ZLIM - Zápočtová úloha č. 6

Miroslav Kubů

12. srpna 2019

1 Úvod

K dispozici máme datovou množinu pocházející ze studie 189 rodiček z určité nemocnice. Každé pozorování je tvořeno 7 údaji o příslušné matce.

- „BIRTH“ - o kolikátý porod dané ženy se jedná
- „SMOKE“ - zda žena kouřila během těhotenství
- „RACE“ – rasa (White, Black, Other)
- „AGE“ - věk matky
- „LWT“ – váha matky v poslední menstruační periodě (v librách)
- „BWT“ – porodní váha dítěte
- „LOW“ – indikátor jevu, že porodní váha je menší než 2500 g

Naší snahou je s pomocí zadaných údajů vytvořit model pro proměnnou LOW. Tvoříme tedy model pro binární statistickou klasifikaci, s jehož využitím lze na základě příslušných údajů predikovat nízkou porodní váhu dítěte. Z povahy úlohy tedy pro predikci nízké porodní hmotnosti sestrojíme model logistické regrese. Následně provádíme detailní analýzu a vyhodnocení zvoleného modelu.

2 Deskriptivní analýza dat

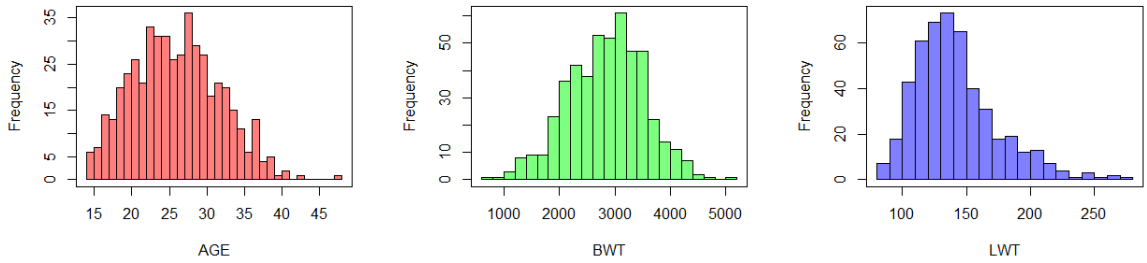
Celkově máme k dispozici 488 pozorování o 7 výše vypsaných proměnných. Pro všechna pozorování jsou vždy k dispozici všechny údaje. Proměnné LOW, BIRTH, SMOKE a RACE uvažujeme jako faktorové. Ostatní proměnné považujeme za spojité, přestože i proměnnou AGE by přísně vzato šlo považovat za faktorovou.

V rámci tabulky 1 prezentujeme vybrané statistiky pro proměnné AGE, LWT a BWT. Z tabulky lze vidět, že pracujeme s rodičkami mezi 14 a 48 lety s průměrným věkem 26,44 let. Výrazné rozdíly jsou poté patrné i pro proměnnou BWT, kde máme k dispozici údaje o dětech s hmotnosti od 798 do 5025 gramu. Podrobnější vhled do údajů AGE, BWT a LWT získáme též z vykreslení histogramů pro

	AGE	LWT	BWT
Min	14,00	80	798
Průměr	26,44	142,8	2842
Medián	26,00	137,0	2883
Max	48,00	272	5025

Tabulka 1: Vybrané statistiky pro proměnné AGE, LWT a BWT.

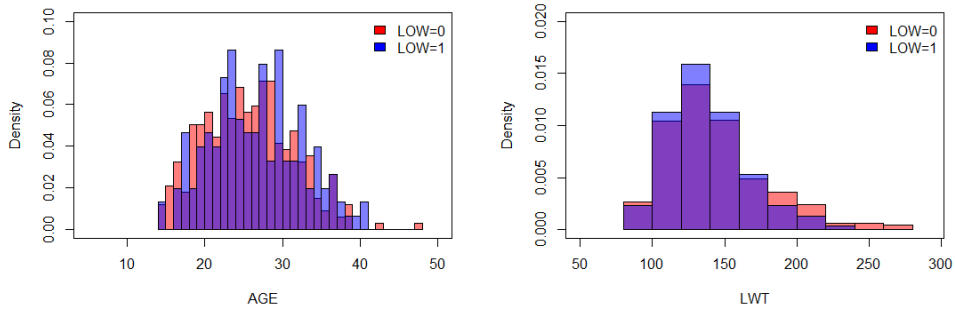
tyto proměnné. Ty jsou vyobrazené na obrázku 1. Z obrázku je patrné, že pro AGE > 45 je již počet rodiček velice malý. Naopak disponujeme poměrně velkým počtem nezletilých rodiček. Poměrně zajímavé je také vykreslení histogramu pro proměnnou LWT. Ta se totiž prezentuje těžkým pravým chvostem odpovídajícím vysokému počtu a spektru rodiček s nadprůměrnou hodnotou LWT. Dále pokračujeme ve vykreslení histogramů s indikací třídy LOW pro proměnné AGE a LWT, jež je zobrazeno na obrázku 2. Na první pohled je zřejmé, že třída LOW $= 1$ dětí s nízkou hmotností je výrazně menší co se mohutnosti týče. V rámci proměnné AGE vidíme výrazně větší zastoupení třídy LOW $= 1$ pro AGE $= 30$ let a dále pro rodičky s věkem okolo 35 let. Přesto není na první pohled zřejmý jasný trend pro rozdělení obou



(a) rozdělení AGE (b) rozdělení BWT (c) rozdělení LWT

Obrázek 1: Vykreslení rozdělení pro proměnné AGE, BWT a LWT. Histogram pro proměnnou AGE byl vytvořen pomocí binů šířkou odpovídajícím jednomu roku. Pro proměnné BWT a LWT byl použit fixní počet 20 binů.

