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

DIPLOMOVÁ PRÁCE

Volba dopravního prostředku

Vedoucí diplomové práce:
RNDr. Ph.D. Tomáš Fürst
Rok odevzdání: 2014

Vypracovala:
Bc. Tereza Šolcová
prezenční studium
AME, II. Ročník

Název práce: Volba dopravního prostředku

Autorka: Bc. Tereza Šolcová

Katedra: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: RNDr. Ph.D. Tomáš Fürst

Akademický rok: 2014/2015

Počet stran: 50

Abstrakt:

Cílem práce je sestavit model multinomické regrese na datech získaných z Jihomoravského kraje, který bude co nejlépe odhadovat pravděpodobnost volby určitého dopravního prostředku pro uskutečnění cesty a zjišťovat, jak tuto pravděpodobnost ovlivňují sociodemografické aspekty. Dopravních módů je pět, a to *chůzi*, *kolo*, *veřejnou dopravu*, *auto jako řidič* a *auto jako spolujezdec*. Díky tomu se podařilo sestavit čtyři modely logistické regrese. V důsledku zjištění, že není k dispozici dostatečný počet dat pro validaci, rozhodla jsem se, že v práci budu pracovat s křížovou validací. Nejlepší modely jsem hledala na základě nejmenšího Akaikeho informačního kritéria. Úpravu dat jsem prováděla pomocí programu Excel a R.

Klíčová slova: logistická regrese, multinomická regrese, křížová validace, volba dopravního prostředku

Title: The choice of transport

Author: Bc. Tereza Šolcová

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: RNDr. Ph.D. Tomáš Fürst

Akademic year: 2014/2015

Numer of pages: 50

Abstract:

The aim is to construct a model of multinomial regression on the data from the South Moravian Region, with the best estimation of the probability a certain choice of transport and find out how socio-demographic aspects influence the likelihood. We have five transport modes - *walking, cycling, public transport, car as driver* and *car as a passenger*. We created four logistic regression models. We find out, that thanks to insufficient number of data we have to validate with cross-validation. The best models we are looking through the smallest Akaike information criterion. Modification of data were carried out using Excel and R.

Keywords: logistic regression, multinomial regression, cross-validation, the choice of transport

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala sama pod vedením pana Tomáše Fürsta a v seznamu literatury jsem uvedla všechny použité zdroje.

V Olomouci dne: Šolcová Tereza

Poděkování

Ráda bych poděkovala vedoucímu diplomové práce panu Tomáši Fürstovi za spolupráci i za čas, který mi věnoval při konzultacích.

Obsah

Obsah ... 6
Seznamy obrázků: ... 7
Seznamy tabulek: ... 7
Seznam příloh: ... 8
Úvod ... 9
1. Seznámení se základními pojmy ... 10
1.1 Kategoriální data ... 10
1.2 Rozdělení užívaná pro kategorická data ... 11
1.2.1 Binomické rozdělení ... 11
1.2.2 Multinomické rozdělení ... 11
1.3 Kontingenční tabulky ... 12
1.4 Logistická regrese ... 14
1.5 Metoda maximální věrohodnosti ... 16
1.6 Křížová validace ... 17
1.7 Testy významnosti regresních koeficientů ... 18
1.7.1 Test poměrem věrohodností ... 18
1.7.2 Waldův test ... 19
1.8 Nástroje pro posouzení vhodnosti modelu ... 19
1.8.1 Akaikeho informační kritérium ... 19
1.8.2 Deviance ... 20
1.8.3 McFaddenův index determinace ... 20
1.8.4 Nagelkerkův index determinace ... 20
2. Průběh výzkumu ... 21
3. Práce s daty ... 22
4. Sestavení modelu ... 29
4.1 Auto jako řidič s chůzí ... 30
4.2 Auto jako řidič s kolem ... 35
4.3 Auto jako řidič s veřejnou dopravou ... 38
4.4 Auto jako řidič s autem jako spolujezdec ... 40
4.5 Test poměrem věrohodností ... 44
4.6 Kvalita proložení dat ... 45
4.6.1 Akaikeho informační kritérium ... 45
4.6.2 Indexy determinace ... 45
Závěr ... 47
Seznam použité literatury ... 49
Přílohy ... 50

Seznamy obrázků:

Obrázek 1: Ukázka kontingenční tabulky……..……..... 12
Obrázek 2: Logistická křivka (S-křivka)……15
Obrázek 3: Histogram účelů cesty…….…….25
Obrázek 4: Volba způsobu dopravy……..... 26
Obrázek 5: Volba způsobu dopravy závislá na pohlaví…….. 26
Obrázek 6: Volba způsobu dopravy závislá na zaměstnání…….…….…….…. 27
Obrázek 7: Volba způsobu dopravy závislá na vzdělání…….……...…27
Obrázek 8: Volba způsobu dopravy závislá na struktuře domácnosti…….……28
Obrázek 9: Výpis modelu auto jako řidič + chůze…….…..…..… 31
Obrázek 10: Výpis upraveného modelu auto jako řidič + chůze……..…...33
Obrázek 11: Výpis druhého upraveného modelu auto jako řidič +chůze...……..….. 34
Obrázek 12: Výpis modelu auto jako řidič + kolo……..… 35
Obrázek 13: Výpis upraveného modelu auto jako řidič + kolo……..…....…...36
Obrázek 14: Výpis druhého upraveného modelu auto jako řidič + kolo…….37
Obrázek 15: Výpis modelu auto jako řidič + VD……..……...38
Obrázek 16: Výpis upraveného modelu auto jako řidič + VD…….……... 39
Obrázek 17: Výpis modelu auto jako řidič + spolujízda…….…… 41
Obrázek 18: Výpis upraveného modelu auto jako řidič + spolujízda……. 42
Obrázek 19: Výpis druhého upraveného modelu auto jako řidič + spolujízda……... 43
Obrázek 20: Test poměrem věrohodnosti prvního modelu……44
Obrázek 21: Test poměrem věrohodnosti druhého modelu...…… 44
Obrázek 22: Test poměrem věrohodnosti třetího modelu..…… 44
Obrázek 23: Test poměrem věrohodnosti čtvrtého modelu……...…… 45

Seznamy tabulek:

Tabulka 1: Matice pravděpodobností …… 13
Tabulka 2: Kontingenční tabulka…… 13
Tabulka 3: Ukázka agregace proměnné vzdělání…….……..……23
Tabulka 4: Ukázka agregace proměnné práce…….…..……..……24
Tabulka 5: Souhrn hodnot AIC…...……..…….. 45
Tabulka 6: Souhrn hodnot indexů determinace u základních modelů…… 46
Tabulka 7: Souhrn hodnot indexů determinace u upravených modelů…….…….. 46

Seznam příloh:

Příloha 1: Dotazník pro domácnosti……..... 50
Příloha 2: Dotazník pro osoby……..…... 53
Příloha 3: Dotazník pro cesty…….. 56

Úvod

Při výběru tématu diplomové práce jsem se soustředila na to, aby práce vycházela spíše z praktického prostředí, aby bylo možno výsledky importovat do běžného života. Jelikož jsem měla problém s výběrem tématu diplomové práce na Katedře matematické analýzy a aplikací matematiky, využila jsem možnosti spolupracovat s Centrem dopravního výzkumu (CDV). Bohužel, představa CDV o zpracování daného tématu se významně neshodovala s představou katedry, rozhodli jsme se proto s panem vedoucím, že využijeme data poskytnutá CDV a pokusíme se téma zpracovat ještě jednou.

Cílem práce je tedy vytvořit co nejvíce vhodný a dobře interpretovatelný model, který bude odhadovat pravděpodobnost, s jakou respondent zvolí určitý dopravní prostředek. Za dopravní prostředek budu v této práci uvažovat *chůzi, kolo, veřejnou dopravu, auto jako řidič* a *auto jako spolujezdec*. Budu zjišťovat, které sociodemografické aspekty mají vliv na respondentovo rozhodnutí. Model bude stavěn na datech poskytnutých CDV, která jsou získána z Jihomoravského kraje. K vypracování tohoto cíle budu využívat multinomickou regresi. Jestliže mluvím o respondentovi, mám na mysli účastníka dotazníkového průzkumu.

Zmíněná data z Jihomoravského kraje byla získána pomocí dotazníkového šetření. Průzkum prováděla agentura FOCUS na jaře roku 2013, přičemž zadavatelem bylo CDV. Ze získaných dat jsem se rozhodla do této práce zahrnout následující vysvětlující proměnné – *věk respondenta, vzdálenost od zastávky, délku cesty, vlastnictví řidičského průkazu, pohlaví respondenta, vzdělání, práci, strukturu domácnosti, ve které respondent žije* a *účel cesty*.

Práci dělím do 4 kapitol. V první kapitole se snažím seznámit čtenáře se základními matematickými pojmy, bez kterých bych se neobešla v praktické části. Ve druhé kapitole bych ráda seznámila čtenáře s tím, jak byla získána data. Načež volně navazuje kapitola třetí, která uvádí některé popisné statistiky a už také detailně popisuje proměnné, vstupující do modelu. O výstavbě samotného modelu pojednává kapitola 4.

Pro to, abych stanoveného cíle dosáhla, je nutné důkladně se seznámit s daty a následně je analyzovat pomocí programu R.

1. Seznámení se základními pojmy

V této kapitole bych ráda objasnila všechny základní matematické pojmy vyskytující se v práci a týkající se dané problematiky. Veškeré poznatky v této kapitole jsou čerpány ze zdrojů [1], [4], [7], nebude-li řečeno jinak. Tato část je ještě dále dělena do dalších osmi podkapitol. V kapitole 1.1 seznámím čtenáře s kategoriálními daty, která mě budou provádět celou prací. V kapitole další potom uvádím rozdělení typické pro kategoriální data. Třetí podkapitola pojednává o kontingenčních tabulkách. Další dvě části se potom postupně zabývají logistickou regresí a metodou maximální věrohodnosti. V kapitole 1.6 se zabývám křížovou validací, která je nezbytná pro správné vypočtení chyby modelu. A v posledních dvou kapitolách jsou zmíněny testy významnosti parametrů a míry pro posouzení kvality modelu.

1.1 Kategoriální data

Než se pustím do představení metod, které v diplomové práci budu užívat, je třeba se pozastavit nad pojmem kategoriální data, na kterých jsou metody založeny. Jsou zde uvedeny hlavní typy kategoriálních dat.

Kategorické proměnné se skládají ze sady kategorií. Mají omezený počet hodnot. Jako příklad kategorické proměnné uvedeme rodinný stav, který můžeme klasifikovat do kategorií – svobodný, ženatý, rozvedený, vdovec. Kategorické proměnné se mohou objevovat v mnoha různorodých odvětvích.

Dále můžeme dělit proměnné na *nominální* a *ordinální* kategorické proměnné. Jestliže máme kategorie bez přirozeného uspořádání, pak je nazýváme nominální. Jako příklad zde můžeme uvést náboženskou příslušnost – katolická, protestantská, židovská, muslimská,… Jak vidíme, pořadí kategorií nemá žádný význam. Objevují se i kategorické proměnné, které mají uspořádané kategorie. Tyto pak nazýváme ordinální. Příkladem může být vzdělání – nedokončené základní, základní, střední s maturitou, vysoká škola. Vidíme zde určité uspořádání, ale nejsme schopni říci, jaké jsou vzdálenosti mezi kategoriemi.

Intervalová proměnná je taková proměnná, která má číselné vzdálenosti mezi každými dvěma hodnotami. Jako příklad zde mohu uvést roční příjem.

Jako poslední mohu uvést dělení na *kvantitativní* a *kvalitativní*. Obecně se dá říci, že nominální proměnné jsou kvalitativní, protože jednotlivé kategorie se liší v kvalitě, nikoli

v množství. Intervalové proměnné jsou naopak kvantitativní. Ordinální proměnné mohou být kvantitativní i kvalitativní.

1.2 Rozdělení užívaná pro kategorická data

Deduktivní analýzy dat vyžadují znalost předpokladů o náhodném mechanismu, který generoval data. V případě že pracuji s kategoriálními proměnnými, je třeba znát binomické a multinomické rozdělení.

1.2.1 Binomické rozdělení

Nechť $y_1, y_2, ...,y_n$ označují výsledky n nezávislých a identických pokusů, $P(Y_i = 1) = \pi$ a $P(Y_i = 0) = 1 - \pi$. Pro tyto výsledky užíváme označení „úspěch“ a „neúspěch“ pro výstupy 1 a 0.

V předchozím odstavci jsem mluvila o nezávislých a identických pokusech. *Identické* pokusy chápeme tak, že pravděpodobnost úspěchu π je stejná v každém pokusu. A dále jestliže jsou označeny výsledky dílčích pokusů $y_1, y_2, ...,y_n$, pak řekneme, že pokus je nezávislý, jestliže pro výsledek ($y_1, y_2, ...,y_n$) sdruženého pokusu platí, že

$$P(y_1, y_2, ..., y_n) = P(y_1) * P(y_2) * ... * P(y_n).$$

Celkový počet úspěchů $Y = \sum_{i=1}^{n} Y_i$ má binomické rozdělení s parametry n, π. Označujeme Bi(n, π). Rozdělení pravděpodobnosti je

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y=0, 1, 2,\ldots, n.$$

Střední hodnota binomického rozdělení je E(Y) = $n\pi$, rozptyl var(Y) = $n\pi(1-\pi)$.

1.2.2 Multinomické rozdělení

Některé pokusy mohou mít více než dva možné výsledky. Předpokládejme, že každý z n nezávislých, identických pokusů má výstup, který můžeme klasifikovat do některé z c kategorií. Označme $y_{ij} = 1$, jestliže i-tý pokus má výstup v kategorii j a $y_{ij} = 0$, jinak. Potom $y_i = (y_{i1}, y_{i2}, ..., y_{ic})$ reprezentují multinomický pokus, kde $\sum_j y_{ij} = 1$. Nechť $N_j = \sum_i y_{ij}$ označuje počet pokusů, u kterých nastal výsledek v kategorii j. Čísla $(N_1, N_2, ..., N_c)$ mají multinomické rozdělení.

