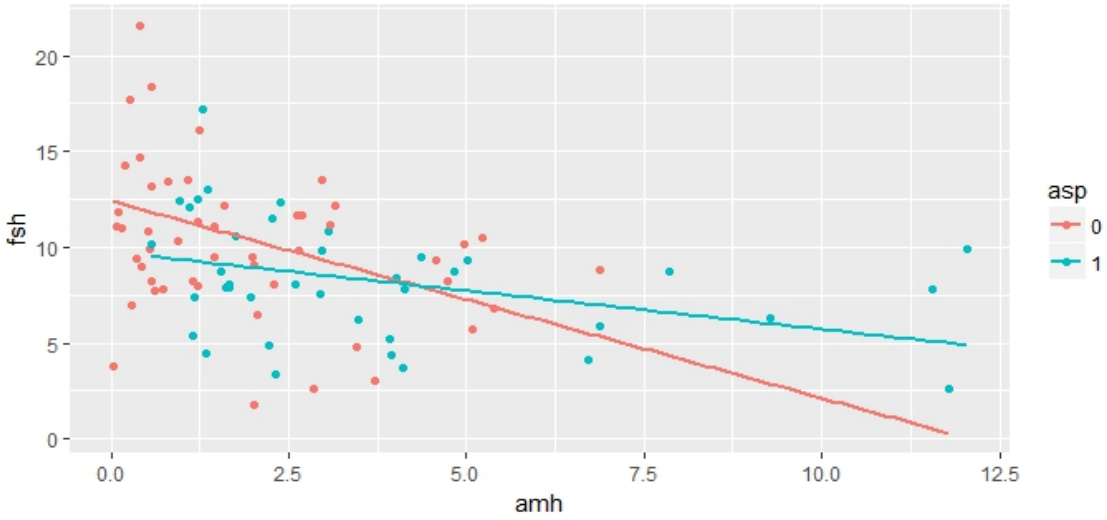
Koeficient	Hodnota
Kendallovo τ_a	0.216
Kendallovo τ_b	0.306
Somersovo D_s	0.431
Goodmanovo-Kruskalovo γ	0.430

Tabulka 5.8: Míry asociace pro model2

5.4.2. Analýza vlivu hladiny AMH a FSH na pravděpodobnost otěhotnění

Opět si nejprve vykresleme graf závislosti hladiny AMH a FSH. Barevně si rozlišme případy s úpěšnou a núspěšnou léčbou. Graf získáme v knihovně `ggplot2` příkazem:

```
> ggplot(newdata, aes(x=amh, y=fsh, color=asp))+
geom_point() +  geom_smooth(method = "lm",se=FALSE)
```



Obrázek 5.16: Graf závislosti AMH a FSH podle výsledku léčby

Z obrázku 5.16 vidíme, že hodnota hladiny hormonu FSH se pro skupinu úspěšných pozorování napříč všemi hladinami hormonu AMH nijak výrazně nemění a kolísá kolem hodnot 5-10 IU/l. Pro neúspěšná pozorování se však hladina tohoto hormonu pohybuje po celé škále (0-20 IU/l) a odpovídající hladina hormonu AMH se až na vyjímky pohybuje do 5 ng/ml.

Pokud bychom chtěli modelovat nejen vliv hladiny hormonu AMH, ale i hladiny hormonu FSH (bez interakcí), zjistili bychom, že regresní koeficient odpovídající právě hormonu FSH není statisticky významný. Je to dáno tím, že hladina hormonu FSH je korelována s hladinou hormonu AMH. Připomeňme, že mezi FSH a AMH je Spearmanův korelační koeficient roven -0.513. Jestliže však model obsahuje interakci mezi oběma proměnnými, je odpovídající regresní koeficient významný. Nejlepším nalezeným modelem obsahující obě proměnné je následující:

$$logit(P(Y = 1)) = \beta_0 + \beta_1 * \text{FSH} + \beta_2 * \text{FSH} * \text{AMH}.$$

V modelu není záměrně zahrnuta proměnná AMH, neboť odpovídající regresní koeficient poté vychází nevýznamný. V softwaru dostaneme tento model následovně:

```
> model3=glm(asp~amh:fsh+fsh,data=newdata,family=binomial)
```

```
> summary(model3) # obě proměnné statisticky významné, intercept ne

Call:
glm(formula = asp ~ amh:fsh + fsh, family = binomial, data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8873  -0.9972  -0.1171   1.0711   1.6593

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.45151    0.55542   0.813  0.41626
fsh         -0.14061    0.05137  -2.737  0.00620 **
amh:fsh      0.03941    0.01299   3.033  0.00242 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.54  on 135  degrees of freedom
Residual deviance: 167.24  on 133  degrees of freedom
AIC: 173.24

Number of Fisher Scoring iterations: 4
```