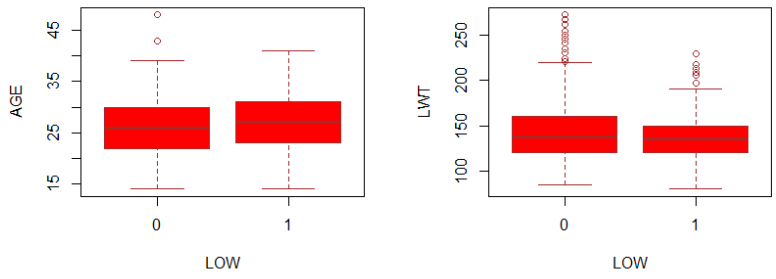
tříd. Pro proměnnou LWT jsou tvary obou histogramů velice podobné. Pozorujeme nicméně, že rodičky s proměnnou LWT > 250 již rodily pouze děti se zdravou porodní vahou. To nicméně nemusí být obecný trend, jelikož tento efekt pozorujeme jen na několika málo pozorováních.



(a) rozdělení AGE (b) rozdělení LWT

Obrázek 2: Histogramy pro proměnné AGE a LWT s indikací třídy LOW.

Na obrázku 3 následně vykreslujeme boxploty pro proměnné AGE a LWT v závislosti na indikaci třídy LOW. Z boxplotů není patrný výrazný rozdíl ve střední hodnotě a příslušných kvantilech pro proměnnou AGE. Pozorování dvou nejstarších rodiček s AGE>45 a LOW $= 0$ byla poté v boxplotech označena za odlehlá. Co se proměnné LWT týče, je patré, že pravý chvost pro rozdělení s LOW $= 0$ je těžší než pro LOW $= 1$.



(a) AGE (b) LWT

Obrázek 3: Boxploty pro proměnné AGE a LWT s indikacemi třídy LOW.

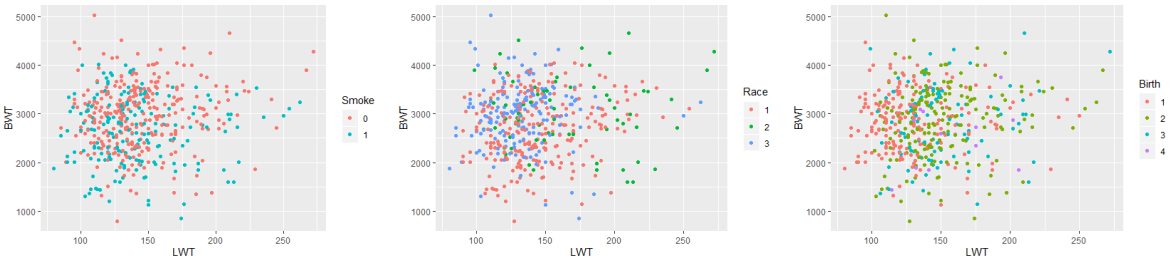
Vzhledem k tomu, že se na porodní váhu nemusíme dívat jen jako na faktorovou proměnnou LOW,

ale též jako na spojitou proměnnou BWT udávající přímo porodní váhu může být prospěšné vykreslit scatter ploty BWT vůči zbylým spojitým proměnným AGE a LWT. Ze scatter plotů pro proměnné BWT a AGE ilustrovaných na obrázku 4 na první pohled nelze pozorovat jasně daná závislost mezi oběma proměnnými. Je nicméně patrná nižší porodní hmotnost mezi kuřačkami a též očekávaná závislost mezi věkem a počtem porodů. Scatter ploty pro proměnné BWT a LWT jsou ilustrovány na obrázku 5. Co se

(a) SMOKE (b) RACE (c) BIRTH

Obrázek 4: Scatter ploty pro proměnné BWT a AGE s indikacemi příslušných kategorií.

samotné závislosti BWT na LWT týče, je lehce patrný rostoucí trend s nárůstem LWT. Zároveň jsou opět výrzně patrné nižší porodní váhy pro kuřačky a bělošky. Ze scatter plotů nicméně můžeme vyčíst i obecně četnější zastoupení černošek a žen s vyšším počtem porodů mezi rodičkami s vyšší hodnotou LWT. Dále



(a) SMOKE (b) RACE (c) BIRTH

Obrázek 5: Scatter ploty pro proměnné BWT a LWT s indikacemi příslušných kategorií.

budeme analyzovat kategorické proměnné a prozkoumáme zastoupení dětí s nízkou porodní vahou napříč jednotlivými kategoriemi. Nejdříve se zaměříme na proměnnou SMOKE. Jednotlivá zastoupení kategorií SMOKE pro LOW jsou uvedena v tabulce 2. Jak dále ukazuje i obrázek 6, mezi kuřáky je pozorovatelně vyšší zastoupení dětí narozených s nízkou vahou.

	Non-Smoker	Smoker	Total
Low = 0	223 (76,1%)	114 (58,5%)	337
Low = 1	70 (23,9%)	81 (41,5%)	151
Total	293	195	488

Tabulka 2: Zastoupení kategorií SMOKE ve třídách LOW.

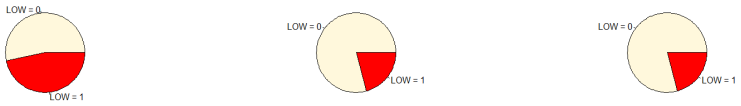
(a) Non-Smoker (b) Smoker

Obrázek 6: Koláčové diagramy pro proměnnou SMOKE s indikací LOW.