Nechť $\pi_j = P(Y_{ij} = 1)$ označuje pravděpodobnost, se kterou nastane výsledek v kategorii j. Nakonec tedy uvažujeme pravděpodobnost

$$p(N_1, N_2, ..., N_{c-1}) = \frac{n!}{N_1! N_2! ... N_c!} \pi_1^{N_1} \pi_2^{N_2} ... \pi_c^{N_c},$$

kde $\sum_j N_j = n$.

Jednotlivé náhodné veličiny N_j mají binomické rozdělení s parametry n a π_j, platí tedy

$E(N_j) = n\pi_j$, $\quad$ $\text{var}(N_j) = n\pi_j(1-\pi_j)$.

1.3 Kontingenční tabulky

Tato kapitola byla vypracována s použitím zdroje [6].

V této kapitole jsou znázorněny tabulky, které zachycují vztah mezi kategorickými proměnnými. Uvažujme dvourozměrný náhodný vektor s náhodnými veličinami *X, Y*, které nabývají hodnot $1,...,r$ a $1,...,s$ s pravděpodobnostmi $p_{ij} = P[X = i, Y = j], i = 1,...,r, j = 1,...,s$. Tato situace vznikne tehdy, snažíme-li se pozorovat dva znaky v jeden okamžik.

Příkladem kontingenční tabulky typu 2x2 může být ukázka poměru mužů a žen, jestliže budeme sledovat, zda jsou leváci či praváci.

	levák	pravák
muži	28	208
ženy	113	150

Obrázek 1: Ukázka kontingenční tabulky

Označíme-li

$$p_{i\bullet} = P[X = i] = \sum_{j=1}^{s} P[X = i, Y = j]$$

a

$$p_{\bullet j} = P[Y = j] = \sum_{i=1}^{r} P[X = i, Y = j],$$

pak můžeme vše zapsat do tabulky

X/Y	1	2	3	…	s	$\sum$
1	p_{11}	p_{12}	p_{13}	$\cdots$	p_{1s}	$p_{1\bullet}$
2	p_{21}	p_{22}	p_{23}	$\cdots$	p_{2s}	$p_{2\bullet}$
⋮	⋮	⋮	⋮		⋮	⋮
r	p_{r1}	p_{r2}	p_{r3}	$\cdots$	p_{rs}	$p_{r\bullet}$
$\sum$	$p_{\bullet 1}$	$p_{\bullet 2}$	$p_{\bullet 3}$	$\cdots$	$p_{\bullet s}$	1

Tabulka 1: Matice pravděpodobností

Kontingenční tabulky využíváme nejčastěji k provedení testu hypotézy H_0, zda jsou veličiny X a Y nezávislé. K tomu je však třeba zavést ještě některá označení. Například symbolem n_{ij} označíme četnost jevu $[X = i, Y = j]$ při provedení dvourozměrného náhodného výběru $(X_1, Y_1),...,(X_n, Y_n)$ příslušného náhodnému vektoru (X, Y) a pro marginální četnosti zaveďme označení

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}, n_{\bullet j} = \sum_{i=1}^{r} n_{ij} .$$

Což můžeme opět vepsat do tabulky, kterou nazýváme kontingenční tabulka.

X/Y	1	2	3	…	s	$\sum$
1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1s}	$n_{1\bullet}$
2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2s}	$n_{2\bullet}$
⋮	⋮	⋮	⋮		⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	$\cdots$	n_{rs}	$n_{r\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$\cdots$	$n_{\bullet s}$	n

Tabulka 2: Kontingenční tabulka

O četnostech můžeme říci, že vyjadřují realizaci náhodného vektoru s multinomickým rozdělením s parametry $p_{11}, p_{12},..., p_{rs}, n$, protože n-krát nezávisle opakujeme pokus s rs možnými výsledky s pravděpodobnostmi $p_{11}, p_{12},..., p_{rs}$.

Po ujasnění si několika základních pojmů se konečně dostáváme k výše zmiňovanému testu nezávislosti. Testová statistika pro test nezávislosti

$$T = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n_{i\bullet}n_{\bullet j}/n\right)^2}{n_{i\bullet}n_{\bullet j}/n}.$$

Hypotézu o nezávislosti veličin X a Y zamítáme, když je $T \geq \chi^2_{(r-1)(s-1)}(1-\alpha)$, kde $(r-1)(s-1)$ je počet stupňů volnosti. Neměli bychom zapomenout na podmínku četnosti tříd, kterou je doporučeno splnit, abychom tento test mohli provést : $\frac{n_{i\bullet}n_{\bullet j}}{n} \geq 5$.

1.4 Logistická regrese

Protože cílem práce je prozkoumat vztah mezi závisle proměnnou a nezávisle proměnnou, rozhodli jsme se použít regresní model. Jelikož hodnoty závisle proměnné *Y* nabývají pouze hodnot 0 a 1, nemůžeme volit „klasickou“ lineární regresi, ale musíme uvažovat regresi logistickou.

Dle typu vysvětlující proměnné se dá rozlišit:

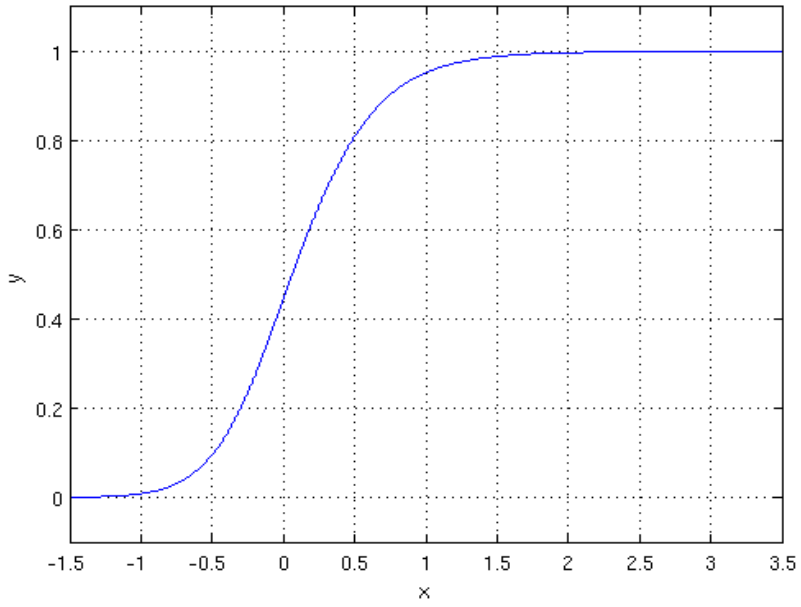
a) *Logistická regrese*, kdy závislá proměnná je binárního typu, a která nabývá pouze dvou možných hodnot, jako je například *muž* a *žena*. Vysvětlující proměnná pak může obsahovat jednu či více proměnných, jak spojitých, tak kategorických.

b) *Multinomická regrese*, týká se závisle proměnné, která nabývá tří a více možných stavů. Buď se můžeme setkat s případem, kdy v závisle proměnné existuje určité přirozené uspořádání, jako je *silný nesouhlas, nesouhlas, souhlas, silný souhlas*. A nebo s případem, kdy u závisle proměnné nepozorujeme žádné uspořádaní a stavy definují pouze odlišnost. Vektor vysvětlujících proměnných se u obou případů může skládat jak ze spojitých tak z kategorických proměnných.

Je nutné nějakým způsobem vytvořit vztah mezi $P(Y = 1/x)$ a x. Kdybychom použili model lineární regrese $\pi(x) = \alpha + \beta x$, pravděpodobnost by mohla nabývat záporných hodnot a hodnot vyšších jak 1, což nedává smysl. Abychom se tomuto problému vyhnuli, používáme funkci, jejíž výstupy se pohybují mezi hodnotami 0 a 1. Volíme logistickou funkci.

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

Křivka logistické funkce, zvaná S-křivka, dává smysluplné odhady při všech možných hodnotách *x*. Při velmi nízkých hodnotách nezávisle proměnné se pravděpodobnost závisle proměnné blíží 0 a při vysokých hodnotách nezávisle proměnné se blíží 1.



Obrázek 2: Logistická křivka (S-křivka) – osa *y* značí odhad pravděpodobnosti jevu a osa *x* hodnoty nezávisle proměnných, zdroj [11]

Nelineární vztah mezi $\pi(x)$ a x bývá obvykle monotónní. $\pi(x)$ se stále zvyšuje nebo stále klesá, jak x roste. Jestliže $x \to \infty$, pak $\pi(x)$ bude klesat k nule, jestliže β<0, a $\pi(x)$ poroste k jedničce, když β>0.

Poté můžeme uvést i vztah

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

Logaritmem předchozí rovnice získáme lineární vztah

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x,$$

čemuž říkáme logit. Logit je logaritmus šance na úspěch v jednotlivých pokusech. Předpokládáme, že je lineární v parametrech. Šance pak udává podíl pravděpodobnosti výskytu jevu ku pravděpodobnosti, že se jev nevyskytne.

V této práci však budu muset řešit multinomickou regresi, kdy výstupní proměnná nabývá více než dvou hodnot. Nepředpokládám žádné přirozené uspořádání těchto hodnot. Uvažuji zde

dopravní módy *chůze, kolo, auto jako řidič, veřejná doprava* a *auto jako spolujezdec*. Pokud bych však zvolila jeden dopravní mód jako referenční, nemusela bych použít multinomickou regresi, ale stačily by čtyři logistické regrese, protože lze podle [8] psát

$$\log\frac{\pi_1(x)}{\pi_0(x)} = \alpha_1 + \beta_1 x \qquad \text{a} \qquad \log\frac{\pi_2(x)}{\pi_0(x)} = \alpha_2 + \beta_2 x$$

Předpokládejme, že index 0 značí referenční hodnotu. Já ve své práci raději než multinomickou regresi volím čtyři regrese logistické.

1.5 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti se užívá k bodovému odhadu koeficientů v regresním modelu. Dalo by se namítnout, že není třeba představovat nové metody odhadu koeficientů, když máme k dispozici metodu nejmenších čtverců, která je hojně využívaná. To je sice pravda, ale tato metoda se využívá pouze u lineárního regresního modelu a u logistického modelu ji nelze použít, proto použijme obecnější metodu maximální věrohodnosti, protože má lepší statistické vlastnosti.

Tato metoda vychází z věrohodnostní funkce výběru. Pro diskrétní náhodné veličiny s rozdělením pravděpodobnosti $p(x_i, \Theta)$, jenž jsou vzájemně nezávislé, je pak věrohodnostní funkce $L(\Theta)$ vyjádřena jako pravděpodobnost současného výskytu všech prvků výběru

$$L(\Theta) = \prod_{i=1}^{n} p(x_i, \Theta), \qquad \text{kde } \Theta \text{ je vektor parametrů.}$$

Pro spojité náhodné veličiny je věrohodnostní funkcí sdružená hustota pravděpodobností $f(x_1,...,x_n,\Theta)$. Jestliže jsou prvky výběru nezávislé, tj. platí, že

$$L(\Theta) = \prod_{i=1}^{n} f(x_i, \Theta),$$

používá se místo funkce $L(\Theta)$ jejího logaritmu, který zachovává polohu extrému.

Maximálně věrohodný odhad $\hat{\Theta}$ vektoru parametrů Θ odpovídá bodu maxima věrohodnostní funkce $L(\Theta)$. Jestliže logaritmus věrohodnostní funkce derivujeme, pak dostáváme k určení maximálně věrohodných odhadů soustavu rovnic

$$\frac{\partial \ln L(\Theta)}{\partial \Theta_j} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \Theta)}{\partial \Theta_j} = 0$$

pro $j = 1,...,m$, kde m je počet odhadovaných parametrů. Pro dostatečně velké rozsahy výběru (asymptoticky pro $n \to \infty$) jsou odhady nejlepší, nestranné a mají normální rozdělení.

1.6 Křížová validace

Tuto kapitolu jsem vypracovala pomocí literatury [5].

Validaci zavádíme, abychom byli schopni vypočítat chybu, která byla způsobena užitím statistické metody k odhadu výsledků na novém pozorování. Pod pojmem nové pozorování máme na mysli taková pozorování, na kterých nebyl model vystavěn. Je pravda, že jestliže budeme chtít vypočítat chybu modelu na datech, na kterých jsme model vystavěli, tak nic nebrání tomu, abychom to udělali. Ovšem tento odhad chyby nebude dobrý, protože nám není schopen říct, jak se bude model chovat pro neznámá data.

Řešení spočívá ve validaci, kdy množinu dat rozdělíme na dvě podmnožiny: trénovací a testovací. Na jedné části dat vytvoříme model a budeme zkoumat, jak dobře bude model pracovat na datech z testovací množiny.

Jelikož jsem v této práci měla k dispozici data, která byla třeba očistit, a rozdělit postupně na 4 části (pro jednotlivé logistické regrese), už nepřicházelo v úvahu, že bych použila klasický přístup validace uvedený výše. Použiji tedy obdobu využívanou v praxi při menším množství vzorků dat.

Data tedy náhodně rozdělíme do navzájem *N* disjunktních podmnožin. Dále data rozdělíme do testovací a trénovací množiny dat. Obvykle volíme za trénovací sadu *N* - 1 podmnožin a poslední užijeme k validaci. Abychom zvýšili objektivitu, provedeme několik opakování, kdy vyměňujeme testovací množiny.

Rozlišujeme několik druhů validace, například:

1) **K-fold cross validace** – původní sadu dat rozdělíme do *N* disjunktních podmnožin (beze zbytku). Většinou volíme *N*=10. Z těchto 10 podmnožin je použito 9 částí jako trénovací sada a poslední část slouží k validaci. Tento proces je opakován 10x a každou z 10 podmnožin využijeme pro validaci pouze jednou. Tudíž všechna data z původního souboru dat využijeme jak k trénování, tak k testování.

