

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

DIPLOMOVÁ PRÁCE

Diskrétní analýza volby dopravního prostředku



Vedoucí diplomové práce:

Ing. Petr Šenk, Ph.D.

Rok odevzdání: 2014

Vypracovala:

Bc. Tereza Šolcová

AME, II.ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně pod vedením pana Ing. Petra Šenka, Ph.D. s použitím uvedené literatury.

V Olomouci dne 20. března 2014

Poděkování

Ráda bych poděkovala vedoucímu diplomové práce panu Ing. Petrovi Šenkovi, Ph.D. za spolupráci i za čas, který mi věnoval při konzultacích.

Obsah

Úvod	6
1. Úvod do problematiky volby dopravního prostředku	7
1.1. Předpověď poptávky v dopravě	7
1.2. Pozadí analýzy diskrétní volby	7
1.3. Aplikace analýz diskrétní volby v dopravě	8
2. Modelování dopravní poptávky	9
2.1. Čtyřstupňový model	9
2.2. Prvky sady možností (choice setu)	11
2.3. Teorie volby založená na utilitě	12
2.3.1. Základní konstrukce funkce utility	12
2.3.2. Koncept deterministické volby	13
2.3.3. Model pravděpodobnostní volby	14
2.3.4. Komponenty deterministické části funkce utility	14
2.3.5. Specifikace chybové složky	16
2.4. Multinomický logitový model	16
2.4.1. Nezávislost nepodstatných vlastností alternativ	19
2.4.2. Odhad modelu	21
2.4.3. Metoda maximální věrohodnosti	21
2.5. Požadavky na data	22
2.5.1. Stated, Revealed preference data	23
2.5.2. Metody pro sběr dat alternativních způsobů dopravy	24
2.6. Posouzení kvality modelu	25
3. Data	28
4. Biogeme	34
4.1. Jak Biogeme funguje	34
4.2. Analýza pomocí programu Biogeme	36
5. Sestavení modelu volby dopravního prostředku	38
5.1. Základní model	38
5.2. Podrobnější model	44
5.3. Modely pro jednotlivé účely	48
5.3.1. Cesta do práce	48

5.3.2. Cesta do školy/za vzděláním	50
5.3.3. Cesta za volno-časovými aktivitami	52
5.3.4. Cesta na nákup	53
Závěr	56
Literatura	57

Úvod

Téma diplomové práce je *Diskrétní analýza volby dopravního prostředku*. Cílem této diplomové práce je vytvoření prvního modelu volby dopravního prostředku v ČR, který bude sestaven výhradně z průzkumu odvozených preferencí. Vytvořený model bude sestaven na základě dat zjištěných z průzkumu, který byl proveden v Jihomoravském kraji. Postupně bude sestaven model, který bude základní – kde budou zahrnuty pravděpodobně jen atributy specifické pro jednotlivé dopravní prostředky. Dále je do modelu zahrnuta i sociodemografie jedinců a nakonec budeme sledovat, jak se budou měnit hodnoty parametrů, jestliže budeme modely uvažovat pro specifické účely (cesta do práce, nákup, volno-časové aktivity,...).

Podobné práce byly již určitě zpracovány, například česká publikace *Cesty městem, O racionalitě každodenního chování* od Markéty Braun Kohlové. Nejen, že se na této publikaci dá ověřit správnost českých statistických názvů, ale tato publikace je jedna z mála, která do modelu postupně zahrnuje velké množství vysvětlujících proměnných a nakonec je precizně interpretovala.

Práce je rozdělena do 5 kapitol. První kapitola nabízí hrubý úvod do problematiky, kterou se budeme v této diplomové práci zabývat. Čtenář zde nalezne hlavně odkazy na zlomová díla analýzy dopravní volby. V kapitole druhé se čtenář seznámí s poměrně zdlouhavou teorií týkající se diskrétní volby dopravního prostředku. V kapitole třetí jsou čtenáři přiblížena data, pro která bude vytvořen model volby dopravního prostředku. Ve čtvrté kapitole je seznámení se softwarem, pomocí kterého jsou data zpracovávána. A v kapitole poslední jsou zveřejněny a interpretovány některé specifické modely volby dopravního prostředku.

1 Úvod do problematiky volby dopravního prostředku

1.1 Předpověď poptávky v dopravě

Předpověď poptávky je důležitým prvkem v analýze dopravních systémů. Snažíme se předpovídat chování uživatelů dopravních prostředků. Zejména jde potom o to, zjistit odezvu uživatelů na změny, které byly vyvolány zavedením nových služeb, investováním do infrastruktury,...