Co se proměnné RACE týče, z tabulky 3 a obrázku 7 je patrné, že ačkoliv mezi kategoriemi Black a Other nejsou výrazné rozdíly v zastoupení tříd LOW, u kategorie White je zastoupení LOW = 1 výrazně vyšší. Z tabulky 3 je poté též patrné, že máme k dispozici výrazně méně údajů o kategorii Black nežli o kategoriích White a Other. V rámci přehledu o třídách proměnné BIRTH uvedeného v tabulce 4 a

	White	Black	Other	Total
Low = 0	144 (59,0%)	57 (79,2%)	136 (79,1%)	337
Low = 1	100 (41,0%)	15 (20,8%)	36 (20,9%)	151
Total	244	72	172	488

Tabulka 3: Zastoupení kategorií RACE ve třídách LOW.



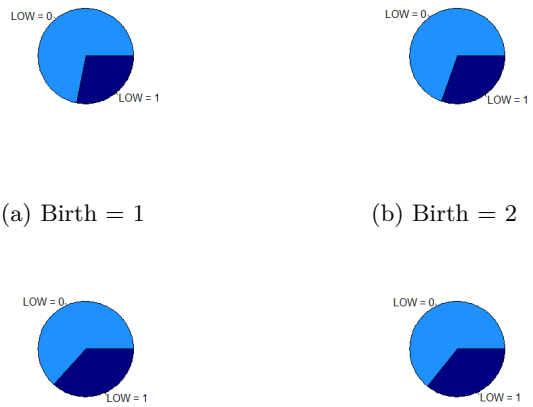
(a) White (b) Black (c) Other

Obrázek 7: Koláčové diagramy pro proměnnou RACE s indikací LOW.

ilustrovaného a obrázku 8 je patrný klesající trend v počtu porodů a rostoucí trend v zastoupení dětí s nízkou porodní vahou s přibývajícími porody. To může zřejmě souviset též s vyšším věkem rodiček s vyšším počtem porodů.

	Birth = 1	Birth = 2	Birth = 3	Birth = 4	Total
Low = 0	135 (71,8%)	131 (69,7%)	62 (63,6%)	9 (64,3%)	337
Low = 1	53 (28,2%)	57 (30,3%)	36 (36,4%)	5 (35,7%)	151
Total	188	188	98	14	488

Tabulka 4: Zastoupení kategorií RACE ve třídách LOW.

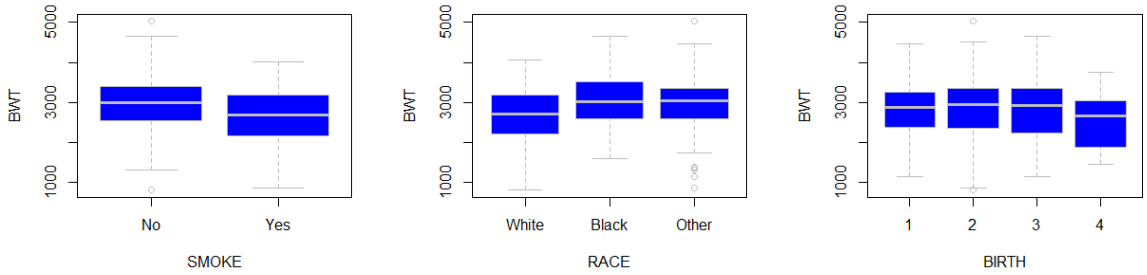


(a) Birth = 1 (b) Birth = 2

(c) Birth = 3 (d) Birth = 4

Obrázek 8: Koláčové diagramy pro proměnnou BIRTH s indikací LOW.

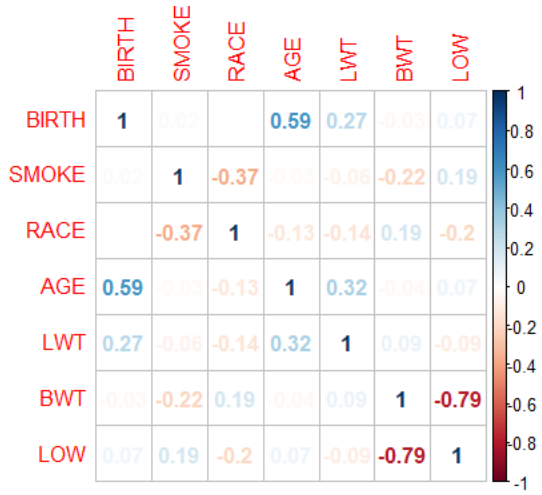
Další grafickou analýzu faktorových proměnných SMOKE, RACE a BIRTH provádíme pomocí boxplotů na obrázku 9 vykreslených vůči proměnné BWT. Na obrázku 9 jsou vidět podobné trendy jako v předchozích koláčových diagramech. Medián porodní váhy pro SMOKE = 1, RACE = White a Birth = 4 je pozorovatelně nižší nežli u dalších kategorií.



(a) SMOKE (b) RACE (c) BIRTH

Obrázek 9: Přehled boxplotů pro faktorové proměnné proti proměnné BWT.

Na závěr na obrázku 10 vykreslujeme ilustraci hodnot korelační matice pro všechny dostupné proměnné. Z obrázku je na první pohled zřejmá závislost mezi veličinami BWT a LOW. To nicméně odpovídá tomu, že LOW je pouze indikace pro překročení hodnoty BWT danou mez. Dále pozorujeme vztah mezi proměnnými BIRTH a AGE. Proměnná BIRTH přitom odpovídá počtu porodů, tedy od prvorodiček skutečně očekáváme nižší věk než u rodiček s vyšším počtem porodů. U dalších proměnných poté již nepozorujeme vysoké hodnoty korelační koeficientů.



Obrázek 10: Ilustrace korelační matice pro všechny dostupné proměnné.