2) **Repeated random subsampling** – velice podobná metoda jako předchozí, opět rozdělujeme data do podmnožin a dělíme je na testovací a trénovací množinu, opět můžeme tuto metodu několikrát opakovat pro zvýšení objektivnosti. Liší se však v tom, že při výběru z originální množiny se náhodně vybere testovací a trénovací sada, díky čemuž může dojít k tomu, že některá data mohou být vybrána vícekrát a jiná naopak vůbec.

3) **Leave-one-out Gross validace** – tato metoda je založena na myšlence, že z celého souboru dat vybereme pouze jeden vzorek pro testování a zbytek bereme jako trénovací množinu. Tato metoda udává nejlepší výsledky, bohužel je velice náročná na výpočet.

V této diplomové práci budeme užívat první uvedený případ validace, tzv. křížovou validaci.

1.7 Testy významnosti regresních koeficientů

Pro to, abychom zjistili, který z vytvořených modelů má větší vypovídací hodnotu co se závisle proměnné týče, zavádíme testy hypotéz. Jestliže jsme užili metodu maximální věrohodnosti, nabízí se v této části využít hodnotu věrohodnostní funkce. Testujeme nulovou hypotézu $H_0 : \beta = \beta_0$.

1.7.1 Test poměrem věrohodností

Uvažujeme zde dva typy modelů. První, což je úplný model, jenž zahrnuje všechny odhadované parametry a druhý, který je speciálním tvarem plného modelu a nazýváme ho redukovaný model.

Pro tento test je třeba vypočítat odhady parametrů jak plného, tak redukovaného modelu. Navíc ještě potřebujeme znát maximální hodnotu věrohodnostní funkce.

Redukovaný model získáme tak, že jeden nebo více parametrů položíme rovny nule. Tím vlastně testujeme nulovost parametrů obsažených v plném modelu. Testová statistika

$$T = -2\log\frac{L_{reduced}}{L_{full}},$$

kde $L_{reduced}$ chápeme jako věrohodnostní funkci redukovaného modelu a L_{full} jako věrohodnostní funkci plného modelu, má při velkých výběrech $T \sim \chi^2_{k_2-k_1}(0,1-\alpha)$. Jestliže

testová statistika nabývá vysoké hodnoty, je vliv parametrů v plném modelu velký. Naopak, jestliže se testová statistika blíží 0, je vhodné parametr z plného modelu vynechat.

1.7.2 Waldův test

Zatímco testem poměrem věrohodnosti se dalo testovat více parametrů zároveň, u Waldova testu tomu tak není a je možné testovat významnost pouze jednoho parametru.

Předpokládáme normované normální rozdělení $W \sim N(0,1)$, potom testová statistika

$$W = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}},$$

kde $\hat{\beta}$ je odhad parametru získaný metodou maximální věrohodnosti a $s_{\hat{\beta}}$ je odhad standardní chyby testovaného parametru. Jestliže bude splněna podmínka dostatečně velkého výběru, nabývá testová statistika Waldova testu přibližně stejných hodnot jako u testu poměru věrohodností.

1.8 Nástroje pro posouzení vhodnosti modelu

Jakmile se nám podaří vytvořit určitý model, je třeba mít určité nástroje k jeho posouzení, kdy chceme zjistit, jak dobře model prokládá data. Takových nástrojů existuje celá řada, v této práci se však zaměříme pouze na pár z nich.

1.8.1 Akaikeho informační kritérium

Obecně se dá říci, že informační kritéria jsou schopna brát v úvahu dvě zásadní věci. Za prvé to, v jaké míře odpovídá odhadnutá hodnota hodnotě naměřené, a za druhé počet vysvětlujících proměnných, které jsou do modelu zahrnovány. V praxi potom tato kritéria používáme tak, že vytvoříme několik modelů a následně platí, že čím nižší je hodnota tohoto kritéria u jednotlivých modelů, tím je model lepší. Konkrétně Akaikeho informační kritérium (AIC) vypadá takto

$$AIC = 2k - 2\ln L,$$

kde k je počet parametrů a L je maximální hodnota věrohodnostní funkce.

1.8.2 Deviance

Deviance slouží podobně jako reziduální součet čtverců u lineární regrese. Devianci vypočteme jako

$$D(\beta) = 2(l_{\max} - l(\beta)),$$

kde $l_{\max}$ vyjadřuje maximální hodnotu věrohodnostní funkce v modelu, který je nejbohatší - obsahuje všechny proměnné a $l(\beta)$ je hodnota věrohodnostní funkce ve sledovaném modelu. V práci potom počítáme devianci nulového modelu, což je deviance pro model, který obsahuje pouze konstantu. Druhá počítaná deviance je deviance zkoumaného modelu. U těchto zmiňovaných charakteristik preferujeme model s co nejnižší deviací. Dále tyto vypočtené hodnoty užijeme při výpočtech indexů determinace.

1.8.3 McFaddenův index determinace

Pro vypracování následujících dvou kapitol jsem čerpala z [10].

V logistické regresi užíváme McFaddenův index determinace. Uvažujeme nulový model, což je model bez vysvětlujících proměnných, pouze model s konstantou. Devianci tohoto modelu označíme jako D_0. Dále uvažujeme libovolný jiný model s deviancí D_b. Potom můžeme psát

$$R_F^2 = 1 - \frac{D_b}{D_0}.$$

Index nabývá hodnot $R_F^2 \in \langle 0,1 \rangle$ s tím, že čím více se hodnota blíží k jedničce, tím je model lepší.

1.8.4 Nagelkerkův index determinace

Druhým koeficientem, který si zde ukážeme je Nagelkerkův koeficient determinace

$$R_N^2 = \frac{1 - e^{\frac{D_b - D_0}{n}}}{1 - e^{-\frac{D_0}{n}}}$$

a platí pro něj stejné závěry jako pro index předešlý.

2. Průběh výzkumu

Než se začnu zabývat daty určenými pro testování, měla bych také říci, jak tato data vznikla a odkud pochází.

Data, na základě kterých se snažím vytvořit model logistické regrese, pocházejí z Jihomoravského kraje. Byla získána pomocí dotazníkového průzkumu. Tento průzkum prováděla agentura FOCUS, přičemž CDV bylo zadavatelem.

Sběr dat proběhl v termínu od 21. května 2013 do 27. června 2013. Výzkum se prováděl vyplněním tří dotazníků se zaměřením zkoumat dopravní chování respondentů. Na základě vyplnění prvního dotazníku jsem schopna vyčíst základní sociodemografické charakteristiky domácností respondentů. V druhém jsou zjišťovány informace o jednotlivých členech domácností. A ve třetím dotazníku se konečně dozvídám, jak se respondent choval v určitý den. Dotazník pro osoby a dotazník pro cesty vyplňovaly všechny osoby v domácnosti starší šesti let.

Nakonec bylo získáno 1092 dotazníků od jednotlivých domácností, 2631 dotazníků od osob žijících ve zmiňovaných domácnostech a bylo podniknuto celkem 5772 cest.

V příloze jsou potom k vidění všechny tři typy dotazníků.

3. Práce s daty

CDV dodalo data v 5-ti excelovských tabulkách. Tudíž bylo třeba za prvé dát data ze souborů dohromady do jednoho souboru a za druhé z tohoto obrovského souboru dat vybrat pouze ta data, která nesou nějakou užitečnou informaci pro mé analýzy. S touto prací mi byl nápomocen software R, díky speciální funkci *merge*. Pro následnou úpravu dat postačil program Excel.

Co se následných úprav dat týče, potřebovala jsem, aby data byla homogenní. Homogenní data jsou stejnorodá data. Tuto stejnorodost by mohla narušit například měření v různých podmínkách nebo například extrémní nestejnoměrnost pozorovaných objektů. Abych splnila tento požadavek, rozhodla jsem se do analýzy zahrnout pouze data pocházející z Brna. Respondent tedy započal svou cestu v Brně a rovněž ji tam dokončil. Bohužel tímto krokem, kdy jsme ze všech pozorování vyloučili ta pozorování, která buď nezačala nebo nekončila v Brně, a nebo v rámci Brna vůbec vykonána nebyla (například cesta z Vyškova do Olomouce), jsem z celkového počtu 5772 ztratila 3189 dat. Pro modelování mi ale pořád zbývá 2583 dat.

Dále bylo třeba upravit typ zvoleného dopravního prostředku. Není rozumné, abych například do svých analýz zahrnovala typ dopravního prostředku *motocykl*, protože se v dotazníkových odpovědích objevil jen 22x. Proto jsem jej z důvodu nedostatku dat z množiny odpovědí úplně vyloučila. Dále jsem se zamýšlela nad způsoby dopravy *vlakem, tramvají, autobusem a trolejbusem*. Jednotlivé způsoby dopravy v podstatě nemají dostatečný počet pozorování na to, aby mohly být testovány jednotlivě. Ale jestliže jsem se rozhodla uvažovat pouze cestování z Brna do Brna, mohu předpokládat, že všechny tyto prostředky budou sloužit jako veřejná doprava. Navíc jsem při studování dat zjistila, že tyto jednotlivé způsoby dopravy byly libovolně kombinovány. Proto jsem tyto čtyři způsoby agregovala pouze do jedné skupiny a nazvala jsem ji *veřejná doprava*.

Další úpravy, které byly provedeny s datovým souborem, jsou zmíněny níže spolu s popisem jednotlivých proměnných:

- **délka cesty** – respondent odhadoval délku cesty v kilometrech. Jedná se o spojitou proměnnou.
- **věk-** zde uvažujeme věk respondentů – 6 - 93 let. Jedná se o kategorickou proměnnou, já s ní pracuji jako se spojitou proměnnou.

- **vzdálenost na zastávku** – odpověď na otázku, jak daleko (pěšky v minutách) se nachází nejbližší zastávka veřejné dopravy. Jedná se o spojitou proměnnou.
- **pohlaví** – tato proměnná může nabývat pouze dvou hodnot. Jestliže respondent odpoví, že je muž – pak používám označení 1, v opačném případě – tedy, že je respondent žena, píšu 0. Proměnnou, která nabývá pouze dvou hodnot, označujeme jako dichotomní. Referenční proměnnou je kategorie 0, tedy v našem případě ženy.
- **vlastnictví řidičského průkazu** – opět dichotomní proměnná popisující vlastnictví řidičského průkazu. Jestliže respondent disponuje řidičským průkazem, označím 1, v opačném případě označím 0. Referenční proměnnou je možnost, kdy respondent nevlastní řidičský průkaz.
- **vzdělání** – jedná se o kategorickou proměnnou, která popisuje dosažené vzdělání respondentů. Nabývá těchto 9 hodnot:

neukončené základní vzdělání	nedokončené základní
základní vzdělání	základní
střední vzdělání včetně vyučení – bez maturity	střední škola
úplné střední vzdělání s maturitou	
nástavbové studium	vyšší odborná škola, nástavba, vysoká škola
vyšší odborné vzdělání	
Bakalářské studium	
Magisterské studium	
Doktorské studium	

Tabulka 3: Ukázka agregace proměnné vzdělání

Původní hodnoty zjišťované v dotaznících jsou vidět v levé části tabulky a protože se mi zdálo zbytečné uvažovat v analýzách všechny tyto kategorie, agregovala jsem je pouze do čtyř hodnot, jak je vidět v pravé části tabulky. Tuto proměnnou považuji za ordinální proměnnou. Dokáže nějakým způsobem porovnat stupeň dosaženého vzdělání od nejnižšího po nejvyšší. V softwaru R ji proto budu prezentovat jako spojitou proměnnou. Díky celočíselnosti program rozpozná, že mám na mysli

ordinální proměnnou a bude s ní takto zacházet. Toto zadání je lepší z důvodu interpretovatelnosti modelu.

- **práce –** i v tomto případě se jedná o kategorickou proměnnou. V dotazníku bylo možné odpovídat následujícími způsoby:

zaměstnanec řadový	pracující
zaměstnanec vedoucí	
podnikatel bez zaměstnanců	
podnikatel se zaměstnanci	
pracující důchodce	
nepracující důchodce	nepracující
nezaměstnaný	
na mateřské dovolené	
student	student

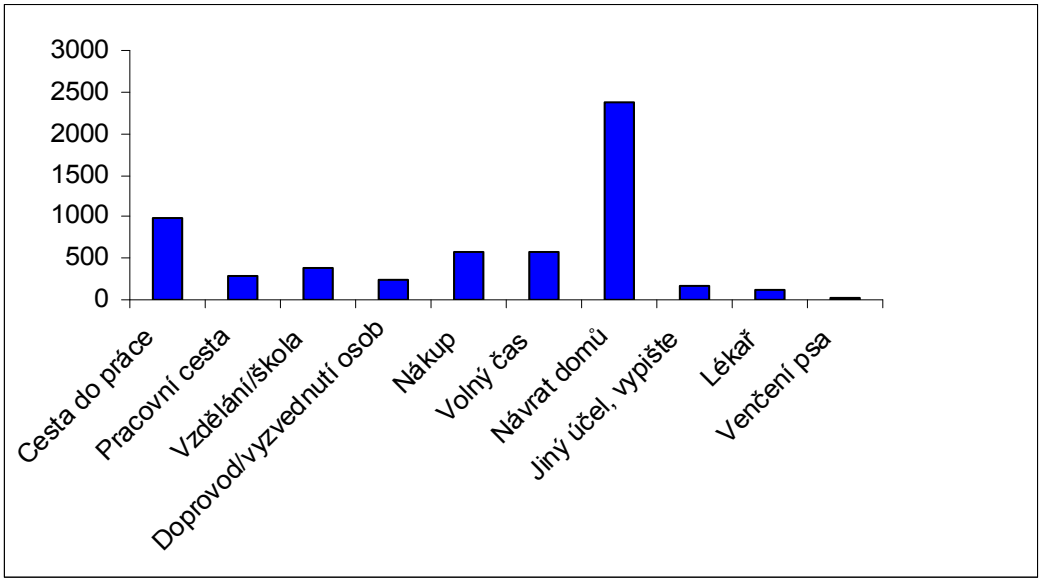
Tabulka 4: Ukázka agregace proměnné práce

Na levé straně tabulky opět vidíme hodnoty zjišťované v dotaznících. Jelikož z pohledu provádění analýzy je i tato struktura velice detailní, moje agregace je potom vidět v pravé části tabulky. Referenční proměnnou volím skupinu pracujících.