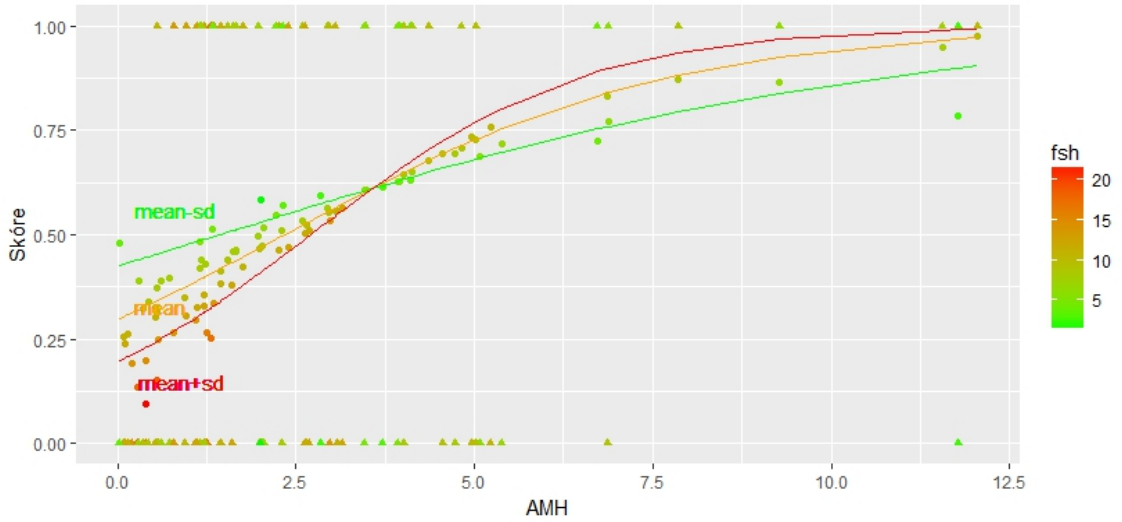
Vidíme, že hodnota AIC je v tomto případě nejmenší. Naopak hodnota AUC je rovna 0.704 a je tak z těchto tří modelů nejnižší.

Podobně jako u předchozího modelu si namísto intepretace jednotlivých regresních koeficientů vykreslíme grafy znázorňující závislost modelem odhadnuté pravděpodobnosti otěhotnění na obou uvažovaných proměnných (obrázky 5.17 a 5.18). Pravděpodobnost je v tomto případě rovna:

$$P(asp = 1) = \frac{\exp\left(0.45 - 0.14 * \mathrm{FSH} + 0.04 * \mathrm{FSH} * \mathrm{AMH}\right)}{1 + \exp\left(0.45 - 0.14 * \mathrm{FSH} + 0.04 * \mathrm{FSH} * \mathrm{AMH}\right)}.$$

Nejprve na osu x vykreslíme hladinu hormonu AMH a na osu y modelem přiřazené skóre. Do grafu opět vykreslíme kromě bodů odpovídajících modelem přiřazenému skóre také napozorované hodnoty. Jednotlivé body v grafu, barevně rozlišíme podle hladiny hormonu FSH (obrázek 5.17). I v tomto případě zvolíme tři hodnoty hladiny hormonu FSH (průměr

a ± směrodatná odchylka), pro které do grafu vykreslíme křivky odpovídající závislosti modelem predikované pravděpodobnosti otěhotnění na hladině hormonu AMH.



Obrázek 5.17: Graf závislosti skóre a hladiny AMH s ohledem na hladinu FSH

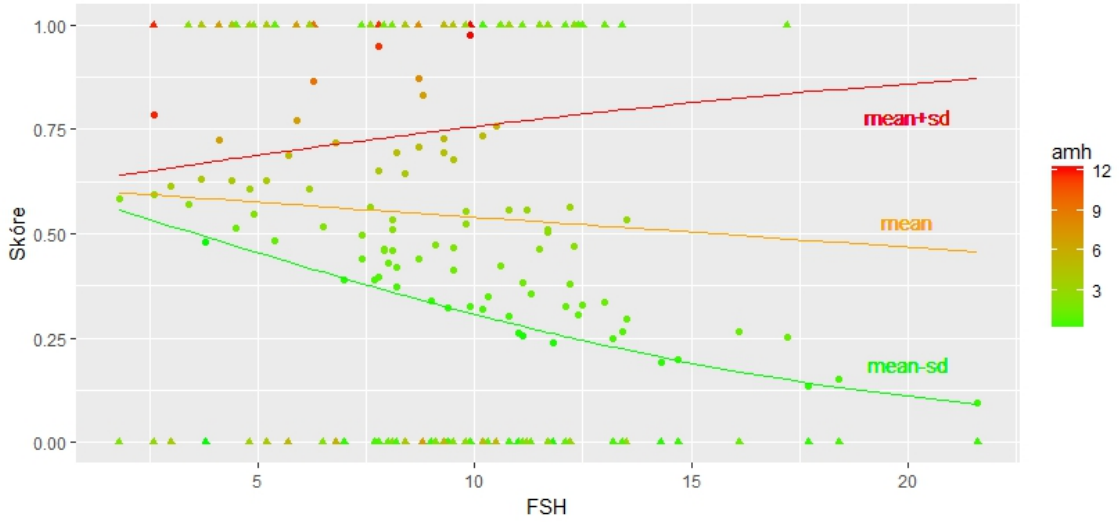
Vidíme, že při hladině hormonu AMH přibližně 3.5 ng/ml je pravděpodobnost otěhotnění pro všechny hodnoty hladiny hormonu FSH přibližně 0.60. Jestliže však budeme uvažovat hodnoty hormonu AMH nižší než 3.5 ng/ml, pak mají vyšší pravděpodobnost otěhotnění ženy s nižší hladinou hormonu FSH a naopak při hodnotách hormonu AMH vyšších než 3.5 ng/ml mají vyšší pravděpodobnost otěhotnění ženy s vyšší hladinou hormonu FSH. Dosaďme si do rovnosti pro pravděpodobnost průměrnou hodnotu hladiny hormonu FSH (9.31 IU/l). Získáme:

$$P(asp = 1|\text{FSH=9.31}) = \frac{\exp\left(0.45 - 0.14 * 9.31 + 0.04 * 9.31 * \text{AMH}\right)}{1 + \exp\left(0.45 - 0.14 * 9.31 + 0.04 * 9.31 * \text{AMH}\right)}.$$

Můžeme tedy opět interpretovat regresní koeficient β_2 odpovídající interakci v modelu jako změnu šance na otěhotnění při jednotkové změně hormonu AMH a při dané hladině hormonu FSH. Konkrétně pro průměrnou hodnotu hormonu FSH dostáváme, že se šance na otěhotnění při jednotkovém zvýšení hladiny hormonu AMH zvýší $e^{0.04*9.31} = 1.45$-krát.

Zbývá zjistit, jak se mění modelem odhadovaná pravděpodobnost otěhotnění pro stejné hodnoty hormonu FSH v závislosti na různých hodnotách hormonu AMH. Opět si vykreslíme graf (obrázek 5.18 kde na osu x vykreslíme hladiny hormonu FSH, na osu y modelem predikovanou pravděpodobnost otěhotnění a body v grafu (pro skóre i napozorované

hodnoty) barevně rozlišíme podle hladiny hormonu AMH. Dále do grafu vykreslíme křivky odpovídající modelem odhadované pravděpodobnosti otěhotnění pro hodnoty hladiny FSH při pevně daných hodnotách AMH (průměr ± směrodatná odchylka).



Obrázek 5.18: Graf závislosti skóre a hladiny FSH s ohledem na hladinu AMH

Z obrázku 5.18 vidíme, že při nižších hodnotách hladiny hormonu FSH je modelem predikovaná pravděpodobnost otěhotnění pro všechny hladiny hormonu AMH přibližně stejná. Se zvyšující se hladinou hormonu FSH se však tato pravděpodobnost pro různé hodnoty hladiny hormonu AMH liší, přičemž vyšší odhadovanou pravděpodobnost otěhotnění při stejné hodnotě hladiny hormonu FSH mají ženy s vyšší hladinou hodmonu AMH (červená křivka). Naopak nižší odhadovanou pravděpodobnost otěhotnění mají při stejné hodnotě hladiny hormonu FSH ženy s nižší hladinou hormonu AMH (zelená křivka).

Vyjádřeme si nyní rovnost pro modelem odhadovanou pravděpodobnost otěhotnění při průměrné hodnotě hladiny hormonu AMH (oranžová křivka).

$$P(asp = 1|\text{AMH=2.83}) = \frac{\exp\left(0.45 - 0.14 * \text{FSH} + 0.04 * \text{FSH} * 2.83\right)}{1 + \exp\left(0.45 - 0.14 * \text{FSH} + 0.04 * \text{FSH} * 2.83\right)}.$$

Z této rovnosti si můžeme vypočítat, jak se při průměrné hladině hormonu AMH (2.83 ng/ml) zvýšením hladiny hormonu FSH o jednotku změní šance na otěhotnění. Výpočet je následovný:

$$\text{OR(FSH)} = e^{\beta_1 + \beta_2 * 2.83} = e^{-0.14 + 0.04 * 2.83} = 0.97.$$

Znamená to tedy, že zvýšení hladiny hormonu FSH o jednotku u ženy, která má průměrnou hladinu hormonu AMH (2.83 ng/ml), znamená 0.97-krát snížení šance na otěhotnění.

Analogickým způsobem jako u modelu1 můžeme opět zjistit, že prahový bod maximalizující přesnost (0.6985) je $P_c = 0.425$. Odpovídající hodnota senzitivity je 0.8529, hodnota specificity je 0.5441, prediktivní hodnota pozitivního testu je 0.6517 a prediktivní hodnota negativního testu je 0.7872. Cohenovo kappa je rovno 0.3971 (vyšší než u předcházejících modelů). I v tomto případě však hypotézu McNemarova testu zamítáme. Hodnoty těchto statistik jsme opět získali pomocí příkazu `confusionMatrix`.