3 Model logistické regrese

V předchozí sekci jsme díky numerické a grafické analýze dat získali elementární představu o významu jednotlivých proměnných pro predikci nízké porodní váhy. V této sekci následně konstruujeme model logistické regrese, s jehož pomocí lze následně ze zadaných údajů predikovat pravděpodobnost nízké porodní hmotnosti. V prvním stádiu zkonstruujeme maximální model, tedy model obsahující všechny dostupné proměnné. Následně se maximální model pokusíme odebráním statisticky nevýznamných proměnných zjednodušit s cílem co nejvíce zachovat predikční vlastnosti z maximálního modelu. Tuto proceduru opakujeme pro tři vybrané spojovací funkce, a následně z nich vybereme nejvhodnější model.

3.1 Spojovací funkce logit

Použijeme-li pro konstrukci zobecněného lineárního modelu spojovací funkci logit, získáme odhady parametrů příslušného maximálního modelu uvedené v tabulce 5. Z tabulky je zřejmé, že proměnné AGE a BIRTH jsou na základě p hodnot při standardní hladině významnosti $\alpha = 0,05$ nevýznamné. Z tohoto důvodu provádíme zpětnou eliminaci proměnných na základě kritéria AIC, čímž získáváme model ve

	Estimate	Std. Error	p-value
Intercept	0,015	0,720	0,982
AGE	0,021	0,023	0,347
BIRTH = 2	0,232	0,260	0,371
BIRTH = 3	0,450	0,336	0,181
BIRTH = 4	0,343	0,681	0,614
SMOKE = 1	0,508	0,223	0,023
RACE = Black	-0,727	0,334	0,030
RACE = Other	-0,855	0,256	0,001
LWT	-0,010	0,004	0,009

Tabulka 5: Odhady parametrů s příslušnými statistikami včetně p hodnot.

	p-value
Intercept	0,5856
AGE	0,0474
SMOKE = 1	0,0142
RACE = Black	0,0363
RACE = Other	0,0014
LWT	0,0123

Tabulka 6: Přehled p hodnot pro významnost jednotlivých proměnných finálního modelu se spojovací funkcí logit.

tvaru

$$\log\left(\frac{p}{1-p}\right) = -0,354 + 0,037 \cdot \text{AGE} + 0,541 \cdot \text{Smoker} - 0,692 \cdot \text{Black} - 0,806 \cdot \text{Other} - 0,01 \cdot \text{LWT}, \quad (1)$$

kde $p = \text{P}\left(\text{LOW} = 1 | \, data\right)$ značí pravděpodobnost příslušnosti k třídě LOW = 1 při zadaných parametrech. Podíváme-li se na výsledné p hodnoty pro model popsaný rovnicí (1) uvedené v tabulce 6, vidíme, že všechny proměnné jsou nyní při hladině významnosti $\alpha = 0,05$ statisticky významné. Model (1) tedy odpovídá našim představám z deskriptivní analýzy ohledně nižší náchylnosti pro narození dítěte s nízkou hmotností mezi nekuřačkami, rodičkami s rasou Black či Other a rodičkami s vyšší hodnotou LWT.

3.2 Spojovací funkce probit

Dle zadání následně analogicky sestrojíme zobecněné lineární modely též pro spojovací funkce probit a cloglog. Podobně jako v předchozím případě nejdříve konstruujeme maximální modely, ze kterých následně odebíráme statisticky nevýznamné proměnné. Pro spojovací funkci probit tímto způsobem získáváme model v podobě

$$p = \Phi\left(0,175 + 0,326 \cdot \text{Smoker} - 0,471 \cdot \text{Black} - 0,502 \cdot \text{Other} - 0,004 \cdot \text{LWT}\right), \quad (2)$$

kde Φ značí distribuční funkci pro normální rozdělení $\mathcal{N}(0,1)$. Tabulku příslušných p hodnot pro statistickou významnost jednotlivých proměnných uvádíme v tabulce 7. Oproti modelu (1) je model (2) zajímavý především absencí vysvětlující proměnné AGE. Ta se totiž při hladině významnosti $\alpha = 0,05$ ukázala býti statisticky nevýznamná.

	p-value
Intercept	0,5891
SMOKE = 1	0,0136
RACE = Black	0,0135
RACE = Other	0,0007
LWT	0,0439

Tabulka 7: Přehled p hodnot pro významnost jednotlivých proměnných finálního modelu se spojovací funkcí probit.

3.3 Spojovací funkce cloglog

Pro model se spojovací funkcí cloglog po odstranění statisticky nevýznamných proměnných z maximálního modelu získáváme model ve tvaru

$$\log\left(-\log(1-p)\right) = -0,680 + 0,031 \cdot \text{AGE} + 0,421 \cdot \text{Smoker} - 0,557 \cdot \text{Black} - 0,663 \cdot \text{Other} - 0,008 \cdot \text{LWT}. \quad (3)$$

Přehled p hodnot pro statistickou významnost proměnných poté uvádíme v tabulce 8.

	p-value
Intercept	0,2017
AGE	0,0392
SMOKE = 1	0,0194
RACE = Black	0,0485
RACE = Other	0,0020
LWT	0,0152

Tabulka 8: Přehled p hodnot pro významnost jednotlivých proměnných finálního modelu se spojovací funkcí cloglog.

3.4 Porovnání modelů

Aktuálně máme k dispozici tři modely pro různé spojovací funkce, které ze zadaných dat predikují pravděpodobnost narození dítěte s nízkou hmotností. Modely se shodují ve znamínkách koeficientů příslušných jednotlivým proměnným. Jediný významný rozdíl je patrný u modelu se spojovací funkcí probit, ve kterém nefiguruje proměnná AGE. Podíváme-li se na srovnání informačních kritérií AIC,BIC a deviace v tabulce 9, zjistíme, že všechny tři modely popisují data prakticky stejně dobře. Model se spojovací funkcí logit je nicméně z daných modelů nejjednodušší co se interpretace týče, a proto pro predikci volíme právě tento model.