- **struktura domácnosti –** kategorická proměnná, která nám odkrývá skutečnost, s kolika osobami respondent sdílí jednu domácnost. V dotaznících jsem zaznamenala odpovědi, že osoby sdílejí domácnost až s 8 osobami. Nechtěla jsem agregovat data pouze podle počtu osob, a jelikož mi to data umožňovala, zamyslela jsem se nad tím, které skupiny osob se chovají naprosto rozdílně. Z důvodu cíle práce potom přichází v úvahu brát rodiny s dětmi a rodiny žijící bez dětí, protože si myslím, že děti výrazně ovlivňují volbu dopravního prostředku (doprava do školy, doprava do zájmových kroužků). Nakonec se jako nejpřijatelnější zdálo rozdělit osoby z dotazníkového šetření do 3 kategorií – buď „žije sám“, nebo „žije v páru“ nebo „žije s dětmi“.

Dále jsem si myslela, že by volbu dopravního prostředku mohl ovlivňovat i účel cesty. Je jasné, že jestliže jede respondent na nákup, nebude jeho nejlepší volba kolo. Proto jsem do

analýzy chtěla tato data zahrnout. Při pohledu na histogram však vidím, že to nejspíš nebude možné.



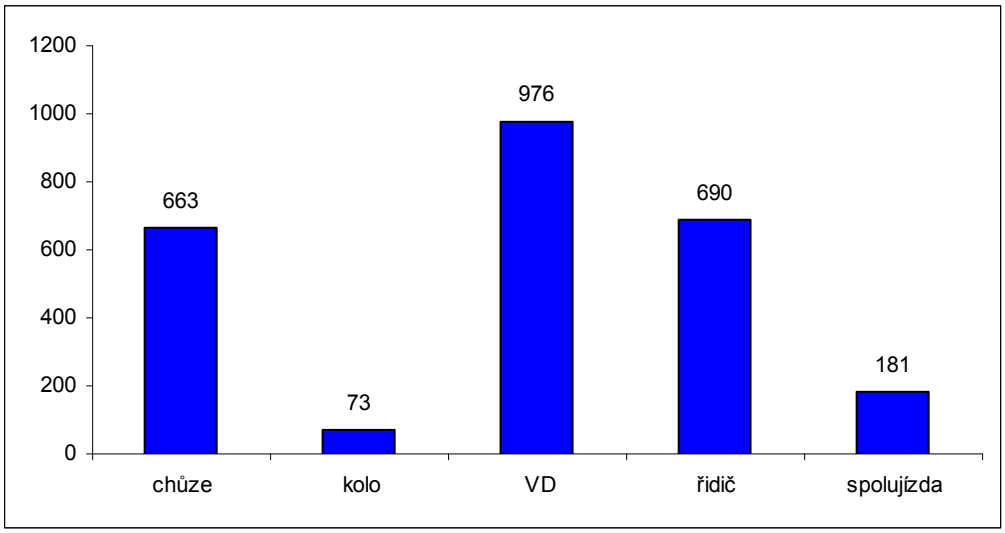
Obrázek 3: Histogram účelů cesty

CDV pravděpodobně nezvolilo nejlepší možnosti volby odpovědí na tuto otázku. Problém spočívá v tom, že většina lidí (2381) uvedla jako účel cesty „návrat domů“. Z čehož se nedá vyčíst, jestli je to návrat domů z práce, nákupu nebo například z volno-časové aktivity. Proto pro mě nemá tato proměnná žádný vypovídací charakter, a z modelu jsem se ji rozhodla vynechat.

Jak je vidno u popisu jednotlivých proměnných, většinou jsem se snažila jako referenční skupinu volit vždy tu skupinu dat, která byla datově nejobsáhlejší. V praxi je to zavedený postup, a to kvůli lepší interpretaci dat. I kdybych zvolila referenční proměnnou jinou, na tvar modelu to nebude mít význam. Pozor si však musím dát při interpretaci odhadů parametrů.

Nakonec bych zde měla poznamenat, že s takto uvedenými proměnnými jsem pracovala pouze na začátku kapitoly sestavování modelu. V průběhu testování jsem odhalovala nedostatky v agregaci nebo specifikaci proměnných (např. proměnnou *vzdělání* jsem v průběhu měnila z ordinální kategorické proměnné na nominální), proto jsem je dále v práci podle potřeby upravovala. Každá úprava však bude u jednotlivých modelů vždy zmíněna.

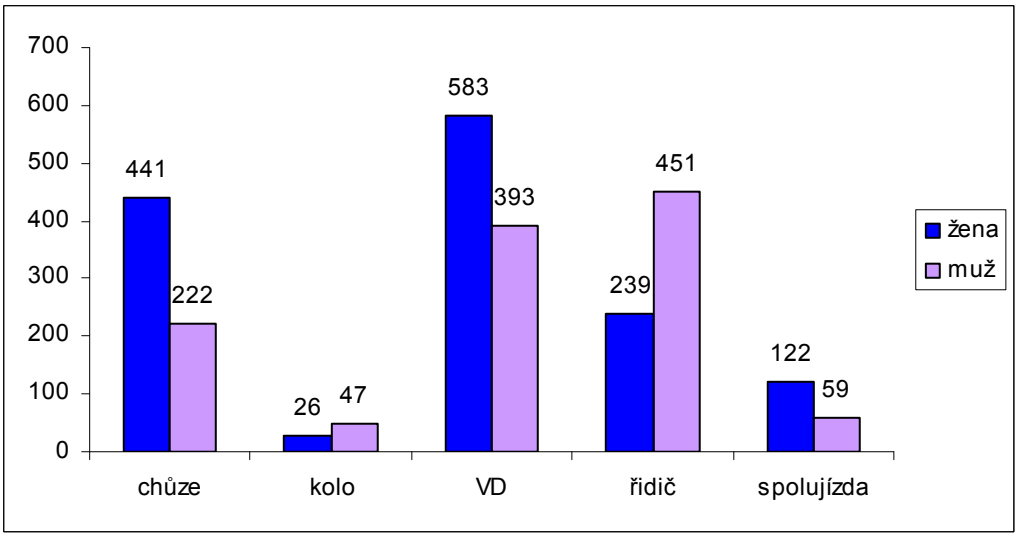
Pro ilustraci situace je níže zobrazeno pár grafů pro dokreslení situace. V následujícím grafu je vidět, jaký způsob dopravy respondenti volili.



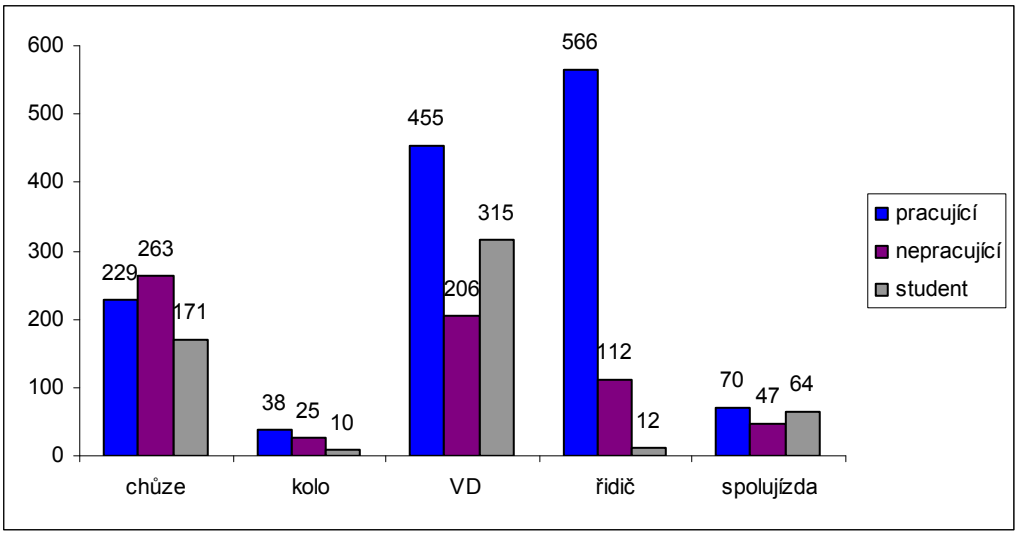
Obrázek 4: Volba způsobu dopravy

Z čehož vidíme, že z celkového počtu 2583 zvolilo pro svou cestu 663 respondentů *chůzi*, 73 respondentů *kolo*, 976 zvolilo *veřejnou dopravu (VD)*, 690 dotázaných lidí uskutečnilo svou cestu *autem jako řidič* a zbytek (181 respondentů) volilo způsob dopravy *autem jako spolujezdec*. Je také evidentní, že dva způsoby dopravy (kolo, spolujízda) nemají velké množství pozorování.

Dále jsem se chtěla podívat na to, zda pohlaví respondenta ovlivňuje volbu dopravního prostředku a z tohoto grafu bez jakýchkoli analýz můžeme říci, že ano, na pohlaví záleží. Hned v prvním sloupci vidíme, že ženy více upřednostňují *chůzi, veřejnou dopravu* a *spolujízdu*. Naopak muži častěji volí způsob dopravy *autem jako řidič*.

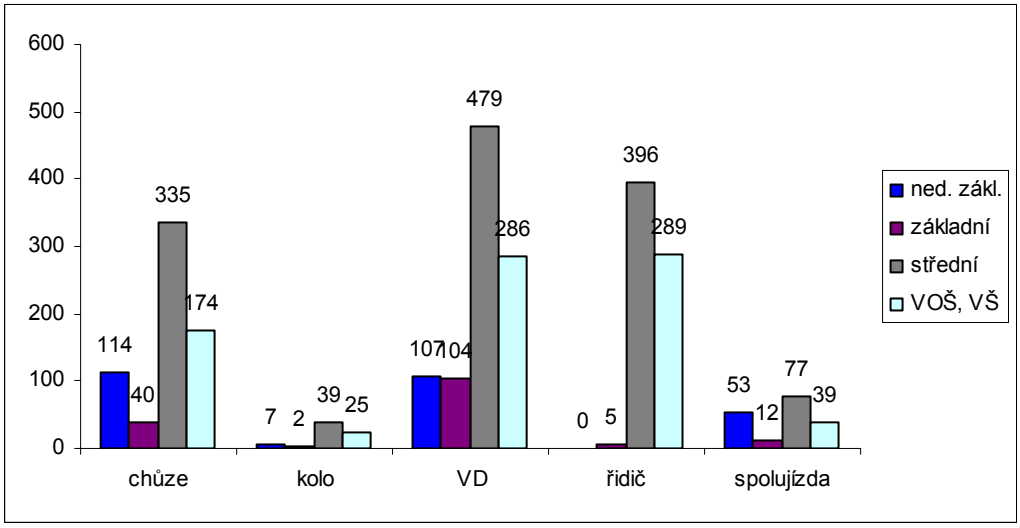


Obrázek 5: Volba způsobu dopravy závislá na pohlaví



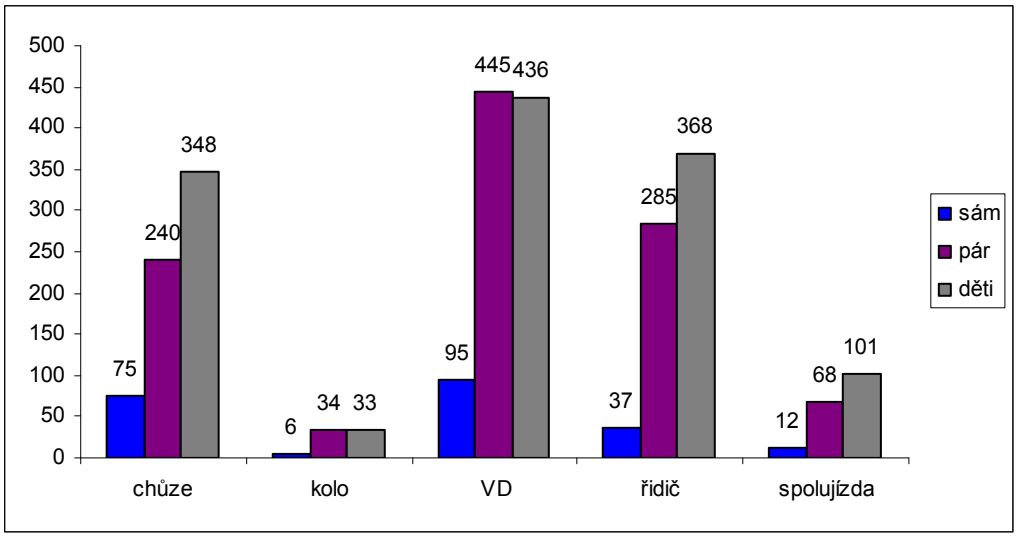
Obrázek 6: Volba způsobu dopravy závislá na zaměstnání

V obrázku 6 je vidět, jak ovlivňuje respondenta skutečnost, zda je *pracující, nepracující* či *studující*. Vidíme také, že největšího rozdílu mezi volbami dopravních prostředků nastává v případě volby dopravního prostředku *auta*, kdy *pracující* lidé volí tento prostředek nejčastěji.



Obrázek 7: Volba způsobu dopravy závislá na vzdělání

Vidíme, že v tomto grafu ve způsobu dopravy *autem jako řidič* se skoro vůbec neobjevují lidé s *nedokončeným základním* a *základním vzděláním*. Celý graf ale ovládli lidé se středním vzděláním.