Dřívější studie velkých investic do regionální silniční sítě a rychlý pokrok ve výpočetní technice vedl k vývoji postupů při prognózách cest a určení objemu dopravy na silniční síti. Tyto modelovací postupy byly neustále lepší a rozšířenější a byly aplikovány celosvětově. Jsou zdokumentovány (např. Manheim 1979, Stoper a Meyburg 1975, Hutchinson 1974).

Na konci 60. let a počátkem 70. let nastal velký zlom v analýze dopravních systémů. Došlo k posunu k multimodální systémům a začalo se uvažovat nad cenami a konstrukcí nových dopravních zařízení. To vedlo k dalšímu zlepšení dopravních prognóz (např. Brand a Manheim 1973, Werner 1983)

Největší inovace však v analýze dopravní poptávky byla, když byly vyvinuty disgregované modely poptávky po dopravě, založené na modelech diskrétní volby. Hlavní myšlenkou v disgregovaných modelech poptávky po dopravě bylo, že se studovaná plocha rozdělila do zón. Potom původní a cílové destinace všech cestujících byly reprezentovány jediným bodem v zóně, obvykle centroidem. Významným efektem tohoto přístupu byla možnost získat spolehlivé odhady pro modely s mnohem širším rozsahem vysvětlujících proměnných (např. Oi a Shuldiner 1962, Fleet a Robertson 1968, McCarthy 1969).

1.2 Pozadí analýzy diskrétní volby

Klasický problém analýzy diskrétní volby je modelování volby dopravního prostředku ze sady vzájemně jedinečných možností, ale na druhou stranu zcela postihující (vyčerpávající) všechny možnosti. Většinou se užívá k řešení analýzy diskrétní volby princip maximalizace užitku.

Je v podstatě nemožné specifikovat a odhadovat modely diskrétní volby, které se vždy shledají s úspěchem v prognózách ve výběru dopravní možnosti pro každého jednotlivce. Proto byl převzat koncept náhodného užitku, který se nejdříve objevil v psychologii (Thurston 1927). Skutečné užitky alternativních možností jsou uvažovány jako náhodné proměnné,

takže pravděpodobnost, že určitá alternativní možnost bude vybrána, je definována jako pravděpodobnost, že má největší utilitu mezi všemi ostatními možnostmi.

Problém diskretní volby je středem zájmu pro výzkumníky mnoho let v různých odvětvích. Počátky pravděpodobnostních modelů volby jsou zakořeněny v matematické psychologii (např. Thurston 1927, Luce 1959, Marschak 1960, Luce a Suppes 1965, Bock a Jones 1968, Tversky 1972). Bylo sepsáno také několik revizí ekonometrických aplikací modelů diskretní volby (např. McFadden 1982, Manski 1981, Amemiya 1981).

1.3 Aplikace analýz diskretní volby v dopravě

Dřívější aplikace modelů diskretní volby v dopravě byly vytvářeny pro binární volbu cestovních módů – pro volbu dopravního prostředku uvažujeme pouze dvě alternativy (např. Warner 1962, Lisco 1967, Stoper 1969, Wigner 1973, Watson 1974). Některé z těchto studií se zaměřovaly na odhad hodnoty času, kompromis mezi cestovním časem a cestovními náklady.

Na počátku 70. let se začaly objevovat modely s více než dvěmi možnostmi volby a do modelů se postupně začaly dostávat frekvence cest, cílová destinace cest, vlastnictví auta, specifikace obytné zóny, typ bydlení (např. Ben-Akiva 1973,1974, McFadden 1974, Lerman a Ben-Akiva 1975).

2 Modelování dopravní poptávky

2.1 Čtyřstupňový model

Čtyřstupňový model, který uvádí [10], je obvykle používán k dopravním prognózám, kdy nás zajímá odhad počtu vozidel nebo lidí, kteří použijí objem výkonů dopravního systému v budoucnosti. Například nám může dodat odhad, který nám bude říkat, jaký bude počet vozidel na silnici nebo pasažérů využívajících vlak. K získání odhadů využíváme data, která jsme zjistili na základě dotazníkového šetření, popřípadě dat zjištěných například z jízdních řádů nebo ceníků. Z těchto dat jsme schopni vyvinout poptávkový model.

Tento model dopravního plánování byl vyvinut Chicago Area Transportation Study a byl od počátku zaměřen na dopravu. Jak název kapitoly napovídá, skládá se ze čtyř kroků.