	AIC	BIC	Deviace
logit	577	602	565,02
probit	579	600	568,69
cloglog	578	602	565,59

Tabulka 9: Přehled hodnot informačních kritérií pro jednotlivé modely.

4 Diagnostika modelu

4.1 Analýza koeficientů modelu

V rámci této sekce provedeme detailní diagnostiku námi zvoleného modelu logistické regrese. Analyzujeme tedy model s logitovou spojovací funkcí ve tvaru

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{Smoker} + \beta_3 \cdot \text{Black} + \beta_4 \cdot \text{Other} + \beta_5 \cdot \text{LWT}, \quad (4)$$

kde odhady jednotlivých koeficientů i s 95% intervaly spolehlivosti uvádíme v tabulce 10. Deviace ma-

	odhad	95% interval spolehlivosti
β_0	-0,354	(-1,624, 0,924)
β_1	0,037	(0,001, 0,075)
β_2	0,541	(0,109, 0,975)
β_3	-0,692	(-1,367, -0,063)
β_4	-0,806	(-1,309, -0,317)
β_5	-0,01	(-0,017, -0,002)

Tabulka 10: Odhady koeficientů modelu logistické regrese s intervaly spolehlivosti.

ximálního modelu činila $D_S = 563,18$. V případě námi vybraného modelu očištěného o statisticky nevýznamné proměnné jsme dosáhli hodnoty deviační statistiky $D = 565,02$. To znamená, že náš model popisuje data téměř stejně dobře jako příslušný maximální model. Shodu dat s modelem lze též testovat pomocí tzv. Hosmer-Lemeshow testu. Nulová hypotéze H_0 Hosmer-Lemeshow testu odpovídá dobré shodě modelu s daty. Výsledná p hodnota Hosmer-Lemeshow testu nicméně činí $p = 0,125$. Při hladině významnosti $\alpha = 0,05$ tak nelze zamítnout dobrou shodu dat s modelem.

Významy jednotlivých vysvětlujících proměnných pro nízkou porodní hmotnost lze zkoumat též za použití odds ratio (OR). Pro hodnoty OR > 1 je příslušná proměnná pozitivní faktor pro třídu LOW $= 1$, v případě OR < 1 je příslušná proměnná faktor protektivní. Přehled OR pro jednotlivé proměnné i s intervaly spolehlivosti uvádíme v tabulce 11. Z té je zřejmé, že věk a kouření bylo vyhodnoceno jako pozitivní faktor pro nízkou porodní hmotnost i v rámci intervalů spolehlivosti.

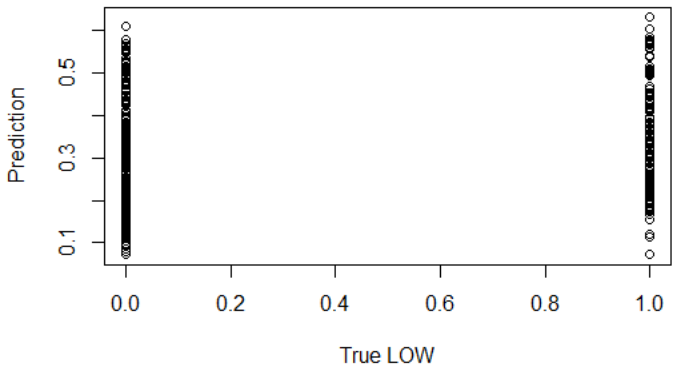
	odds ratio	95% interval spolehlivosti
Intercept	0,702	(0,197, 2,518)
AGE	1,038	(1,000, 1,078)
SMOKE=1	1,718	(1,115, 2,651)
RACE=Black	0,500	(0,255, 0,939)
RACE=Other	0,447	(0,270, 0,728)
LWT	0,991	(0,983, 0,998)

Tabulka 11: Přehled OR pro proměnné z vybraného modelu.

4.2 Analýza predikce modelu

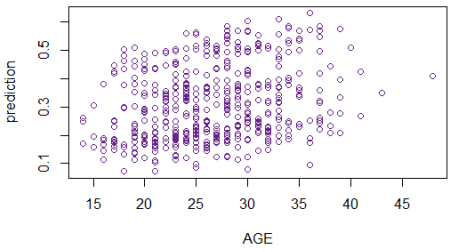
V dalším kroku prozkoumáme samotné predikční vlastnosti modelu. V první fázi na obrázku 11 vykreslujeme hodnoty predikovaných pravděpodobností vůči skutečným indikacím třídy LOW. Predikované pravděpodobnosti u pozorování třídy LOW $= 0$ se koncentrují pod hranicí 0,5. To odpovídá dobré předpovědi pro třídu LOW $= 0$. Na druhou stranu pozorování třídy LOW $= 1$ by se v ideálním případě měla koncentrovat nad hranicí 0,5, což nepozorujeme. To znamená, že může být těžké obě dvě třídy oddělit. V pozdější fázi tak bude nutné zvolit hraniční hodnotu $c \in (0,1)$ určující pro predikci $p \in (0,1)$ odhad příslušnosti k třídě LOW způsobem

$$\hat{\text{LOW}} = \begin{cases} 1 \text{ pro } p > c. \\ 0 \text{ pro } p \leq c. \end{cases} \tag{5}$$

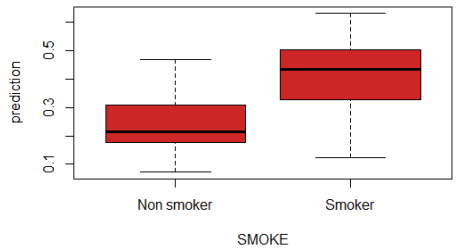


Obrázek 11: Vykreslení predikcí modelu oproti indikacím třídy LOW.