V tabulce 5.9 jsou uvedeny hodnoty testovacích statistik pro testy dobré shody včetně kritických hodnot a výsledku testu.

Test	Hodnota test. statistiky	kritická hodnota	Závěr
Pearson	132.21	159.81	Nezamítáme
Roz. test deviance	167.24	159.81	Zamítáme
Hosmer-Lemeshow C_g	11.03	15.51	Nezamítáme
Hosmer-Lemeshow H_g	7.80	15.51	Nezamítáme

Tabulka 5.9: Testy dobré shody pro model3

I v případě tohoto modelu tedy zamítáme hypotézu o shodě pozorovaných a očekávaných hodnot pouze podle rozdílového testu deviance.

Hodnota Kolmogorovova-Smirnovova testu KS = 0.397, odpovídající kritická hodnota je opět 0.233. I v tomto případě tedy zamítáme hypotézu o shodě distribučních funkcí skóre úspěšných a neúspěšných případů.

Konkordantních párů máme 70.24%, diskorantních 29.39% a vázaných 0.37%. Hodnoty koeficientů hodnotících míry asociace mezi pozorovanými a očekávanými hodnotami jsou uvedeny v tabulce 5.10. Tyto hodnoty jsou nižší než hodnoty pro oba předcházející modely.

Koeficient	Hodnota
Kendallovo τ_a	0.206
Kendallovo τ_b	0.291
Somersovo D_s	0.410
Goodmanovo-Kruskalovo γ	0.409

Tabulka 5.10: Míry asociace pro model3

5.5. Výběr modelu

Protože přidání dalších proměnných do modelu nevede k jeho vylepšení, zůstaneme u uvedených tří modelů. V této části práce si vypočítáme další hodnotící statistiky, pomocí kterých uvažované modely porovnáme.

Připomeňme, že nejnižší hodnotu AIC má model3 (173.24) a model2 (173.25). Nejvyšší přesnost má model3 (69.85%) a model1 (69.12%). Naopak nejvyšší hodnotu AUC má model1 (0.723) a model2 (0.715).

Vypočítejme si ještě hodnoty Bayesova informačního kritéria pro dané modely. Příkazem `BIC("název modelu")` zjistíme, že nejnižší hodnotu má opět model3 (181.979) a model2 (181.983), zatímco pro model1 je tato hodnota rovna 182.587 .

Podívejme se na hodnoty koeficientů determinace pro jednotlivé modely (tabulka 5.11). Tyto hodnoty získáme v knihovně `pscl` příkazem `pR2("název modelu")`.

	McFaddenův	Coxové-Snellův	Nagelkerkův
Model1	0.084	0.110	0.146
Model2	0.113	0.145	0.193
Model3	0.113	0.145	0.193

Tabulka 5.11: Koeficienty determinace

Z výsledků vidíme, že nejlépe vychází hodnoty druhého a třetího modelu. I s přihlédnutím k předchozím hodnotícím statistikám tedy můžeme říci, že druhý a třetí model srovnatelně dobře popisují dostupná data.

Shrňme ještě na závěr nalezené modely, které jsou ve tvarech:

$$\text{Model2} : logit(P(Y = 1)) = 2.486 - 0.104 * \text{VĚK} + 0.008 * \text{AMH} * \text{VĚK}.$$

$$\text{Model3} : logit(P(Y = 1)) = 0.452 - 0.141 * \text{FSH} + 0.039 * \text{AMH} * \text{FSH}.$$

V tabulce 5.12 jsou pro přehlednost a snažší porovnání uvedeny některé vlastnosti těchto modelů.

Na základě tabulky 5.12 můžeme říct, že model2 má oproti modelu3 vyšší hodnotu senzitivity a prediktivní hodnotu negativního testu. Jinými slovy, model2 má vysokou úspěšnost ve správné klasifikaci úspěšných případů (jako úspěšná označí 97.1% pozorování ze skutečně úspěšných) a jestliže bylo pozorování klasifikováno jako úspěšné, potom máme pravděpodobnost 61.1%, že je skutečně úspěšné. Dále jestliže bylo pozorování klasifikováno jako neúspěšné, potom pravděpodobnost že je skutečně neúspěšné je 92.9%. Na

	Model2	Model3
Přesnost	0.677	0.699
Senzitivita	0.971	0.853
Specificita	0.382	0.544
Pred. hod. poz. testu	0.611	0.652
Pred. hod. neg. testu	0.929	0.787

Tabulka 5.12: Shrnutí výsledných modelů

druhou stranu je však hodnota specificity tohoto testu pouze 0.382, což znamená, že model označí pouze 38.2% skutečně neúspěšných případů jako neúspěšné. Oproti tomu model3 vykazuje vyšší hodnotu přesnosti, specificity a prediktivní hodnoty pozitivního testu. Tento model označí 54.4% pozorování ze skutečně neúspěšných jako neúspěšná a 78.7% pozorování z modelem označených jako neúspěšné je skutečně neúspěšných. Navíc můžeme říci, že 65.2% modelem klasifikovaných pozorování jako úspěšná jsou skutečně úspěšná a přitom tento model označí jako úspěšná celkem 85.3% pozorování ze skutečně úspěšných.