- Generování objemů dopravy v zónách – tato analýza se zaměřuje na bydliště respondenta a generováním objemů dopravy je pak myšlena funkce sociálních a ekonomických atributů domácnosti
- Distribuce přepravních vztahů mezi zónami – tento krok spojuje výchozí a cílovou destinaci cestovatele pomocí tzv. matice cest. Tato matice zobrazuje počet cest uskutečněných z každého počátku do každé cílové destinace. Tento krok byl vyvinut jako poslední komponenta čtyřstupňového modelu.
- Dělbá přepravní práce (model volby dopravního prostředku pro jednotlivé mezizonální vztahy) – umožní analytikovi určit, který dopravní mód bude použit. To díky matici z druhého kroku, která nám říká, které cesty budou uskutečněny.
- Rozvrhování přepravních vztahů (v osobách) na síť linek HD, resp. dopravních vztahů (ve vozidlech) na komunikační síť.

Tato práce se zabývá pouze třetím krokem čtyřstupňového modelu, proto se dále budu vyjadřovat pouze k tomuto kroku.

Uvažujme zde jednotlivce, který se potřebuje dostat za určitým účelem a rozhoduje se mezi dostupnými dopravními módy. Dostupná sada dopravních alternativ se nazývá sada možností (choice set). Úkolem je porozumět rozhodnutí, které učinil jedinec v případě výběru určitého dopravního módu a popřípadě být schopen určit předpověď rozhodnutí jednotlivce. Můžeme se také zajímat o určení vlivu na pravděpodobnost volby některého z dopravních prostředků. Tyto úkoly lze rozuzlit s daty, která jsou napozorovaná, a která má analytik k dispozici. Tento přístup, kdy si jednatel musí vybrat jednu možnost ze sady dostupných módů nazýváme diskretní analýzou volby dopravního prostředku. Teorii diskretní volby

můžeme použít i v jiných oblastech než v dopravě. Každý z nás často musí čelit výběru mezi diskrétními alternativami, například při výběru zaměstnání, školy,...

Pokud bychom pozorovali jednotlivce, který se každý den rozhoduje, jaký dopravní mód zvolí, při stejných podmínkách, pak můžeme předpokládat, že jeho volby se mohou v průběhu času lišit. To vysvětlujeme tak, že analytik není schopen zjistit úplně všechna data a ani respondent nemusí mít perfektní informace o alternativních módech dopravy. S tímto však v modelu počítáme.

Diskrétní modely volby mohou nabývat mnoha forem, například:

- Binární logit
- Binární probit
- Multinomický logit
- Multinomický probit
- Podmíněný logit
- Nested logit
- Mixed logit

Který model zvolíme, to záleží na datech a na tom, co od modelu očekáváme a požadujeme. Nejčastěji bývá ale využíván kvůli své jednoduchosti Multinomický logitový model. O Multinomickém logitovém modelu se budu rozepisovat níže, kde popíšu i důvody, proč je zmiňovaný model nejpoužívanější.

Vytvářet model dopravy je užitečná věc. Zde by se mohl čtenář zarazit a ptát se, proč složitě modelovat model volby dopravního prostředku, když můžeme provést například dopravní průzkum na pozemních komunikacích a železnicích. Dopravní průzkum obvykle bývá docela drahý a výsledky nám ukazují pouze současnou situaci. Z těchto dat nejsme schopni předvídat nebo zjišťovat, co by se stalo, kdybychom například zrušili jednu linku autobusů. Oproti tomu dobře vytvořený model dopravy nám je schopen říci, jak je který cestující citlivý na jaké atributy a jak významně tyto atributy ovlivňují chování cestujících. Navíc nám zde stačí prozkoumat pouze vzorek obyvatelstva, znát zájem o dopravu ve sledovaném úseku a mít jakýsi názor cestujícího na jednotlivé atributy.

Pokud nebude řečeno jinak, je většina zbytku druhé kapitoly převzata z [8].

Existují dva způsoby, kterými můžeme modelovat chování poptávky po dopravě. Jeden přístup přímo modeluje agregátní podíl všech nebo určité části rozhodovatelů, kteří si vybrali alternativu na základě funkce charakteristik alternativ a sociodemografických atributů skupiny. Tento přístup bývá označován jako agregovaný přístup. Druhý přístup rozpoznává,

že celková poptávka po dopravě je výsledkem chování každého jednotlivce. Do modelu se zaznamenávají odpovědi jednotlivců jako funkce dostupných charakteristik alternativ a sociodemografické vlastnosti každého jednotlivce. Tento přístup nazýváme disagregovaný.