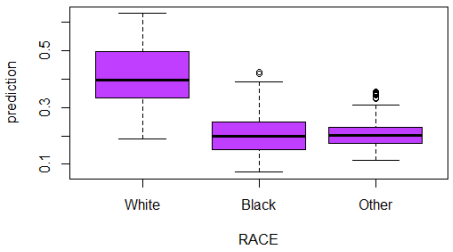
Dále vůči predikcím vykreslujeme též vybrané vysvětlující proměnné. Na obrázku 12 tak můžeme pozorovat, že s rostoucím věkem se zvyšuje pravděpodobnost nízké porodní hmotnosti. Podobně poté pozorujeme, že pro kuřačky a matky bílé rasy jsou predikovány pozorovatelně vyšší hodnoty predikcí. To je v souladu s analýzou dat ze sekce 1.



(a) AGE



(b) SMOKE

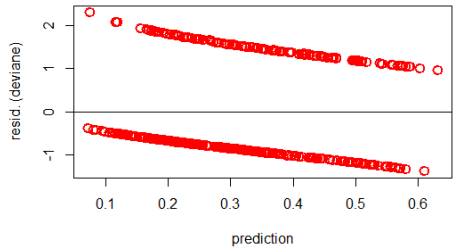


(c) RACE

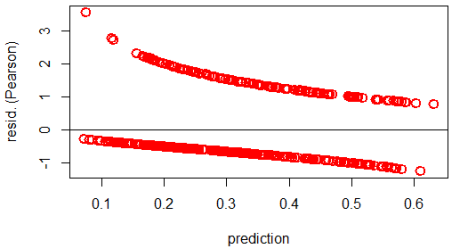
Obrázek 12: Vykreslení proměnných AGE, SMOKE a RACE vůči predikovaným pravděpodobnostem příslušnosti ke třídě LOW = 1.

4.3 Analýza reziduí

Budeme-li zkoumat rezidua při zvoleném modelu, z obrázku 13 pozorujeme mírně klesající hodnoty reziduí pro vyšší hodnoty predikcí.



(a) deviační rezidua



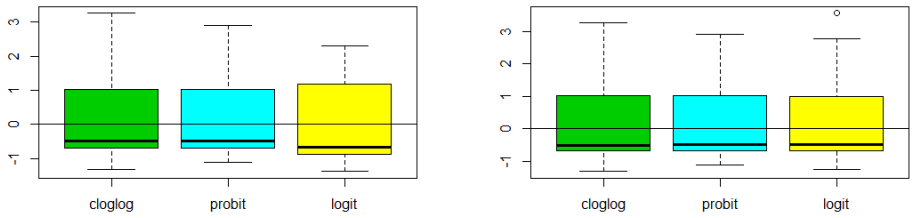
(b) Pearsonova rezidua

Obrázek 13: Vykreslení reziduí vůči predikcím

Dále na obrázku 14 vykreslujeme boxplot pro hodnoty reziduí. V ideálním případě bychom totiž chtěli, aby se hodnoty reziduí koncentrovaly okolo 0. V našem případě se hodnoty koncentrují okolo hodnoty nižší než 0, nicméně jak obrázek ukazuje, tento problém nelze odstranit ani volbou jiného modelu.

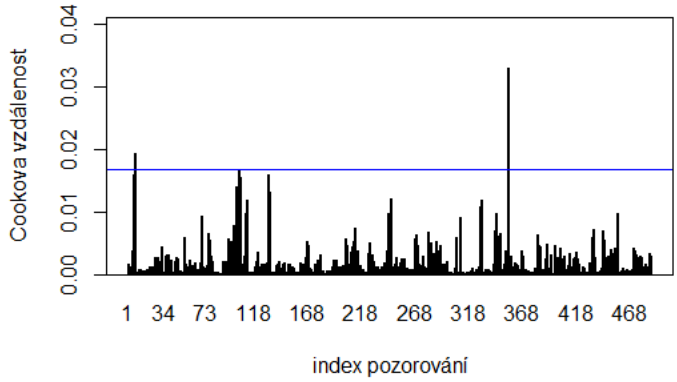
4.4 Influenční a pákové body

V dalším bodě se zaměříme na identifikaci podezřelých pozorováních pomocí Cookovy vzdálenosti a tzv. hat values. Na obrázku 15 vykreslujeme hodnoty Cookovy vzdálenosti pro jednotlivá pozorování se zvýrazněním hraniční hodnoty, od které považujeme pozorování za influenční. Hraniční hodnotu v tomto případě výrazně přesahuje pozorování s indexem 355. To je tvořeno nekuřačkou rasy Other s nízkým věkem a vysokou hodnotou LWT, které se nicméně navzdory všem protektivním faktorům narodilo dítě s nízkou porodní vahou. Z tohoto důvodu tak bylo pozorování vyhodnoceno jako influenční. Na druhou stranu, co se týče teoretických poznatků, neshledáváme nutné pozorování z modelu odstranit.



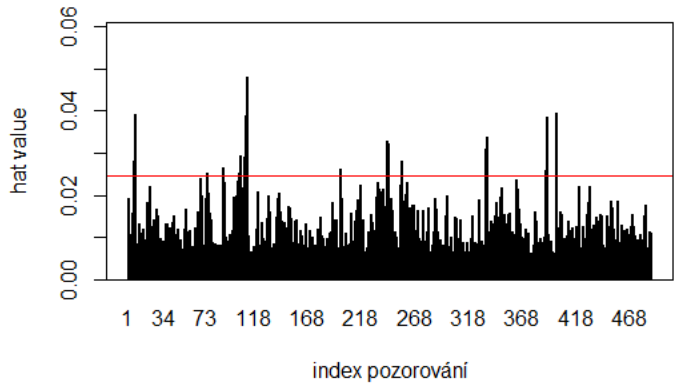
(a) SMOKE (b) RACE

Obrázek 14: Přehled boxplotů pro faktorové proměnné proti proměnné BWT.