Obrázek 8: Volba způsobu dopravy závislá na struktuře domácnosti

Musím se přiznat, čekala jsem, že tento graf bude mít více vypovídající výsledky pro respondenty, kteří žijí sami, že budou ve větší míře preferovat módy *chůze, kolo, veřejná doprava*. Každopádně to může být způsobeno skutečností, že v dotazníkovém šetření nemám dostatečně mnoho osob, které žijí samy. Ale u dopravního módu *auto jako řidič* je vidět, že tento prostředek volí nejčastěji rodiny s dětmi, jak se předpokládalo.

Na základě vytvořených grafů můžeme následně porovnávat interpretované výsledky, zda dávají smysl.

4. Sestavení modelu

V této části se dostávám k samotnému jádru práce. Ráda bych zde vytvořila model dopravního chování, sestavený na základě odpovědí od respondentů v Jihomoravském kraji. Snažím se odhadnout pravděpodobnost, se kterou respondent zvolí určitý dopravní mód pro svoji cestu a chci zjišťovat, jak ovlivňují výše uvedené proměnné tuto pravděpodobnost. Dopravních módů mám pět, a to *chůzi*, *kolo*, *veřejnou dopravu*, *auto jako řidič* a *auto jako spolujezdec*.

Jak jsem již zmínila výše, v této části práce budu pracovat nikoli s multinomickou regresí, ale s klasickými logistickými regresemi. Jestliže mám pět vysvětlovaných proměnných, budu provádět vždy čtyři logistické regrese, protože jeden dopravní mód bude sloužit jako referenční. V mém případě jsem se rozhodla volit za referenční mód *auto jako řidič*. Z čehož plyne, že v první logistické regresi budu sestavovat model, který se bude snažit zjistit, s jakou pravděpodobností zvolí respondent způsob dopravy *chůzi*. Pokud předpokládáme, že referenční mód je *auto jako řidič*, potom takto vzniklý model označujeme jako *Auto jako řidič* s *chůzí.* V dalších třech případech postupujeme obdobně a níže jsou vypsány jednotlivé logistické regrese, které budeme postupně provádět:

- *Auto jako řidič* s *chůzí*
- *Auto jako řidič* s *kolem*
- *Auto jako řidič* s *veřejnou dopravou*
- *Auto jako řidič* s *autem jako spolujezdec*

Protože po očištění dat a po rozdělení do 4 logistických regresí nezbylo mnoho dat k analýze, zvolila jsem k validaci modelu matematický aparát zvaný křížová validace. Křížová validace, jak už bylo zmiňováno výše, rozdělí soubor na dvě podmnožiny – testovací a trénovací sadu dat. Z trénovací množiny software vypočte sadu koeficientů specifických pro daný model a následně díky testovací množině má možnost ověřit přesnost tohoto modelu. V tomto případě používáme k=10, což znamená, že datový soubor rozdělím na deset podmnožin, kdy každé pozorování náleží právě do jedné podmnožiny. Natrénování modelu proběhne na devíti podmnožinách a na desáté vždy ověřím, jak je vytvořený model dobrý.

Mluvila jsem o zjišťování přesnosti modelu. To provádím na základě měření odchylky skutečných hodnot testovací sady od predikce sestaveného modelu. To znamená, že při deseti opakováních logistické regrese jsem schopna vypočíst deset chyb predikce a zprůměrováním

dostanu celkovou cross-validační chybu. Samozřejmě, že se budu snažit volit takovou strukturu modelu, aby byla celková chyba co nejmenší.

Pro utvoření celkové představy o zmiňované chybě uvádím i vzorec pro výpočet

$$MSE = (y_1 - \hat{y}_1)^2 \ ,$$

kde y_1 jsou skutečné naměřené hodnoty a $\hat{y}_1$ značí hodnotu predikce sestaveného modelu.

Pro celkovou cross – validační chybu potom platí

$$CV = \frac{1}{k}\sum_{i=1}^{k} MSE_i \ ,$$

kde *k* značí počet podmnožin, na které jsme rozdělila datovou množinu při cross-validaci.

Celou křížovou validaci vlastně provádím proto, abychom zabránili tzv. přefitování modelu, což může způsobit právě malé množství dat či přehnané množství vložených regresních koeficientů do modelu.

Při vytváření jednotlivých modelů se budu řídit zásadou, že nejdříve do modelu zařadím všechny proměnné, které jsou k dispozici a postupně budu zkoumat, které proměnné jsou nevýznamné a je třeba je upravit, či úplně vyřadit z modelu.

Jednotlivé logistické regrese zpracovávám pomocí programu R a velkým pomocníkem při vypracování této kapitoly mi byla kniha [5].

4.1 Auto jako řidič s chůzí

Jako první se budu snažit sestavit model, kdy budu zjišťovat, s jakou pravděpodobností zvolí respondent dopravní prostředek *auto jako řidič* nebo *chůzi*. V této části jsem měla k dispozici 1353 dat.

Ráda bych zde nejdříve ukázala model, který obsahuje všechny vysvětlující proměnné zmíněné výše. Detailně si nechám tento model vypsat a budu sledovat význam regresních koeficientů. To proto, abych si ověřila, zda jednotlivé regresní koeficienty dávají smysl a tím pádem zda i celý model dává smysl.

```
Call:
glm(formula = MODE ~ P_age + H_pt_dist + T_length + P_dlic +
    P_gen + P_edu + as.factor(sám_pár_deti) + as.factor(P_work),
    family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1611  -0.0493   0.0000   0.2687   3.4835

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                8.411761   1.076423   7.815 5.52e-15 ***
P_age                     -0.031784   0.009731  -3.266  0.00109 **
H_pt_dist                 -0.022627   0.031327  -0.722  0.47012
T_length                  -1.130474   0.083332 -13.566  < 2e-16 ***
P_dlic                    -3.942638   0.617634  -6.383 1.73e-10 ***
P_gen                     -0.482246   0.239734  -2.012  0.04426 *
P_edu                      0.182130   0.221525   0.822  0.41098
as.factor(sám_pár_deti)1 -0.456589   0.462165  -0.988  0.32319
as.factor(sám_pár_deti)2 -1.376832   0.475434  -2.896  0.00378 **
as.factor(P_work)1         1.515431   0.276281   5.485 4.13e-08 ***
as.factor(P_work)2         1.146340   0.671656   1.707  0.08787 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1875.12  on 1352  degrees of freedom
Residual deviance:  520.27  on 1342  degrees of freedom
AIC: 542.27

Number of Fisher Scoring iterations: 8
```

Obrázek 9: Výpis modelu auto jako řidič + chůze

Tento výpis můžeme pomyslně rozdělit na dvě části. První část, kdy program R zopakoval můj příkaz a druhou část, která začíná slovy „deviance residuals“, kterou software na základě mého příkazu vypočítal.

K vypočtení charakteristik modelu jsem použila funkci s následující strukturou

$$glm(Y \sim x_1 + x_2 + ... + x_n, family = binomial).$$

Funkce *glm* se používá při výpočtech zobecněných lineárních modelů, Y je označení pro závislou proměnnou, $x_1, x_2, ..., x_n$ odděleny znamínkem plus značí nezávisle proměnné vstupující do modelu a nakonec $family = binomial$ označuje skutečnost, že je jedná o binomickou závislou proměnnou. Ve výpisu modelu výše je ještě uvedeno $data = datarch$, což říká, jaká data byla pro výpočet koeficientů použita. V obrázku výpisu modelu se objevil

také příkaz *as.factor*(), čímž jsem dala softwaru informaci, že vysvětlující proměnná, obsažená v závorce má být chápána jako nominální kategorická proměnná.

V části, která byla vypočtena vidím, že zde nebyly vypsány pouze odhady parametrů a jejich významnosti, které jsou značeny '***', '**', '*', '.', ' ', postupně pro hladiny významnosti 0, 0.001, 0.01, 0.05, 0.1, 1. Umožňuje mi ale nahlédnout i na další charakteristiky modelu jako je například AIC, deviance nulového modelu a deviance skutečného modelu. Poslední dvě charakteristiky mi slouží ke zhodnocení vytvořeného modelu.

Ještě než se pustím do rozboru významu jednotlivých regresních koeficientů, měla bych vysvětlit, co se pod jednotlivými označeními skrývá za koeficienty. Označení jednotlivých nezávislých proměnných jsem totiž nechala tak, jak s nimi zacházelo CDV.

- *P_age* – věk respondenta
- *H_pt_dist* – vzdálenost od zastávky
- *T_length* – délka cesty
- *P_dlic* – vlastnictví řidičského průkazu, referenční proměnnou jsou respondenti, kteří nevlastní řidičský průkaz
- *P_gen* – pohlaví, referenční proměnnou jsou ženy
- *P_edu* - vzdělání
- *Sám_pár_deti* – struktura domácnosti, referenční proměnnou jsou lidé žijící sami
- *P_work* – práce, referenční skupinou jsou pracující lidé

Nyní se konečně podívám na jednotlivé vypočtené koeficienty. Pravděpodobnost, že respondent zvolí *chůzi*, bude klesat spolu s tím, jak respondentovi roste věk. Dále je vidět, že pravděpodobnost volby dopravního prostředku *chůze* bude také klesat v závislosti na rostoucí vzdálenosti respondenta od zastávky. Tento parametr je však statisticky nevýznamný a není se čemu divit. Jestliže se totiž rozhoduji mezi dopravním módem *chůze* a *auto jako řidič*, tak do svého rozhodování určitě vzdálenost k zastávce nezahrnuji. Dále se dá říci, že s rostoucí délkou cesty bude opět klesat pravděpodobnost, že respondent zvolí jako způsob dopravy *chůzi*. Model také říká, že vlastnictví řidičského průkazu výrazně ovlivňuje pravděpodobnost volby. Jestliže je respondent vlastníkem řidičského průkazu, bude s větší pravděpodobností volit *auto* jako dopravní prostředek. Také skutečnost, že je respondent muž, bude zvyšovat pravděpodobnost volby dopravního prostředku *auto*. Se zvyšujícím se vzděláním respondenti

raději volí *chůzi*. Tento koeficient však není statisticky významný. Dále se dívám na strukturu domácnosti, která nám říká, že pravděpodobnost volby dopravního prostředku *chůze* se bude snižovat jak pro respondenty, kteří žijí v páru, tak pro ty, kteří žijí v domácnosti s dítětem. Referenční skupina tu byly osoby, které žijí samy. A nakonec byla zkoumána skupinu pracujících, nepracujících a studentů, u kterých se prokázalo, že lidé nepracující a studující raději zvolí k uskutečnění své cesty *chůzi* oproti lidem pracujícím. Tento model byl namodelován s celkovou cross – validační chybou 0,1992884.

Nyní se budu snažit z modelu odstranit statisticky nevýznamné proměnné a budu se snažit získat model s lepším AIC.

```
Call:
glm(formula = MODE ~ P_age + T_length + P_dlic + P_gen + as.factor(sám_p
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2103  -0.0498   0.0000   0.2717   3.4592

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                8.64379    0.99451   8.692  < 2e-16 ***
P_age                     -0.03225    0.00964  -3.346 0.000821 ***
T_length                  -1.11865    0.08203 -13.638  < 2e-16 ***
P_dlic                    -3.84165    0.60827  -6.316 2.69e-10 ***
P_gen                     -0.50348    0.23841  -2.112 0.034700 *
as.factor(sám_pár_deti)1 -0.45508    0.46446  -0.980 0.327185
as.factor(sám_pár_deti)2 -1.35197    0.47625  -2.839 0.004529 **
as.factor(P_work)1         1.48213    0.27119   5.465 4.62e-08 ***
as.factor(P_work)2         1.09883    0.66652   1.649 0.099229 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1875.12  on 1352  degrees of freedom
Residual deviance:  521.56  on 1344  degrees of freedom
AIC: 539.56

Number of Fisher Scoring iterations: 8
```

Obrázek 10: Výpis upraveného modelu auto jako řidič + chůze

Tento model se od předchozího odlišuje tím, že jsem z původního modelu odstranila dva nevýznamné prediktory – vzdálenost na zastávku a vzdělání. Vidím, že hodnota AIC klesla, což je v tomto případě žádoucí. Dalším krokem v úpravě modelu bude zkoumat to, zda nebude lepší uvažovat jiné kategorie u prediktoru struktura domácnosti.

```
Call:
glm(formula = MODE ~ P_age + T_length + P_dlic + P_gen + as.factor(sám_
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0922  -0.0494   0.0000   0.2664   3.4601

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               8.262352   0.902554   9.154  < 2e-16 ***
P_age                    -0.031926   0.009617  -3.320 0.000901 ***
T_length                 -1.121236   0.082204 -13.640  < 2e-16 ***
P_dlic                   -3.855530   0.604005  -6.383 1.73e-10 ***
P_gen                    -0.490573   0.237460  -2.066 0.038836 *
as.factor(sám_pár_deti)1 -0.965798   0.258670  -3.734 0.000189 ***
as.factor(P_work)1        1.466137   0.270142   5.427 5.72e-08 ***
as.factor(P_work)2        1.104328   0.665022   1.661 0.096796 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1875.12  on 1352  degrees of freedom
Residual deviance:  522.56  on 1345  degrees of freedom
AIC: 538.56

Number of Fisher Scoring iterations: 8
```

Obrázek 11: Výpis druhého upraveného modelu auto jako řidič + chůze

Data jsem upravila tak, že jsem u prediktoru struktura domácnosti uvažovala pouze dvě kategorie. Zda respondent „má děti“ nebo „nemá děti“. Je vidět, že AIC opět kleslo, proto tuto změnu v modelu zachovám. Beru v úvahu, že tento model byl sestaven s celkovou cross–validační chybou 0,1953165.