Disagregovaný přístup má několik důležitých výhod před agregátním přístupem. Za prvé, disagregovaný přístup vysvětluje, proč jednotlivci dělají konkrétní volby za určitých okolností a je schopen lépe odrážet změny v chování jednotlivce. Je to díky tomu, že dochází ke změnám právě a pouze v charakteristikách jednotlivce. Za druhé, disagregovaný přístup má velkou pravděpodobnost, že bude přenositelný na jiné místo a na jiný čas, a to díky jeho kauzální povaze, což je důležité pro předpověď. Za třetí, modely diskrétní volby se začínají ve velké míře používat k porozumění chování. Za čtvrté, disagregovaný přístup je více efektivní, co se týče spolehlivosti sesbíraných dat vůči ceně. Disagregovaná data poskytují podstatnou změnu v chování zájmu a v determinantách chování, což umožňuje efektivní odhad parametrů v modelu. A konečně za páté, disagregované modely, jestliže jsou správně specifikované, tak získáme nevychýlené odhadnuté parametry, zatímco agregátní modely produkují vychýlené parametry. V naší analýze je použit disagregovaný přístup.

2.2 Prvky sady možností (choice setu)

Ráda bych se zde zmínila také o prvcích, co vystupují v sadě alternativních možností. Pozorujeme jednotlivce, kteří si vybírají jednu volbu z široké pestrosti rozhodovacích kontextů. Proces rozhodování by se měl sestávat z uvažování nad dostupnými alternativami jednotlivce, následně by se měly zhodnotit atributy každé alternativy jednotlivce a nakonec by měl jednotlivce užít rozhodovací pravidlo, aby byl schopen vybrat jednu nejlepší možnost ze všech uvažovaných. V praxi se to však neděje takto strukturovaně, jedinec vybírá konkrétní variantu aniž by prošel strukturovaným procesem. Teď už ale k jednotlivým prvkům trochu detailně:

- Rozhodovatel – může to být jedinec, skupina, instituce, která má odpovědnost učinit rozhodnutí. Rozhodnutí rozhodovatele záleží na určité situaci, jediné, co je jasné, že různí rozhodovatelé čelí různým rozhodovacím situacím a rozhodují se jinak, protože mají jiný vkus, náklonnost, chuť.
- Alternativy – každý jednotlivce má sadu možností složenou z těch možností, které má k dispozici. Tyto alternativy nabývají konečného počtu, proto bychom měli dobře zvážit, které alternativy budeme uvažovat. Občas se zde setkáváme s určitými

omezeními. Například vlastnictví řidičského průkazu, ekonomické omezení nebo charakteristiky jedince.

- Atributy alternativ – alternativy jsou v procesu volby charakterizovány sadou hodnot atributů. Vyjadřujeme jimi atraktivitu alternativy. Atributy alternativ mohou být generické (k alternativám se přiřadí stejné atributy) nebo specifické pro každou variantu. I zde bychom se měli zamýšlet nad tím, které atributy alternativ do modelu zahrneme. Měli bychom mít model dobře popsany, avšak neměli bychom zacházet do zbytečných podrobností.
- Rozhodovací pravidla – jednotlivce se odvolává na rozhodovací pravidla, aby byl schopen vybrat jednu možnost ze sady možností. Říkáme, že jednotlivce používá racionální rozhodování, jestliže jsou v procesu zachovány dvě podmínky – konzistenci a tranzitivitu. Pod slovem konzistence si můžeme představit, že za stejných podmínek si bude rozhodovatel vybírat v opakovaných výběrech stejné volby. Tranzitivita potom znamená, že alternativy můžeme jednoznačně porovnat podle preferenčního měřítka. Je zde několik pravidel, která připadají v úvahu použít, nejčastěji se však používá rozhodovací pravidlo známé jako maximalizace utility. Ta je založena na dvou pojmech. Každá alternativa je charakterizována vektorem atributů. Tento vektor musíme být schopni převést na skalár (tzv. užitek, který nám přinese použití určité dostupné alternativy). A za druhé, jednotlivce vybírá alternativu s nejvyšší hodnotou užitku.