Z provedené analýzy vyplývá, že druhý a třetí model srovnatelně dobře popisují naše data. Navzájem se však liší v schopnostech správné predikce úspěšných a neúspěšných pozorování. Jestliže by tedy bylo naším cílem správně určit co nejvíce úspěšných pozorování, vybrali bychom si model2. Pokud by nám však záleželo i na správné klasifikaci neúspěšných případů, pak je pro nás vyhodnější model3, který má zároveň dobrou schopnost rozpoznat úspěšná pozorování.

Závěr

Cílem této práce byla analýza vlivu hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodonost. Analýzu jsme provedli na základě reálných dat metodou logistické regrese, se kterou jsme se nejdříve seznámili na teoretické rovině.

V první kapitole jsme si představili princip tvorby logistického regresního modelu a seznámili jsme se s pojmy jako je například šance, logit nebo poměr šancí, které jsou s logistickou regresí spojeny.

V kapitole druhé jsme se věnovali teoretickým postupům hledání odhadů regresních koeficientů. Konkrétně jsme se seznámili s metodou maximální věrohodnosti, která má své využití nejen v logistické regresi. Tato metoda však v případě logistické regrese vyústí v soustavu nelineárních rovnic, které je třeba řešit iteračními metodami. V této práci jsme si představili dvě znich, konkrétně Newtonovu-Raphosonovu metodu a iterativní váženou metodu nejmenších čtverců.

Ve třetí kapitole jsme se věnovali metodám pro výběr vhodného modelu. Probrali jsme možnosti testování hypotéz o podmodelech a seznámili se s některými statistikami, které slouží k porovnání kvality více modelů mezi sebou.

Ve čtvrté kapitole se seznámili se statistikami, které hodnotí model jako takový. Hodnoty těchto statistik nás informují například o tom, jak dobrý model se nám podařilo nalézt nebo jak dobře popisuje a klasifikuje data.

Konečně v poslední (páté) kapitole jsme se věnovali praktické analýze dat. Nejprve jsme se seznámili s datovým souborem a představili si všechny dostupné proměnné. S pomocí nástrojů popisné statistiky jsme si ukázali, jak se jednotlivé proměnné chovají a až poté jsme začali samotnou analýzu.

Nejprve jsme provedli analýzu vlivu hladiny antimülleriánského hormonu na pravděpodobnost otěhotnění u žen léčených na neplodnost včetně výpočtu hodnotících statistik. Poté jsme se pokusili přidáváním dalších dostupných proměnných vytvořit model, který by data popisoval ještě lépe. Podařilo se nám nalézt dva takové modely. První model uvažuje vliv nejen antimülleriánského hormonu, ale také vliv věku ženy. Druhý nalezený model

uvažuje vliv antimülleriánského a folikulostimulačního hormonu. Tyto dva modely jsme na základě vypočtených statistik vyhodnotili jako lepší než model původní. Při pokusu o srovnání těchto dvou modelů navzájem jsme však dospěli k závěru, že oba modely srovnatelně dobře popisují naše data, liší se však ve schopnostech správné klasifikace úspěšných a neúspěšných případů.

Model analyzující vliv věku a hladiny hormonu AMH na pravděpodobnost otěhotnění má velmi dobrou schopnost správně predikovat pozorování jako úspěšná, jestliže jsou opravdu úspěšná. Dalším kladem tohoto modelu je vysoká pravděpodobnost, že bude pozorování neúspěšné, jestliže bylo jako neúspěšné označeno, nicméně velikou nevýhodou tohoto modelu je nízká specificita, tedy schopnost označit jako neúspěch pozorování, které je skutečně neúspěšné. Z toho plyne, že nejen úspěšná pozorování, ale i velká část neúspěšných pozorování jsou klasifikována jako úspěšná a klesá tedy pravděpodobnost, že jestliže bylo pozorování označeno jako úspěšné, bude skutečně úspěšné. Model analyzující vliv hladiny hormonů AMH a FSH na pravděpodobnost otěhotnění má oproti předchozímu modelu horší schopnost označit jako úspěch skutečně úspěšná pozorování, nicméně schopnost správné klasifikace neúspěšných pozorování je vyšší.

Při výběru modelu pro predikci pravděpodobnosti otěhotnění je tedy třeba brát zřetel na výše uvedené vlastnosti obou modelů. Model2 je výhodnější pro případ, kdy chceme především správně klasifikovat úspěšná pozorování nebo chceme mít vysokou pravděpodobnost neúspěchu při klasifikaci pozorování jako neúspěch. Nicméně pro situace, kdybychom chtěli především správně klasifikovat neúspěšná pozorování je vhodnější model3, který zároveň disponuje dobrou schopností správně klasifikovat úspěšná pozorování.