Výsledný model vypadá takto

$$\log\frac{\pi(x)}{1-\pi(x)} = 8{,}26 - 0{,}03 * \textit{věk} - 1{,}12 * \textit{délka cesty} - 3{,}86 * \textit{vlastnictví řidičského průkazu} - 0{,}49 * \textit{pohlaví} - 0{,}97 * \textit{struktura domácnosti (děti)} + 1{,}47 * \textit{práce (nepracující)} + 1{,}1 * \textit{práce (studenti)}$$

a pravděpodobnost

$$\pi(x) = \frac{e^{8{,}26 - 0{,}03 * \text{vek} - 1{,}12 * \text{délka cesty} - 3{,}86 * \text{vlastnictví ridicského prùkazu} - 0{,}49 * \text{pohlaví} - 0{,}97 * \text{deti} + 1{,}47 * \text{nepracující} + 1{,}1 * \text{studenti}}}{1 + e^{8{,}26 - 0{,}03 * \text{vek} - 1{,}12 * \text{délka cesty} - 3{,}86 * \text{vlastnictví ridicského prùkazu} - 0{,}49 * \text{pohlaví} - 0{,}97 * \text{deti} + 1{,}47 * \text{nepracující} + 1{,}1 * \text{studenti}}}.$$

4.2 Auto jako řidič s kolem

Dále se budu snažit sestavit model, kdy se budu snažit zjistit, s jakou pravděpodobností zvolí respondent dopravní prostředek *auto jako řidič* nebo *kolo*. V této části jsem měla k dispozici 763 dat.

Nejdříve opět necháme vypsat model se všemi vysvětlujícími proměnnými.

```
Call:
glm(formula = MODE ~ P_age + H_pt_dist + T_length + P_dlic +
    P_gen + P_edu + as.factor(sám_pár_deti) + as.factor(P_work),
    family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3771  -0.4388  -0.2555  -0.0740   3.7010

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.544513   1.201561   2.118  0.03420 *
P_age                     -0.036877   0.014087  -2.618  0.00885 **
H_pt_dist                 -0.007021   0.038180  -0.184  0.85410
T_length                  -0.245863   0.044560  -5.518 3.44e-08 ***
P_dlic                    -2.773903   0.493675  -5.619 1.92e-08 ***
P_gen                      0.705597   0.338474   2.085  0.03710 *
P_edu                      0.266672   0.282269   0.945  0.34479
as.factor(sám_pár_deti)1   0.005051   0.576231   0.009  0.99301
as.factor(sám_pár_deti)2  -1.118964   0.622265  -1.798  0.07214 .
as.factor(P_work)1         1.111773   0.381973   2.911  0.00361 **
as.factor(P_work)2         0.357780   0.756591   0.473  0.63630
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 481.41  on 762  degrees of freedom
Residual deviance: 343.95  on 752  degrees of freedom
AIC: 365.95

Number of Fisher Scoring iterations: 7
```

Obrázek 12: Výpis modelu auto jako řidič + kolo

A opět se podívám na význam jednotlivých koeficientů. Z vytvořeného modelu mohu vyčíst, že pravděpodobnost, že respondent zvolí jako způsob dopravy *kolo*, bude klesat s rostoucím věkem. Dále je vidět, že čím vzdálenější bude zastávka hromadné dopravy, tak tím menší bude pravděpodobnost, že respondent zvolí *kolo*. Všímám si však toho, že tento parametr není statisticky významný a je to nejspíše opět způsobeno tím, že respondent při rozhodování mezi způsoby dopravy *auto* a *kolo* neuvažuje vzdálenost na zastávku. Se zvyšující se délkou cesty bude klesat pravděpodobnost volby *kola*. Dále je vidět koeficient, který opět snižuje

pravděpodobnost výběru způsoby dopravy *kolem,* a to v případě, že respondent vlastní řidičský průkaz. Oproti modelu minulému, kdy jsem zkoumala způsob dopravy *auto* a *chůze*, se zde zvyšuje pravděpodobnost volby *kola*, jestliže je respondent muž. Pravděpodobnost volby *kola* se dále zvyšuje při zvyšujícím se stupni vzdělání. Tento parametr není statisticky významný. Skutečnost, že respondent žije v domácnosti spolu s dětmi, snižuje pravděpodobnost volby *kola* jako dopravního prostředku, naopak lidé žijící v páru se na *kole* projedou raději. Vždy uvažujeme referenční skupinu respondentů, kteří žijí sami. Nakonec mohu říci, že lidé nepracující a studující volí raději jako způsob dopravy *kolo* s tím, že referenční skupina jsou pracující lidé. Celková cross-validační chyba je 0,09913573.

Opět z modelu odstraním statisticky nevýznamné proměnné.

```
Call:
glm(formula = MODE ~ P_age + T_length + P_dlic + P_gen + as.factor(sá
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4016  -0.4323  -0.2611  -0.0820   3.6852

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.97010    1.08472   2.738  0.00618 **
P_age                     -0.03566    0.01372  -2.600  0.00933 **
T_length                  -0.23676    0.04278  -5.534 3.13e-08 ***
P_dlic                    -2.62852    0.46819  -5.614 1.97e-08 ***
P_gen                      0.64325    0.32898   1.955  0.05055 .
as.factor(sám_pár_deti)1 -0.01209    0.57403  -0.021  0.98319
as.factor(sám_pár_deti)2 -1.09264    0.61713  -1.771  0.07664 .
as.factor(P_work)1         1.04802    0.37700   2.780  0.00544 **
as.factor(P_work)2         0.20058    0.76247   0.263  0.79250
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 481.41  on 762  degrees of freedom
Residual deviance: 344.90  on 754  degrees of freedom
AIC: 362.9

Number of Fisher Scoring iterations: 7
```

Obrázek 13: Výpis upraveného modelu auto jako řidič + kolo

I přes důkladné zkoumání, zda není prediktor vzdělání významný, jsem byla nucena jej i v tomto případě vyloučit. Dále v tomto modelu chybí vzdálenost na zastávku. Hodnota AIC klesla. Opět se zde nabízí možnost pozměnit agregaci prediktoru struktura domácnosti.

```
Call:
glm(formula = MODE ~ P_age + T_length + P_dlic + P_gen + as.factor(sám_
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3972  -0.4324  -0.2611  -0.0820   3.6849

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.95958    0.96280   3.074  0.00211 **
P_age                     -0.03566    0.01372  -2.600  0.00933 **
T_length                  -0.23676    0.04278  -5.534 3.13e-08 ***
P_dlic                    -2.62897    0.46764  -5.622 1.89e-08 ***
P_gen                      0.64329    0.32893   1.956  0.05050 .
as.factor(sám_pár_deti)1 -1.08201    0.35552  -3.043  0.00234 **
as.factor(P_work)1         1.04815    0.37695   2.781  0.00543 **
as.factor(P_work)2         0.20081    0.76239   0.263  0.79224
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 481.41  on 762  degrees of freedom
Residual deviance: 344.90  on 755  degrees of freedom
AIC: 360.9

Number of Fisher Scoring iterations: 7
```

Obrázek 14: Výpis druhého upraveného modelu auto jako řidič + kolo

Opět jsem brala v úvahu pouze to, zda respondenti „mají děti“ či „nemají děti“. Nejen, že se stal koeficient týkající se struktury domácnosti statisticky významný, ale opět zde kleslo AIC, proto tento model považuji za konečný. Tento model má celkovou cross-validační chybu 0,08011372.

Výsledný model vypadá takto

$$\log\frac{\pi(x)}{1-\pi(x)} = 2{,}96 - 0{,}03 * \textit{věk} - 0{,}24 * \textit{délka cesty} - 2{,}63 * \textit{vlastnictví řidičského průkazu} + 0{,}64 * \textit{pohlaví} - 1{,}08 * \textit{struktura domácnosti (děti)} + 1{,}05 * \textit{práce (nepracující)} + 0{,}2 * \textit{práce (studenti)}$$

a pravděpodobnost

$$\pi(x) = \frac{e^{2{,}96 - 0{,}03 * \text{vek} - 0{,}24 * \text{délka cesty} - 2{,}63 * \text{vlastnictví ridicského prùkazu} + 0{,}64 * \text{pohlaví} - 1{,}08 * \text{deti} + 1{,}05 * \text{nepracující} + 0{,}2 * \text{studenti}}}{1 + e^{2{,}96 - 0{,}03 * \text{vek} - 0{,}24 * \text{délka cesty} - 2{,}63 * \text{vlastnictví ridicského prùkazu} + 0{,}64 * \text{pohlaví} - 1{,}08 * \text{deti} + 1{,}05 * \text{nepracující} + 0{,}2 * \text{studenti}}}.$$

4.3 Auto jako řidič s veřejnou dopravou

Další v pořadí je model, kdy se zabývám zjištěním pravděpodobnosti, se kterou zvolí respondent dopravní prostředek *auto jako řidič* nebo *veřejná doprava*. V této části jsem měla k dispozici 1666 pozorování.

Jak je už zvykem, hned zpočátku této kapitoly si nechám vypsat model, který obsahuje všechny vysvětlující proměnné.

```
Call:
glm(formula = MODE ~ P_age + H_pt_dist + T_length + P_dlic +
    P_gen + P_edu + as.factor(sám_pár_deti) + as.factor(P_work),
    family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0593  -0.9202   0.1658   0.9275   1.8713

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.554871   0.486903   7.301 2.86e-13 ***
P_age                      0.001460   0.005067   0.288 0.773165
H_pt_dist                 -0.067968   0.018884  -3.599 0.000319 ***
T_length                  -0.004844   0.007340  -0.660 0.509324
P_dlic                    -2.453174   0.250085  -9.809  < 2e-16 ***
P_gen                     -0.727842   0.125009  -5.822 5.80e-09 ***
P_edu                     -0.037594   0.114683  -0.328 0.743059
as.factor(sám_pár_deti)1 -0.474797   0.230699  -2.058 0.039583 *
as.factor(sám_pár_deti)2 -1.104125   0.239049  -4.619 3.86e-06 ***
as.factor(P_work)1         0.407474   0.159499   2.555 0.010627 *
as.factor(P_work)2         2.645687   0.340686   7.766 8.11e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2260.2  on 1665  degrees of freedom
Residual deviance: 1645.7  on 1655  degrees of freedom
AIC: 1667.7

Number of Fisher Scoring iterations: 6
```

Obrázek 15: Výpis modelu auto jako řidič + VD

U tohoto modelu pozoruji změnu oproti minulým dvěma modelům. Koeficient týkající se věku mi říká, že jestliže se věk zvyšuje, bude se lehce zvyšovat i pravděpodobnost volby dopravního prostředku *veřejná doprava*. Věk je v tomto případě statisticky nevýznamný. Vzdálenost od zastávky se konečně stala statisticky významnou a říká, že jestliže se bude zvyšovat vzdálenost k zastávce, pak bude klesat pravděpodobnost volby *veřejné dopravy* jako dopravního prostředku. Délka cesty říká to samé – jestliže bude narůstat délka cesty, bude

respondent preferovat jízdu *autem*. Tento parametr je statisticky nevýznamný. Stejně jako v předchozích dvou modelech je vidět, že vlastnictví řidičského průkazu snižuje pravděpodobnost volby jakéhokoli jiného dopravního prostředku než *auta*. Skutečnost, že je respondent muž opět snižuje pravděpodobnost volby *veřejné dopravy* pro uskutečnění cesty. Se zvyšujícím se vzděláním také klesá pravděpodobnost volby *veřejné dopravy*. Tento parametr je statisticky nevýznamný. Jestliže domácnost obývá více členů (pár, děti) pak pravděpodobnost volby dopravního prostředku *veřejná doprava* klesá. Jako referenční proměnnou uvažujeme respondenty žijící samostatně. Nakonec mohu říci, že nepracující a studující lidé volí s větší pravděpodobností *veřejnou dopravu*, když referenční skupinou jsou pracující lidé. Tento model má celkovou cross-validační chybu 0.3838426.

Opět jsem se snažila odstranit statisticky nevýznamné parametry.

```
Call:
glm(formula = MODE ~ H_pt_dist + P_dlic + P_gen + as.factor(sám_pár_d
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0603  -0.9189   0.1705   0.9470   1.8652

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.52413    0.33972  10.374  < 2e-16 ***
H_pt_dist                 -0.06893    0.01872  -3.682 0.000231 ***
P_dlic                    -2.47801    0.24505 -10.112  < 2e-16 ***
P_gen                     -0.72574    0.12196  -5.951 2.67e-09 ***
as.factor(sám_pár_deti)1 -0.48810    0.22975  -2.124 0.033633 *
as.factor(sám_pár_deti)2 -1.13631    0.23052  -4.929 8.25e-07 ***
as.factor(P_work)1         0.43518    0.14869   2.927 0.003425 **
as.factor(P_work)2         2.63032    0.31295   8.405  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2260.2  on 1665  degrees of freedom
Residual deviance: 1646.3  on 1658  degrees of freedom
AIC: 1662.3

Number of Fisher Scoring iterations: 6
```

Obrázek 16: Výpis upraveného modelu řidič + VD

Snížit AIC z 1667,7 na 1662,3 se mi podařilo díky vyřazení statisticky nevýznamných proměnných. Bylo nutno odstranit prediktory věk, vzdělání a délku cesty. Takto sestavený model má celkovou cross-validační chybu 0.2828171.

Výsledný model tedy vypadá takto

$$\log\frac{\pi(x)}{1-\pi(x)} = 3{,}52 - 0{,}07 * \textit{vzdálenost od zastávky} - 2{,}48 * \textit{vlastnictví řidičského průkazu} - 0{,}73 * \textit{pohlaví} - 0{,}49 * \textit{struktura domácnosti (pár)} - 1{,}14 * \textit{struktura domácnosti (děti)} + 0{,}44 * \textit{práce (nepracující)} + 2{,}63 * \textit{práce (studenti)}$$

a pravděpodobnost

$$\pi(x) = \frac{e^{3.52 - 0{,}07 * \text{vzdálenost od zastávky} - 2{,}48 * \text{vlastnictví rid. průkazu} - 0{,}73 * \text{pohlaví} - 0{,}49 * \text{pár} - 1{,}14 * \text{děti} + 0{,}44 * \text{nepracující} + 2{,}63 * \text{studenti}}}{1 + e^{3.52 - 0{,}07 * \text{vzdálenost od zastávky} - 2{,}48 * \text{vlastnictví rid. průkazu} - 0{,}73 * \text{pohlaví} - 0{,}49 * \text{pár} - 1{,}14 * \text{děti} + 0{,}44 * \text{nepracující} + 2{,}63 * \text{student}}}$$

4.4 Auto jako řidič s autem jako spolujezdec

Nakonec budu sestavovat model, kdy se budu snažit zjistit, s jakou pravděpodobností zvolí respondent dopravní prostředek *auto jako řidič* nebo *auto jako spolujezdec*. K sestavení tohoto modelu jsem měla dispozici 871 pozorování.