Obrázek 15: Přehled Cookovy vzdálenosti pro jednotlivé indexy pozorování.

Na obrázku 16 poté analogicky vykreslujeme hodnoty hat values pro jednotlivé indexy pozorování opět s hraniční hodnotou. V tomto případě detekujeme výrazně více podezřelých pozorování, přičemž nejvyšší hodnoty hat values jsou vyhodnoceny u pozorování s indexy 400, 8, 111, 112, 391 a 392. Ani v tomto případě nicméně po nahlédnutí na údaje příslušných pozorování nevidíme důvod k jejich vyřazení z modelu.



Obrázek 16: Přehled hat values pro jednotlivé indexy pozorování.

Pokud nicméně pozorování o zmíněných indexech z modelu vyjmeme, získáme model s koeficienty vypsanými v tabulce 12. V porovnání s originálními koeficienty z tabulky 10 poté mimo proměnné LWT pozorujeme mírné rozdíly oproti původnímu modelu. Význam proměnných se nicméně malými rozdíly

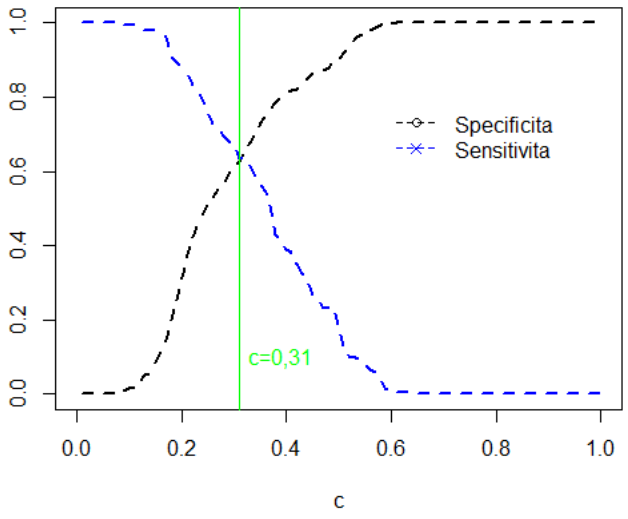
v koeficientech výrazně nezmění. Oproti původnímu modelu je nicméně zajímavá změna pro rasu Black jakožto nyní silnější protektivního faktoru nežli Other.

	odhad	95% interval spolehlivosti
β_0	-0,148	(-1,460, 1,175)
β_1	0,029	(-0,009, 0,067)
β_2	0,467	(0,027, 0,906)
β_3	-0,933	(-1,670, -0,260)
β_4	-0,857	(-1,361, -0,366)
β_5	-0,009	(-0,016, -0,001)

Tabulka 12: Koeficienty s 95% intervaly spolehlivosti pro model bez podezřelých pozorování o indexech 400, 8, 111, 112, 391 a 392.

5 Vyhodnocení úspěšnosti modelu

V této sekci pro vybraný a analyzovaný model provedeme výběr hraniční hodnoty c a pro takto finalizovaný model následně provedeme jeho vyhodnocení. Pro výběr optimální hodnoty c využijeme tzv. specificitu a senzitivitu, které určují kvalitu klasifikace pro jednu či druhou třídu. Na obrázku 17 poté vykreslujeme průběh specificity a senzitivity v závislosti na volbě c. V našem případě poté volíme optimální hodnotu c jako bod, kde se křivky kříží. Alternativně by poté šlo optimální hodnotu c volit též jako průsečík tzv. recall a precision, či jako bod maxima F1 skóre. Náš model tedy pro predikci příslušnosti k



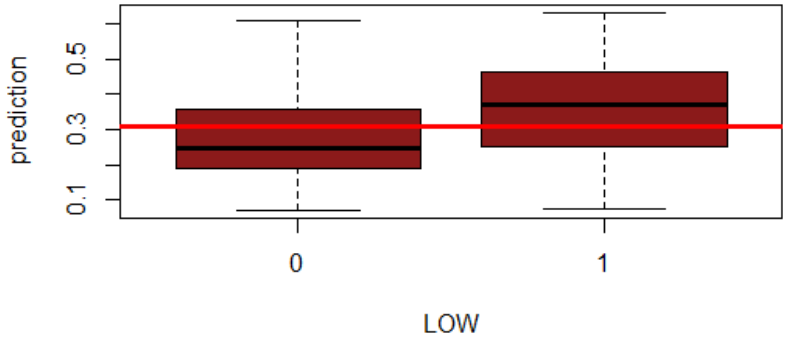
Obrázek 17: Křivky senzitivity a specificity pro různou volbu hraniční hodnoty c s vyznačenou optimální hodnotou c.

dané třídě používá rozhodovací pravidlo ve tvaru

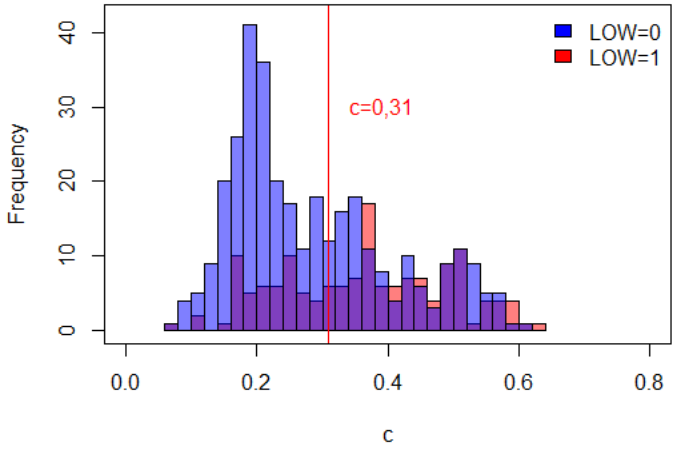
$$\widehat{\text{LOW}} = \begin{cases} 1 \text{ pro } p > 0,31. \\ 0 \text{ pro } p \leq 0,31. \end{cases} \tag{6}$$