Přílohy

Obrázek 5.6 pravděpodobnost otěhotnění v závislosti na AMH

```
library(dplyr)
newdata$amh_sk <-case_when(amh <=1.5 ~ "1",
                           amh > 1.5 & amh <=3 ~ "2",
                           amh > 3 & amh <=4.5 ~ "3",
                           amh > 4.5& amh <=6 ~ "4",
                           amh > 6 & amh <=7.5 ~ "5",
                           amh > 7.5 & amh <=9 ~ "6",
                           amh > 9 & amh <=10.5 ~ "7",
                           amh > 10.5& amh <=12.5 ~ "8")
attach(newdata)
table(asp,amh_sk)
cetnosti=c(sum(asp[amh_sk==1])/sum(amh_sk==1),
sum(asp[amh_sk==2])/sum(amh_sk==2),
sum(asp[amh_sk==3])/sum(amh_sk==3),
sum(asp[amh_sk==4])/sum(amh_sk==4),
sum(asp[amh_sk==5])/sum(amh_sk==5),
sum(asp[amh_sk==6])/sum(amh_sk==6),
sum(asp[amh_sk==7])/sum(amh_sk==7),
sum(asp[amh_sk==8])/sum(amh_sk==8))
cetnosti
xko=seq(1.5,12,by=1.5)
dat=cbind(cetnosti,xko)
dat1=as.data.frame(dat)
model$coefficients
b0=model$coefficients[1]
b1=model$coefficients[2]
a=exp(b0+1.5*b1)/(1+exp(b0+1.5*b1))
```

```
b=exp(b0+3*b1)/(1+exp(b0+3*b1))
c=exp(b0+4.5*b1)/(1+exp(b0+4.5*b1))
d=exp(b0+6*b1)/(1+exp(b0+6*b1))
e=exp(b0+7.5*b1)/(1+exp(b0+7.5*b1))
f=exp(b0+9*b1)/(1+exp(b0+9*b1))
g=exp(b0+10.5*b1)/(1+exp(b0+10.5*b1))
h=exp(b0+12*b1)/(1+exp(b0+12*b1))
dat1$prob=c(a,b,c,d,e,f,g,h)
attach(dat1)
library(ggplot2)
ggplot( dat1, aes(x=xko, y=cetnosti)) +
  geom_point()+
  geom_line(aes(x=xko,y=prob))+
  labs(y = "Podíl otěhotnění",x="AMH")
```

Křížová validace

```
>library(caret)
>crossValSettings=trainControl(method = "repeatedcv",number=10,
                               savePredictions = TRUE)
>crossval1=train(as.factor(asp)~amh,data=newdata, family="binomial",
 method="glm",trControl=crossValSettings)
>crossval1
```

ROC křivka a hodnota AUC

```
>prob=predict(model1,newdata, type="response")
>prob # odhadnuté psti pro jednotlivá pozorování
>pred=prediction(prob, asp)
>pred # predikce pro poz. na základě odh. pstí
>roc=performance(pred,measure = "tpr",x.measure = "fpr")
>plot(roc, colorize=T,
     ylab="Senzitivita",
     xlab="1-Specificita",
     main="ROC křivka")
>abline(a=0,b=1)
>auc= performance(pred, measure = "auc")
>auc@y.values
```

Volba prahového bodu

```
# Prahový bod maximalizující přesnost
>acc=performance(pred,"acc")
>ac.val=max(unlist(acc@y.values))
>ac.val # přesnost
>cut=unlist(acc@x.values)[unlist(acc@y.values)==ac.val]
>cut # prahový bod
# Grafické znázornění
>plot(acc, xlab="Prahový bod",ylab="Přesnost")
>abline(v=cut, col='blue',lty=2)

# Prahový bod - senzitivita a specificita vyvážené
>opt.cut = function(perf, pred){
  cut.ind = mapply(FUN=function(x, y, p){
    d = (x - 0)^2 + (y-1)^2
    ind = which(d == min(d))
    c(sensitivity = y[[ind]], specificity = 1-x[[ind]],
      cutoff = p[[ind]])
  }, perf@x.values, perf@y.values, pred@cutoffs)
}
>print(opt.cut(roc, pred))
```

Obrázek 5.9 odhady pro pravděpodobnost otěhotnění

```
newdata$asp=factor(newdata$asp)
library(ggplot2)
ggplot(newdata, aes(amh, prob)) +
  geom_point(aes(color = asp))+
  scale_color_manual(breaks = c("0", "1"), values=c("red", "blue"))+
  labs(y = "Skóre",x="AMH")+
  geom_hline(yintercept =cutOff, linetype="dashed", color = "black")+
  annotate(geom="text", label="Prahový bod", x=10, y=cutOff, vjust=-1)
```