Držím se zavedené struktury, proto na tomto místě uvádím model, který obsahuje všechny vysvětlující proměnné.

```
Call:
glm(formula = MODE ~ P_age + H_pt_dist + T_length + P_dlic +
    P_gen + P_edu + as.factor(sám_pár_deti) + as.factor(P_work),
    family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2540  -0.5270  -0.3055  -0.1898   2.9411

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.965210   0.984624   4.027 5.65e-05 ***
P_age                     -0.021037   0.010666  -1.972 0.048568 *
H_pt_dist                  0.012593   0.030208   0.417 0.676756
T_length                  -0.005016   0.014668  -0.342 0.732365
P_dlic                    -2.590682   0.362005  -7.156 8.28e-13 ***
P_gen                     -1.554344   0.245002  -6.344 2.24e-10 ***
P_edu                     -0.684256   0.197921  -3.457 0.000546 ***
as.factor(sám_pár_deti)1  0.120005   0.470580   0.255 0.798712
as.factor(sám_pár_deti)2 -0.876884   0.512424  -1.711 0.087035 .
as.factor(P_work)1         0.553568   0.270995   2.043 0.041080 *
as.factor(P_work)2         1.115234   0.538188   2.072 0.038247 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 890.23  on 870  degrees of freedom
Residual deviance: 550.92  on 860  degrees of freedom
AIC: 572.92

Number of Fisher Scoring iterations: 5
```

Obrázek 17: Výpis modelu řidič + spolujízda

Ke čtvrtému a zároveň poslednímu modelu mohu poznamenat, že pravděpodobnost volby dopravního prostředku *auto jako spolujezdec* bude klesat se zvyšujícím se věkem. Vzdálenost na zastávku bude zvyšovat pravděpodobnost volby *spolujízdy* s rostoucí vzdáleností k zastávce. Tento parametr však není statisticky významný, předpokládám, že při rozhodování se mezi tím, zda budu auto *řídit* nebo pojedu jako *spolujezdec* nejsem ovlivňována vzdáleností na zastávku. S rostoucí délkou cesty klesá pravděpodobnost volby *auta jako spolujezdec*. Tento parametr není statisticky významný. Skutečnost, že respondent je vlastníkem řidičského průkazu, snižuje pravděpodobnost, že pojede jako *spolujezdec*. Dále je také vidět, že jestliže je respondent muž, pak bude s menší pravděpodobností volit *spolujízdu* jako způsob dopravy. Model mi dále říká, že se zvyšujícím se vzděláním bude respondent také s menší pravděpodobností volit *spolujízdu*. Nutno podotknout, že parametr vzdělání je statisticky významný pouze v tomto modelu. Parametry týkající se struktury domácnosti mi říkají, že jestliže respondent žije v páru, tak se zvyšuje pravděpodobnost, že pojede jako

spolujezdec, a jestliže jsou v domácnosti i děti, tak se naopak pravděpodobnost volby *spolujízdy* snižuje. Co se pracovního zařazení týče, u studujících a nepracujících lidí je větší pravděpodobnost, že zvolí k uskutečnění cesty *auto jako spolujezdec*. Referenční skupina i v posledním modelu jsou pracující lidé. Tento model má celkovou cross-validační chybu 0,1337235.

Nyní se budu snažit model opět zjednodušit.

```
Call:
glm(formula = MODE ~ P_age + P_dlic + P_gen + P_edu + as.factor(sám_p
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2544  -0.5338  -0.2999  -0.1901   2.9411

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.93847    0.93866   4.196 2.72e-05 ***
P_age                     -0.02047    0.01059  -1.933 0.053253 .
P_dlic                    -2.58442    0.36017  -7.176 7.20e-13 ***
P_gen                     -1.56327    0.24427  -6.400 1.56e-10 ***
P_edu                     -0.67992    0.19735  -3.445 0.000571 ***
as.factor(sám_pár_deti)1  0.11940    0.46418   0.257 0.797002
as.factor(sám_pár_deti)2 -0.87528    0.50171  -1.745 0.081058 .
as.factor(P_work)1         0.56446    0.26851   2.102 0.035537 *
as.factor(P_work)2         1.14463    0.53593   2.136 0.032698 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 890.23  on 870  degrees of freedom
Residual deviance: 551.20  on 862  degrees of freedom
AIC: 569.2

Number of Fisher Scoring iterations: 5
```

Obrázek 18: Výpis upraveného modelu řidič + spolujízda

Zde jsem odstranila statisticky nevýznamné parametry vzdálenost na zastávku a délku cesty. Protože jsem usoudila, že změna proměnné struktura domácnosti by mohla model vylepšit jako v modelech minulých, provedla jsem další změnu.

```
Call:
glm(formula = MODE ~ P_age + P_dlic + P_gen + P_edu + as.factor(sám_pár_
    as.factor(P_work), family = "binomial", data = datarch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2506  -0.5294  -0.3009  -0.1903   2.9419

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                4.06176    0.80767   5.029 4.93e-07 ***
P_age                     -0.02079    0.01053  -1.975 0.048308 *
P_dlic                    -2.58396    0.36005  -7.177 7.14e-13 ***
P_gen                     -1.56084    0.24411  -6.394 1.62e-10 ***
P_edu                     -0.68045    0.19739  -3.447 0.000566 ***
as.factor(sám_pár_deti)1 -0.98386    0.27012  -3.642 0.000270 ***
as.factor(P_work)1         0.55750    0.26735   2.085 0.037047 *
as.factor(P_work)2         1.12796    0.53183   2.121 0.033930 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 890.23  on 870  degrees of freedom
Residual deviance: 551.27  on 863  degrees of freedom
AIC: 567.27

Number of Fisher Scoring iterations: 5
```

Obrázek 19: Výpis druhého upraveného modelu řidič + spolujízda

Opět jsem u prediktoru struktura domácnosti udělala stejnou změnu jako v minulých případech. Místo rozdělení na tři kategorie „žije sám“, „žije v páru“, „žije s dětmi“, jsem vytvořila pouze dvě kategorie „žije bez dětí“ a „žije s dětmi“. Tento model je sestaven s celkovou cross-validační chybou 0.1305305.

Výsledný model vypadá takto

$$\log\frac{\pi(x)}{1-\pi(x)} = 4{,}06 - 0{,}02 * \textit{věk} - 2{,}58 * \textit{vlastnictví řidičského průkazu} - 1{,}56 * \textit{pohlaví} - 0{,}68 * \textit{vzdělání} - 0{,}98 * \textit{struktura domácnosti (děti)} + 0{,}56 * \textit{práce (nepracující)} + 1{,}13 * \textit{práce (studující)}$$

a pravděpodobnost

$$\pi(x) = \frac{e^{4{,}06 - 0{,}02 * \text{vek} - 2{,}58 * \text{vlastnictví ridicského průkazu} - 1{,}56 * \text{pohlaví} - 0{,}68 * \text{vzdelání} - 0{,}98 * \text{deti} + 0{,}56 * \text{nepracující} + 1{,}13 * \text{studující}}}{1 + e^{4{,}06 - 0{,}02 * \text{vek} - 2{,}58 * \text{vlastnictví ridicského průkazu} - 1{,}56 * \text{pohlaví} - 0{,}68 * \text{vzdelání} - 0{,}98 * \text{deti} + 0{,}56 * \text{nepracující} + 1{,}13 * \text{studující}}}.$$

4.5 Test poměrem věrohodností

Nyní použiji test poměrem věrohodností, pro zjištění, zda se od sebe základní a upravené modely statisticky významně liší. Tento test provedu v programu R a výstup je následující. Obsahuje výpis modelů, které testujeme, počet stupňů volnosti jednotlivých modelů a také jednotlivé hodnoty maximálně věrohodné funkce. Dále je vidět rozdíl stupňů volnosti, vypočtenou chí-kvadrát statistiku, hodnotu p-value a nakonec rozhodnutí o statistické významnosti.

Postupně zde testuji všechny čtyři modely.

```
Likelihood ratio test

Model 1: MODE ~ P_age + H_pt_dist + T_length + P_dlic + P_gen + P_edu +
    as.factor(sám_pár_deti) + as.factor(P_work)
Model 2: MODE ~ P_age + T_length + P_dlic + P_gen + as.factor(sám_pár_deti)
    as.factor(P_work)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -260.13
2   8 -261.28 -3 2.2856     0.5153
```

Obrázek 20: Test poměrem věrohodnosti prvního modelu

```
Likelihood ratio test

Model 1: MODE ~ P_age + H_pt_dist + T_length + P_dlic + P_gen + P_edu +
    as.factor(sám_pár_deti) + as.factor(P_work)
Model 2: MODE ~ P_age + T_length + P_dlic + P_gen + as.factor(sám_pár_deti)
    as.factor(P_work)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -171.98
2   8 -172.45 -3 0.9434     0.8149
```

Obrázek 21: Test poměrem věrohodnosti druhého modelu

```
Likelihood ratio test

Model 1: MODE ~ P_age + H_pt_dist + T_length + P_dlic + P_gen + P_edu +
    as.factor(sám_pár_deti) + as.factor(P_work)
Model 2: MODE ~ H_pt_dist + P_dlic + P_gen + as.factor(sám_pár_deti) +
    as.factor(P_work)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -822.86
2   8 -823.17 -3 0.6136     0.8933
```

Obrázek 22: Test poměrem věrohodnosti třetího modelu

```
Likelihood ratio test

Model 1: MODE ~ P_age + H_pt_dist + T_length + P_dlic + P_gen + P_edu +
    as.factor(sám_pár_deti) + as.factor(P_work)
Model 2: MODE ~ P_age + P_dlic + P_gen + P_edu + as.factor(sám_pár_deti) +
    as.factor(P_work)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -275.46
2   8 -275.63 -3 0.3415     0.9521
```

Obrázek 23: Test poměrem věrohodnosti čtvrtého modelu

Bohužel jsem zjistila, že ani jeden test nevyšel statisticky významný.

4.6 Kvalita proložení dat

4.6.1 Akaikeho informační kritérium

	AIC	
	Základní model	**Upravený model**
Auto + Chůze	542,27	538,56
Auto + Kolo	365,95	360,9
Auto + VD	1667,7	1662,3
Auto + Spoluj.	572,92	567,27

Tabulka 5: Souhrn hodnot AIC

Jestliže sestavujeme model, je žádoucí, aby AIC klesalo. V tomto případě je vidět klesající tendence AIC u všech modelů.

4.6.2 Indexy determinace

Hodnoty indexů determinace jsou shrnuty v následující tabulce.

	Základní model	
	McFaddenův index determinace	Nagelkerkův index determinace
Auto + Chůze	0,72254	0,84361
Auto + Kolo	0,28554	0,35233

Auto + VD	0,27188	0,41546
Auto + Spoluj.	0,38115	0,50398

Tabulka 6: Souhrn hodnot indexů determinace u základních modelů

	Upravený model	
	McFaddenův index determinace	Nagelkerkův index determinace
Auto + Chůze	0,72132	0,84278
Auto + Kolo	0,28356	0,35011
Auto + VD	0,27161	0,41512
Auto + Spoluj.	0,38076	0,50359

Tabulka 7: Souhrn hodnot indexů determinace u upravených modelů

Což jsou zajímavé a velice rozdílné výsledky. Mohu však říci, že co se týče proložení kvality dat základních a upravených modelů, jsou na tom velice podobně.

První sestavený model se zdá být proložen dobře. Druhý a čtvrtý model nejsou vůbec proloženy tak dobře jako model první, což si ale zdůvodňuji nízkým počtem pozorování, která byla k dispozici v těchto částech sestavování modelu. Proč je ale tak špatně proložen model třetí, je mi záhadou. Možná nebylo správné hned na začátku agregovat do dopravního módu *veřejná doprava* hned čtyři dopravní prostředky (*tramvaj, vlak, trolejbus, autobus*). Nebo zde může docházet ke zkreslení díky tomu, že někteří respondenti použili pro uskutečnění cesty více než jeden způsob dopravy (například autobus a vlak). Každopádně, pro jednotlivé dopravní prostředky, které tvoří skupinu dopravního módu veřejná doprava, bych nebyla schopna sestavit modely dopravního chování, protože k tomu nemám dostatečný počet dat.

Chtěla jsem však poznamenat, že popsat lidské chování pomocí matematických modelů není jednoduché, protože každý člověk je unikátní.

Závěr

V závěru bych ráda shrnula výsledky, čeho bylo dosaženo a čeho ne. U té příležitosti připomenu, co vlastně bylo cílem práce. Cílem bylo vytvoření modelu, který bude odhadovat pravděpodobnost, s jakou respondent zvolí určitý dopravní prostředek k uskutečnění své cesty. Dále potom zjišťovat, které sociodemografické aspekty mají vliv na respondentovo rozhodnutí. Tento model jsem měla sestavit na základě dat z Jihomoravského kraje, která mi byla poskytnuta CDV. Předpokládala jsem, že budu užívat multinomickou regresi.

Po očištění dat jsem se rozhodla volit místo multinomické regrese čtyři obyčejné logistické regrese, poté jsem chtěla rozdělit data na trénovací a testovací sadu dat a zjistila jsem, že k tomu nemám dostatečné množství dat. Problém jsem vyřešila pomocí cross validace, která právě umí pracovat s malými vzorky dat. Takže jsem sledovala významnost jednotlivých parametrů, pomocí Akaikeho informačního kritéria sestavila nejlepší model a nakonec jsem vypočítala celkovou cross – validační chybu modelu, která vypovídá o výkonnosti modelu na datech, která software nikdy neviděl.