Rozdělení dvou množin volbou hraniční hodnoty $c = 0,31$ je možno ilustrovat na obrázcích 18 a 19. Zde je vidět, že model sice neseparuje obě třídy zcela dokonale, ale v rámci mezí dokáže zaručit nám vyhovující přesnost klasifikace obou tříd.

S takto zadefinovanou hraniční hodnotou nyní můžeme dále zkoumat predikci modelu. Výsledky lze velmi přehledně formulovat v tzv. matici záměn zobrazené v rámci tabulky 13. Diagonální prvky tabulky tvoří počty správně klasifikovaných pozorování, mimo diagonálu se naopak nacházejí počty pozorování třídy LOW=1 vyhodnocených jako LOW=0 či naopak. Z tabulky 13 je tak zřejmé, že u obou tříd náš



Obrázek 18: Boxploty pro predikované pravděpodobnosti s indikací vybrané hraniční hodnoty.



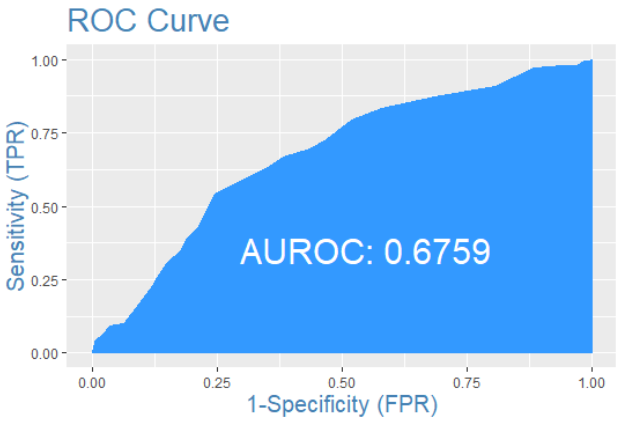
Obrázek 19: Histogramy predikovaných výstupů s indikací třídy LOW a vyznačenou optimální hodnotou c.

	LOW=1	LOW=0	Total
$\widehat{\text{LOW}} = 1$	96 (63,58%)	102 (30,27%)	198
$\widehat{\text{LOW}} = 0$	55 (36,42%)	235 (69,73%)	290
Total	151	337	488

Tabulka 13: Matice záměn pro třídy LOW=0 a LOW=1.

model klasifikoval výrazně lépe nežli náhodný klasifikátor. To lze ukázat například na metrice accuracy vypočtené jako poměru správně klasifikovaných pozorování a celkového počtu pozorování. Pro třídu LOW=1 totiž bylo docíleno accuracy $ACC = 63,58\%$, zatímco pro třídu LOW=0 činila hodnota accuracy $ACC = 69,73\%$. Celková accuracy pak činí $ACC = 67,83\%$. Vzhledem k nevybalancovanosti tříd je zde významný vliv volby hraniční hodnoty $c = 0,31$. Při standardní hodnotě $c = 0,5$ bychom totiž docílili lepší accuracy pro mohutnější třídu, ale výrazně horší accuracy pro třídu méně hojně zastoupenou.

Další metrikou pro úlohu statistické binární klasifikace je též AUC, tedy plocha pod tzv. ROC křivkou. Tu vykreslujeme na obrázku 20. Z výsledné hodnoty $AUC = 67,59\%$ lze poté vyvodit poměrně dobré klasifikační vlastnosti našeho modelu.



Obrázek 20: ROC křivka s vypsanou hodnotou AUC.

6 Závěr

V rámci této úlohy jsme použili model logistické regrese pro řešení úlohy binární klasifikace pro predikci nízké porodní hmotnosti ze zadaných údajů o rodičkách. V první části práce jsme provedli numerickou a grafickou analýzu dat, ve které jsme podrobně popsali specifika datové množiny a získali podezření na faktory pozitivní pro výskyt nízké porodní váhy. Následně jsme sestavili tři zobecněné lineární modely s různými spojovacími funkcemi a na základě jejich shody s daty i interpretovatelnosti jsme se rozhodli použít model využívající logitovou spojovací funkci. Následně jsme analyzovali význam jednotlivých koeficientů, rezidua a influenční pozorování. Pro zvolený model jsme na základě metrik binární klasifikace zvolili optimální hraniční hodnotu $c = 0,31$. S jejím využitím jsme poté dosáhli celkové přesnosti klasifikace $ACC = 67,83\%$ a $AUC = 67,59\%$.

Pokud bychom chtěli dosáhnout lepších výsledků klasifikace, bylo by dále možno použít pokročilejší klasifikační metody jako rozhodovací stromy či neuronové sítě. Dalším způsobem jak potenciálně zlepšit úspěšnost klasifikace by bylo místo klasifikační úlohy pro proměnnou LOW řešit regresní úlohu pro proměnnou BWT. V proměnné LOW totiž ztrácíme informaci o skutečnosti, jak moc se liší skutečná porodní váha od prahové hodnoty 2500 g. Tímto způsobem bychom tak případně mohli odhadnout též faktory zapřičiňující extrémně nízkou porodní hmotnost, neboť rozdělení do pouhých dvou tříd by v praxi mohlo být nedostatečné.