Testy dobré shody

```
# Pearsonův test dobré shody
>si=fitted.values(model1)
```

```
>P=sum(((asp-si)^2)/(si*(1-si)))
>m=length(model1\$coefficients)
>n=length(asp)
>p=n-m-1
>qchisq(0.95,p) # Krit. hodnota

# Rozdílový test deviance
>D=-2*(sum((si*log(si/(1-si)))+log(1-si)))
>qchisq(0.95,p)

# Hosmer-Lemeshow
>install.packages("MKmisc")
>library(MKmisc)
>HLgof.test(fit = fitted(model1), obs = newdata$asp)
```

Kolmogorovův-Smirnovův test dobré shody

```
KS=max(roc@y.values[[1]]-roc@x.values[[1]])
KS
n1=sum(asp==1)
n0=sum(asp==0)
1.36*(sqrt((n1+n0)/(n1*n0))) # krit. hodnota->zamítáme
```

Míry asociace modelu

```
OptimisedConc=function(model)
{
  Data = cbind(model$y, model$fitted.values)
  ones = Data[Data[,1] == 1,]
  zeros = Data[Data[,1] == 0,]
  conc=matrix(0, dim(zeros)[1], dim(ones)[1])
  disc=matrix(0, dim(zeros)[1], dim(ones)[1])
  ties=matrix(0, dim(zeros)[1], dim(ones)[1])
  for (j in 1:dim(zeros)[1])
  {
    for (i in 1:dim(ones)[1])
    {
      if (ones[i,2]>zeros[j,2])
      {conc[j,i]=1}
```

```
      else if (ones[i,2]<zeros[j,2])
      {disc[j,i]=1}
      else if (ones[i,2]==zeros[j,2])
      {ties[j,i]=1}
    }
  }
  Pairs=dim(zeros)[1]*dim(ones)[1]
  PercentConcordance=(sum(conc)/Pairs)*100
  PercentDiscordance=(sum(disc)/Pairs)*100
  PercentTied=(sum(ties)/Pairs)*100
  PercentConcordance=(sum(conc)/Pairs)*100
PercentDiscordance=(sum(disc)/Pairs)*100
PercentTied=(sum(ties)/Pairs)*100
N<-length(model$y)
gamma<-(sum(conc)-sum(disc))/Pairs
Somers_D<-(sum(conc)-sum(disc))/(Pairs-sum(ties))
k_tau_a<-2*(sum(conc)-sum(disc))/(N*(N-1))
return(list("Percent Concordance"=PercentConcordance,
            "Percent Discordance"=PercentDiscordance,
            "Percent Tied"=PercentTied,
            "Pairs"=Pairs,
            "Gamma"=gamma,
            "Somers D"=Somers_D,
            "Kendall's Tau A"=k_tau_a))
  return(list("Percent Concordance"=PercentConcordance,
"Percent Discordance"=PercentDiscordance,
"Percent Tied"=PercentTied,
"Pairs"=Pairs))
}
OptimisedConc(model1)
```

Obrázek 5.13 Graf závislosti skóre a hladiny AMH s ohledem na věk

```
model2$coef
b0 <- model2$coef[1]
b1 <- model2$coef[2]
b2 <- model2$coef[3]
```

```
prob=fitted.values(model2)
mean(vek)
mean(vek)+sd(vek)
mean(vek)-sd(vek)
# do grafu chceme přidat:
#########################
# 1. křivku pro vek=30 (mean)
# 2. křivku pro vek=25 (mean-sd)
# 3. křivku pro vek=36 (mean+sd)

# výpočet:
# 1. zafix. věk=30 -> logit(P(asp=1))=b0+b1*30+b2*30*amh
# -> P(asp=1|vek=0)=exp(logit)/(1+exp(logit))
logit1=b0+(b1*30)+(b2*(30*amh))
logit1
pst1=(exp(logit1))/((1+exp(logit1)))
pst1

# 2. zafix. věk=25 -> logit(P(asp=1))=b0+b1*25+b2*25*amh
# -> P(asp=1|vek=0)=exp(logit)/(1+exp(logit))
logit2=b0+(b1*25)+(b2*(25*amh))
logit2
pst2=(exp(logit2))/((1+exp(logit2)))
pst2

# 3. zafix. věk=36 -> logit(P(asp=1))=b0+b1*36+b2*36*amh
# -> P(asp=1|vek=0)=exp(logit)/(1+exp(logit))
logit3=b0+(b1*36)+(b2*(36*amh))
logit3
pst3=(exp(logit3))/((1+exp(logit3)))
pst3


## výsledný graf
ggplot(newdata,aes(x=amh,y=prob,color=vek))+
  geom_point()+
  geom_point(aes(y=asp))+
```

```
scale_color_continuous(low="green",high="red")+
geom_line(aes(x=amh,y=pst1),colour="orange")+
geom_line(aes(x=amh,y=pst2),colour="green")+
geom_line(aes(x=amh,y=pst3),colour="red")+
geom_text(aes(x=0.8,y=0.56,label="mean-sd"),color="green")+
geom_text(aes(x=0.6,y=0.33,label="mean"),color="orange")+
geom_text(aes(x=0.9,y=0.15,label="mean+sd"),color="red")+
xlab("AMH") + ylab("Skóre")
```

Literatura

[1] Agresti, A.: *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc., 2002.