Ze všech dat, která byla poskytnuta, jsem vybrala pouze data týkající se věku respondenta, vzdálenosti od zastávky, délky cesty, pohlaví respondenta, vzdělání respondenta, struktury domácnosti, práce a vlastnictví řidičského průkazu. Na základě těchto dat, která jsem upravila a vložila do modelu, jsem po interpretaci vzniklých regresních koeficientů zjistila, že sestavené modely dávají smysl.

Nyní už k samotnému sestavování modelů. Sestavila jsem čtyři modely logistické regrese pro *chůzi, kolo, veřejnou dopravu* a *spolujízdu* s tím, že referenční proměnná ve všech modelech byl dopravní mód *auto jako řidič*. Podařilo se mi sestavit modely a postupně je upravit na základě vylepšení či vyřazování nevýznamných parametrů a snižování Akaikeho informačního kritéria. Model druhý, kdy uvažujeme model *auto* + *kolo* má nejmenší celkovou chybu zatímco model *auto* + *veřejná* doprava má chybu největší. Dle indexů determinace ale vidím, že model *auto* + *chůze* nejkvalitněji prokládá data. Podle mého názoru se mi nejlépe podařilo namodelovat model první *auto* + *chůze*, protože tyto způsoby dopravy jsou od sebe v rámci uvažovaných dopravních módů charakterově nejdále. Výsledky druhého a čtvrtého modelu bych raději brala s rezervou, protože pro vytváření těchto modelů jsem neměla dostatečný počet dat. A co se třetího modelu týče, jak už jsem zmínila výše, zde musela nastat nějaká chyba nejspíše při agregaci dat.

Na této práci jsem nejvíce ocenila, že jsem mohla aplikovat nabyté teoretické poznatky do praxe, vše si vyzkoušet na datech, vidět, že výsledky dávají smysl a pokud ne, existuje vysvětlení, proč model nefunguje, tak jak má a také si vyzkoušet práci s úpravou dat.

Data k práci nepřikládám, jelikož jsem zavázána CDV tato data nešířit.

Seznam použité literatury

[1] A. Agresti: Categorical Data Analysis, John Wiley & Sons, 2. vydání, New Jersey 2002

[2] J. Anděl: Základy matematické statistiky, Univerzita Karlova v Praze, Praha 2002

[3] Dokoupil P., Aplikovaná logistická regrese, diplomová práce, rok 2012, dostupná na: http://theses.cz/id/a5zu7h/Diplomov_prce_-_Dokoupil_Petr.pdf

[4] D. Hosmer, S.Lemeshow: Applied Logistic Regression, 2. vydání John Wiley and Sons, Inc. 2000

[5] James G., Witten D., Hastie T., Tibshirani R., An Introduction to Statistical Learning, Springer, New York 2013

[6] P. Kunderová, Základy pravděpodobnosti a matematické statistiky, Univerzita Palackého v Olomouci, Olomouc 2004

[7] M. Meloun, J. Mlitký: Statistická analýza experimentálních dat, 2. vydání Academia 2004

[8] Pecáková, I., Logistická regrese s vícekategoriální vysvětlovanou proměnnou. Acta Oeconomica Pragensia 1. 2007

[9] F. Zlámal, Logistická regrese v R, bakalářská práce, rok 2013, dostupná na: http://is.muni.cz/th/78448/prif_b_b1/bakalarska_prace_Filip_Zlamal.pdf

[10] Zvára, K., Regrese, Matfyzpress, Praha 2008

[11] http://en.wikipedia.org

[12] http://oldweb.izip.cz/ds3/hypertext/AJDLQ.htm

Přílohy

Příloha 1: Dotazník pro domácnosti

DOTAZNÍK PRO DOMÁCNOSTI

- Do domácnosti se počítají všechny osoby (včetně Vás), které spolu dlouhodobě žijí.
- Domácnost může tvořit i jediná osoba (jednočlenná domácnost).

QA1) Kolik osob včetně dětí žije soustavně ve Vaší domácnosti? (včetně Vás)

POKYN: VYZNAČTE POČET. KONTROLA: QA1_1 = QA1_2 + QA1_3

QA1	

QA2) Kolik z těchto osob je mladších 6 let?

POKYN: VYZNAČTE POČET.

QA2	

QA3) Kolik z těchto osob je ve věku 6-18 let?

POKYN: VYZNAČTE POČET.

QA3	

QA4) Žijí ve Vaší domácnosti „částečně" ostatní osoby? (např. studující děti, děti z dřívějších vztahů, osoby, o které pečujete atd.)

POKYN: VYZNAČTE JEDNU Z MOŽNOSTÍ

1	Ano
2	Ne

QA5) Které z těchto osob žijí ve vaší domácnosti?

POKYN: VYZNAČTE JEDNU NEBO VÍCE MOŽNOSTÍ

1	Partner(ka)
2	Dítě/děti
3	Rodiče/prarodiče
4	Ostatní příbuzní

5	Ostatní osoby

QA6) Jak daleko (pěšky) od vašeho domova je nejbližší zastávka veřejné dopravy? (Autobus, vlak, tramvaj, trolejbus)

POKYN: VYZNAČTE POČET V MINUTÁCH

.... minut

QA7) O jaké zastávky jde?

POKYN: VYZNAČTE JEDNU NEBO VÍCE MOŽNOSTÍ

1	Autobusová
2	Tramvajová
3	Trolejbusová
4	Vlaková

QA8) Uveďte prosím čistý měsíční příjem Vaší domácnosti.

POKYN: VYZNAČTE JEDNU Z MOŽNOSTÍ

1	do 15 000 Kč
2	15 001 Kč – 20 000 Kč
3	20 001 Kč – 25 000 Kč
4	25 001 Kč – 30 000 Kč
5	30 001 Kč – 35 000 Kč
6	35 001 Kč – 50 000 Kč
7	50 001 Kč – 60 000 Kč
8	Nad 60 000 Kč

99	Neví / nechce odpovědět

QA9) Kolik jízdních kol je ve Vaší domácnosti?

POKYN: VYZNAČTE POČET

QA9	

QA10) Kolik mopedů/motocyklů je ve Vaší domácnosti?

POKYN: VYZNAČTE POČET

QA10	

Uveďte prosím jednotlivé automobily ve Vaší domácnosti a doplňte k nim příslušné parametry.

	Číslo vozu	Vůz 1	Vůz 2	Vůz 3	Vůz 4	Vůz 5
QA11	Značka/model:					
QA12	Služební vůz	ano/ne	ano/ne	ano/ne	ano/ne	ano/ne
QA13	Najeté km za rok:					
QA14	Typ paliva:					
QA15	Roční dálniční známka:	ano/ne	ano/ne	ano/ne	ano/ne	ano/ne
QA16	Vestavěná navigace:	ano/ne	ano/ne	ano/ne	ano/ne	ano/ne

Příloha 2: Dotazník pro osoby

DOTAZNÍK PRO OSOBY

Odpovězte prosím na následující otázky pro všechny členy domácnosti starší 6 let (včetně).

	Podle stáří	Nejstarší	2. nejstarší	3. nejstarší	4. nejstarší	5. nejstarší
QB1	Číslo osoby	1	2	3	4	5
QB2	Rok narození					
QB3	Pohlaví	m/ž	m/ž	m/ž	m/ž	m/ž

QB4) Jaké je Vaše nejvyšší ukončené vzdělání

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ

	Číslo osoby	1	2	3	4	5
1	Neukončené základní vzdělání					
2	Základní vzdělání					
3	Střední vč. vyučení - bez maturity					
4	Úplné střední s maturitou (odborné i všeobecné)					
5	Nástavbové studium (vč. pomaturitního studia)					
6	Vyšší odborné vzdělání					
7	Bakalářské studium					
8	Magisterské studium					
9	Doktorské studium					

QB5) Jaké je Vaše ekonomické postavení?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ

	Číslo osoby	1	2	3	4	5
1	Zaměstnanec řadový					
2	Zaměstnanec vedoucí					
3	Podnikatel bez zaměstnanců, OSVČ					
4	Podnikatel se zaměstnanci					
5	Nezaměstnaný					
6	Nepracující důchodce					
7	Mateřská, rodičovská, v domácnosti					
8	Student, žák, učeň					

QB5) Kolik hodin týdně tato osoba pracuje?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE POČET HODIN

Číslo osoby	1	2	3	4	5
QB5					

QB6) Můžete si tato osoba zvolit začátek pracovní doby?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Číslo osoby	1	2	3	4	5
QB6	a/n	a/n	a/n	a/n	a/n

QB7) Může tato osoba část práce vykonávat i z domu?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Číslo osoby	1	2	3	4	5
QB7	a/n	a/n	a/n	a/n	a/n

QB8) Má tato osoba na cestách přístup k internetu (chytrý telefon, tablet, PC, notebook atd.)?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Číslo osoby	1	2	3	4	5
QB8	a/n	a/n	a/n	a/n	a/n

QB9) Má tato osoba řidičský průkaz skupiny B?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Číslo osoby	1	2	3	4	5
QB9	a/n	a/n	a/n	a/n	a/n

QB10) Využívá tato osoba nějakou slevu na hromadnou dopravu?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

	Číslo osoby	1	2	3	4	5
1	Ne	a/n	a/n	a/n	a/n	a/n
2	Časová jízdenka (týdenní, měsíční, roční)	a/n	a/n	a/n	a/n	a/n
3	Slevová jízdenka, speciální průkaz atd.	a/n	a/n	a/n	a/n	a/n

QB11)Má tato osoba na pracovišti k dispozici soukromé parkovací místo?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Číslo osoby	1	2	3	4	5
QB11	a/n	a/n	a/n	a/n	a/n

QB12) Má tato osoba doma k dispozici soukromé parkovací místo (pronajaté nebo vlastní)?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Číslo osoby	1	2	3	4	5
QB12	a/n	a/n	a/n	a/n	a/n

QB13) Má tato osoba k dispozici tyto dopravní prostředky jako řidič?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

	Číslo osoby	1	2	3	4	5
1	Motocykl nebo moped	a/n	a/n	a/n	a/n	a/n
2	Jízdní kolo	a/n	a/n	a/n	a/n	a/n

QB14) Jak často má tato osoba (jako řidič) k dispozici osobní automobil?

POKYN: PRO KAŽDOU Z OSOB VYZNAČTE JEDNU Z MOŽNOSTÍ.

	Číslo osoby	1	2	3	4	5
1	Vždy					
2	Občas					
3	Nikdy					

Příloha 3: Dotazník pro cesty

DOTAZNÍK PRO CESTY

QC1) Osoba číslo:

POKYN: UVEĎTE ČÍSLO OSOBY Z DOTAZNÍKU PRO DOMÁCNOSTI.

QC1	

QC2) Dnešní datum.

POKYN: UVEĎTE DNEŠNÍ DATUM.

QC2	

QC3) Vyšel/a tato osoba dnes z domu?

POKYN: VYZNAČTE JEDNU Z MOŽNOSTÍ.

1	Ano
2	Ne. Uveďte prosím důvod.

QC4) Kde byl výchozí bod pro první cestu?

POKYN: PROSÍM VYPLŇTE

1	Domov, bydliště
3	Jiné místo Adresa:

QC5) Kdy tato první cesta začala?

POKYN: PROSÍM VYPLŇTE HODINU A MINUTU (HH:MM)

Pořadí cesty	1.	2.	3.	4.	5.	6.	7.
QC5	h:m	h:m	h:m	h:m	h:m	h:m	h:m

QC6) Za jakým ÚČELEM jste cestu podnikli?

POKYN: PRO KAŽDOU CESTU PROSÍM UVEĎTE JEN JEDNU MOŽNOST.

	Pořadí cesty	1.	2.	3.	4.	5.	6.	7.
1	Cesta do práce							
2	Pracovní cesta							
3	Vzdělání/škola							
4	Doprovod/vyzvednutí osob							
5	Nákup							
6	Volný čas							
7	Návrat domů							
8	Jiný účel a sice:							

QC7) Jaký DOPRAVNÍ PROSTŘEDEK jste v průběhu této cesty použili?

POKYN: POKUD JSTE POUŽILI VÍCE PROSTŘEDKŮ, UVEĎTE JE PROSÍM VŠECHNY.

	Pořadí cesty	1.	2.	3.	4.	5.	6.	7.
1	Pěšky							
2	Jízdní kolo							
3	Autobus							
4	Tramvaj, metro							

5	Vlak							
6	Moped, motocykl							
7	Osobní automobil, řidič							
8	Osobní automobil, spolujezdec							
9	ostatní, a sice:							

QC8) Doprovázeli jste na této cestě JINÉ OSOBY?

POKYN: PRO KAŽDOU Z CEST VYZNAČTE JEDNU Z MOŽNOSTÍ (ANO/NE)

Pořadí cesty	1.	2.	3.	4.	5.	6.	7.
QC7	a/n	a/n	a/n	a/n	a/n	a/n	a/n

QC9) Kolik z nich bylo dětí a kolik dospělých?

POKYN: ODPOVÍDAJÍ JEN TI, KTEŘÍ U PŘÍSLUŠNÉ CESTY V QC8 UVEDLI "ANO". UVEĎTE POČET.

	Pořadí cesty	1.	2.	3.	4.	5.	6.	7.
QC9_1	Děti							
QC9_2	Dospělých							

QC10) KDE byl CÍL této cesty?

POKYN: PROSÍM VYPLŇTE ADRESU CO NEJPŘESNĚJI

QC10	Adresa:

QC11) KDY daná osoba do cíle této cesty dorazila?

POKYN: PROSÍM VYPLŇTE ČASOVÝ ÚDAJ V HODINÁCH A MINUTÁCH (HH:MM)

QC9	hh:mm

QC12) Odhadněte prosím DÉLKU této cesty (v kilometrech)

POKYN: PROSÍM VYPLŇTE VZDÁLENOST V KILOMETRECH.

QC10	... km

POKYN: DALŠÍ CESTY NEBO CESTU ZPĚT UVEĎTE DO DALŠÍHO SLOUPCE.