[2] Agresti, A.: *Foundations of Linear and Generalized Linear Models.* Hoboken, New Jersey: John Wiley & Sons, Inc., 2015.

[3] Allison, P.: *What's the Best R-Squared for Logistic Regression?* [online]. 2013, [cit. 2018-01-20]. Dostupné z: https://statisticalhorizons.com/r2logistic.

[4] Bortlíček, Z.: *ROC křivky.* Brno, 2008. Diplomová práce. Masarykova univerzita v Brně, Přírodovědecká fakulta. Dostupné z: https://is.muni.cz/th/jwg4n/Diplomova_prace.pdf.

[5] Chang, W.: *Multiple graphs on one page (ggplot2).* [online]. [cit. 2017-11-24]. Dostupné z: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/.

[6] Cox, D. R.: *The regression analysis of binary sequences.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 20, No. 2 (1958), p. 215-242.

[7] Czepiel, S.A.: *Maximum likelihood estimation of logistic regression models: theory and implementation.* [online]. 2002, [cit. 2018-01-10]. Dostupné z: https://czep.net/stat/mlelr.pdf.

[8] Dobson, A.J.: *Introduction to generalized linear models, Second edition.* Chapman & Hall/CRC, 2002.

[9] Hallet, D.C.: *Goodness of fit tests in logistic regression.* National Library of Canada, 1999.

[10] Holčák, L.: *Statistická analýza souborů s malým rozsahem.* Brno, 2008. Diplomová práce. Vysoké učení technické v Brně, Fakulta strojního inženýrství. Dostupné z: https://dspace.vutbr.cz/handle/11012/25225.

[11] Horová, I.; Zelinka, J.: *Numerické metody.* 2. rozš. vyd. Praha: Masarykova univerzita, 2008 dotisk.

[12] Hron, K.; Kunderová, P.: *Základy počtu pravděpodobnosti a metod matematické statistiky.* Univerzita Palackého v Olomouci, Přírodovědecká fakulta, Olomouc. 2015.

[13] IMALAB s.r.o.. Imunochemie. *FSH: folikulostimulační hormon.* [online]. ©2009, [cit. 2017-11-21]. Dostupné z: http://www.imalab.cz/clanek/184-fsh-folikulostimulacni-hormon.aspx.

[14] Meloun, M.: *Základy logistické regrese.* Sborník celostátního semináře Analýza dat 2006/II (Pokročilé statistické metody pro praxi), Lázně Bohdaneč 24. - 27. 10. 2006, Trilobyte statistical software, Pardubice, 2007. p. 79 - 90.

[15] Menard, S.: *Logistic Regression: From Introductory to Advanced Concepts and Applications.* Sam Houston State University, SAGE. 2010.

[16] Minka, T.P.: *Algorithms for maximum-likelihood logistic regression.* [online]. 2001, [cit. 2018-01-06]. Dostupné z: https://tminka.github.io/papers/logreg/.

[17] Ondrušková, M.: *Odhadování a kritéria těsnosti modelu logistické regrese.* Praha, 2011. Bakalářská práce. Univerzita Karlova v Praze, Matematicko-fyzikální fakulta. Dostupné z: https://dspace.cuni.cz/handle/20.500.11956/38622?show=full.

[18] Pecáková, I.: *Logistická regrese s vícekategoriální vysvětlovanou proměnnou.* Acta Oeconomica Pragensia (2007). 15(1), p. 86-96.

[19] Procházka, B.: *Biostatistika pro lékaře: Principy základních metod a jejich interpretace s využitím statistického systému R.* Charles University in Prague, Karolinum Press, 2015.

[20] R-bloggers. *A small introduction to the ROCR package.* [online]. 2014, [cit. 2017-1-15]. Dostupné z: https://www.r-bloggers.com/a-small-introduction-to-the-rocr-package/.

[21] Shashiasrblog. Blog [R]. *Binary logistic Regression on R : Concordance and Discordance.* [online]. 2014, [cit. 2017-2-16]. Dostupné z: http://shashiasrblog.blogspot.cz/2014/01/binary-logistic-regression-on-r.html.

[22] Spěváková, D.: *Biologické hodiny z pohledu anti-müllerian hormonu.* Zlín, 2015. Bakalářská práce. Univerzita Tomáše Bati ve Zlíně, Fakulta humanitních studií. Dostupné z: http://digilib.k.utb.cz/handle/10563/33232.

[23] Wood, S.N.: *Generalized additive models: An introduction with R.* Chapman & Hall/CRC, 2006.

[24] Zou, K.H.; O'Malley, A.J.: *Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models.* Circulation, 2007, 115.5: p. 654-657. Dostupné z: http://circ.ahajournals.org/content/115/5/654.figures-only/.