2.3 Teorie volby založená na Utilitě

2.3.1 Základní konstrukce funkce utility

Udává hodnotu vztahující se k jednotlivci. Obecně je utilita odvozena od atributů alternativ. Pravidlo maximální utility je založeno na tom, že jednotlivce vybere jednu určitou alternativu. Tato alternativa musí být dostupná pro jednotlivce a musí maximalizovat užitek jednotlivce. To znamená, že jestliže je vybrána určitá alternativa, tak žádná jiná alternativní možnost nemá hodnotu utility vyšší. Dále pravidlo říká, že funkce obsahuje atributy alternativ a charakteristiky jednotlivců a tím vlastně popisuje jednotlivcovo ocenění utilitou pro každou alternativu.

Model se bude sestávat z parametrů funkce utility, což jsou napozorované nezávislé proměnné a neznámé parametry a jejich hodnoty jsou odhadnuty ze vzorku, který byl získán při známé a dostupné sadě možností od respondenta, který čelil specifické situaci.

Koncept utility nám dovoluje do rozhodování zařadit celou řadu alternativ a určit pouze jednu, která má největší užitek. V podstatě jediná hodnota, která nás zajímá, je rozdíl mezi utilitami páru alternativ, protože utilita nemá žádnou jednotku. Zejména nás zajímá, zda je tento rozdíl kladný či záporný.

2.3.2 Koncept deterministické volby

Pravidlo maximální utility neobsahuje nejistotu v jednotlivcově rozhodovacím procesu, což znamená, že jedinec si je jistý, že vybírá alternativu pod určitými podmínkami s nejvyšší utilitou, která byla do výběru zařazena. Modely utility, které dávají určitou předpověď volby jednotlivce, jsou nazývány deterministické modely utility. Pokud modely deterministické volby popisují chování správně, můžeme očekávat, že ti samí jednotlivci budou za stejných podmínek a stejných alternativách volit to samé rozhodnutí v různých časech pořád stejně. V praxi však pozorujeme i změny v jednotlivcově volbě. Pozorujeme, že jednotlivci si vybírají různé volby, i když působí zdánlivě stejně a čelí podobným, ne-li stejným alternativám. Tyto nevysvětlitelné volby zvyšují otázku vhodnosti modelů deterministické volby pro modelování dopravy nebo jiného lidského chování. Je to však způsobeno neschopností pozorovatele/analytika popsat rozhodovací situaci. Budoucí vizí je vyvinout model strukturovaný tak, že poskytne obstojnou reprezentaci těchto nevysvětlitelných odchylek v cestovním chování.

Víme, že se v aplikaci deterministických funkcí utility objevují tři zdroje chyb. První chybou je, že jednotlivci, kteří provádějí volbu, nemají kompletní informace nebo mají chybné informace, a nebo správně nepochopili informace o attributech alternativ. Výsledkem je, že se různí jedinci s jinými informacemi o alternativách budou pravděpodobně rozhodovat pro jiné volby. Druhou chybou je, že analytik nebo pozorovatel má jiné nebo nekompletní informace o vlastnostech alternativ ve vztahu k jednotlivci. Třetí chybou je, že analytik pravděpodobně neví, nebo nepočítá se speciálními okolnostmi spojenými s jedincovým cestovním rozhodnutím. V modelu tedy musíme počítat s tím, že na straně analytika je určitá nevědomost. Tyto modely jsou nazývány modely náhodné utility nebo model pravděpodobnostní volby.

2.3.3 Model pravděpodobnostní volby

Jestliže analytik plně rozumí všem aspektům interního rozhodovacího procesu jednotlivce, a také jeho vnímání alternativ, bude schopen precizně popsat tento proces a předvídat výběr způsobu dopravy použitím deterministického modelu utility. Zkušenosti ukázaly, že pokud analytik nemá takové znalosti, tak by modely měly mít takovou formu, která připouští analytikův nedostatek informací. Teorie pravděpodobnosti nám spíše pomůže odhalit výběr alternativy, než předpovídat, jestli jednotlivec vybere tu kterou alternativu s jistotou. Tyto pravděpodobnosti mohou efektivně odrážet pravděpodobnosti populace, že lidé s danou sadou charakteristik a čelící stejné sadě alternativ vyberou určitou alternativu. Stejně jako v deterministické teorii volby se předpokládá, že si jednotlivec vybere určitou alternativu pouze tehdy, je-li její utilita větší než utility ostatních alternativ. Pravděpodobnost předpovídat z výsledků analytiků lze z rozdílu mezi odhadovanou hodnotou utility a hodnotou utility použitou cestovatelem. My reprezentujeme tento rozdíl rozkladem utility alternativy na dvě složky. Jedna složka reprezentuje podíl dodaný pozorováním analytika, často nazývaná deterministická složka funkce utility. Druhá složka pak reprezentuje rozdíl mezi neznámou utilitou použitou jednotlivcem a utilitou odhadnutou analytikem. Takže píšeme:

$$U_{it} = V_{it} + \varepsilon_{it}$$

kde U_{it} je pravá utilita alternativy i rozhodovatele t

V_{it} je deterministická (pozorovatelná) složka

ε_{it} je náhodná chyba, šum

Analytik nemá žádnou informaci o chybové složce. Celková chyba, které je součtem několika dílčích chyb je reprezentována náhodnou proměnnou. Vyskytují se různé předpoklady na rozdělení náhodné proměnné, tím se budeme zabývat dále.

2.3.4 Komponenty deterministické části funkce utility

Deterministická část utility, která je různá pro každou alternativu je matematická funkce atributů alternativ a charakteristik rozhodovatele. Užítky jednotlivých alternativ nejsou analytikovi známy. Může mít různé matematické formy, ale funkce je obecně formulována jako aditivní. To proto, abychom si zjednodušili odhadovací proces. Funkce zahrnuje neznámé parametry, které jsou odhadovány v procesu modelování. V deterministické funkci utility můžeme pozorovat další tři složky. Složka spojená výhradně s atributy alternativ, složka spojená výhradně s charakteristikami rozhodovatele a třetí složka reprezentuje

interakce mezi atributy alternativ a charakteristikami rozhodovatele. Jednotlivé složky bych ráda trochu rozvedla.

Utilita sdružená s atributy alternativ. Taková utilita zahrnuje proměnné, které popisují atributy alternativ. Tyto atributy mají vliv na utilitu každé alternativy pro každého člověka ve zkoumané populaci. Atributy alternativ jsou měřitelné a mezi atributy řadíme například cestovní čas, cestovní náklady, vzdálenost od zastávky, spolehlivost dojetí na čas. Tyto hodnoty se mohou lišit napříč alternativami pro stejné jedince, a také mohou nastat rozdíly mezi jednotlivci a to díky různým počátečním a konečným destinacím.

Než se dostaneme k utilitě sdružené s charakteristikami rozhodovatele, měli bychom se také zmínit o problému, kdy rozhodovatel zvolil preference alternativ a pozorovatel tuto volbu nemohl nijak vysvětlit na základě pozorování. Došlo se k názoru, že se jedná o určitou náklonnost. Rozdíly v náklonnosti u jednotlivců můžeme vysvětlit pomocí začlenění osob a proměnných týkající se domácností. Právě tím se dostáváme k utilitě spojené s charakteristikami rozhodovatele. K tomuto účelu nejčastěji používáme proměnné jako například příjem domácnosti cestujícího, věk, pohlaví cestujícího, počet automobilů v domácnosti, počet pracujících lidí v domácnosti. Tyto proměnné můžeme zadávat různými způsoby. Můžeme například vytvářet model a brát v úvahu počet automobilů v domácnosti nebo můžeme pracovat s tím, zda jednotlivec má k dispozici auto či ne. Je však důležité si uvědomit, že musíme rozpoznat, že parametr, který popisuje vliv cestovatelovy charakteristiky se může lišit v každé alternativě, i když proměnné jsou stejné napříč všemi alternativami.

Jako poslední si představíme utilitu, která je složená jak z atributů alternativ, tak z rozhodovatelových charakteristik. Náš model se snažíme podchytit právě touto cestou. Například vliv příjmu, popsany dříve, zvyšuje preferenci pro cestování osobním automobilem. Tato reprezentace znamená, že preference osobního automobilu roste se zvyšováním příjmu. Další cesta jak reprezentovat vliv příjmu je, že zvyšování příjmu redukuje důležitost nákladů v hodnocení alternativních dopravních módů. Myšlenku, že vyšší příjem cestovatelů stanoví menší důležitost na náklady, můžeme reprezentovat rozdělením nákladů na cestování podle alternativ z ročního příjmů nebo nějakou funkcí ročního příjmu cestovatele nebo domácnosti. Jiný interakční efekt může být, že různí cestovatelé vnímají cestovní čas jinak. Například ženy, které mají zodpovědnost za úklid a péči o děti, hodnotí čas strávený v dopravním prostředku více negativně než muži.

2.3.5 Specifikace chybové složky

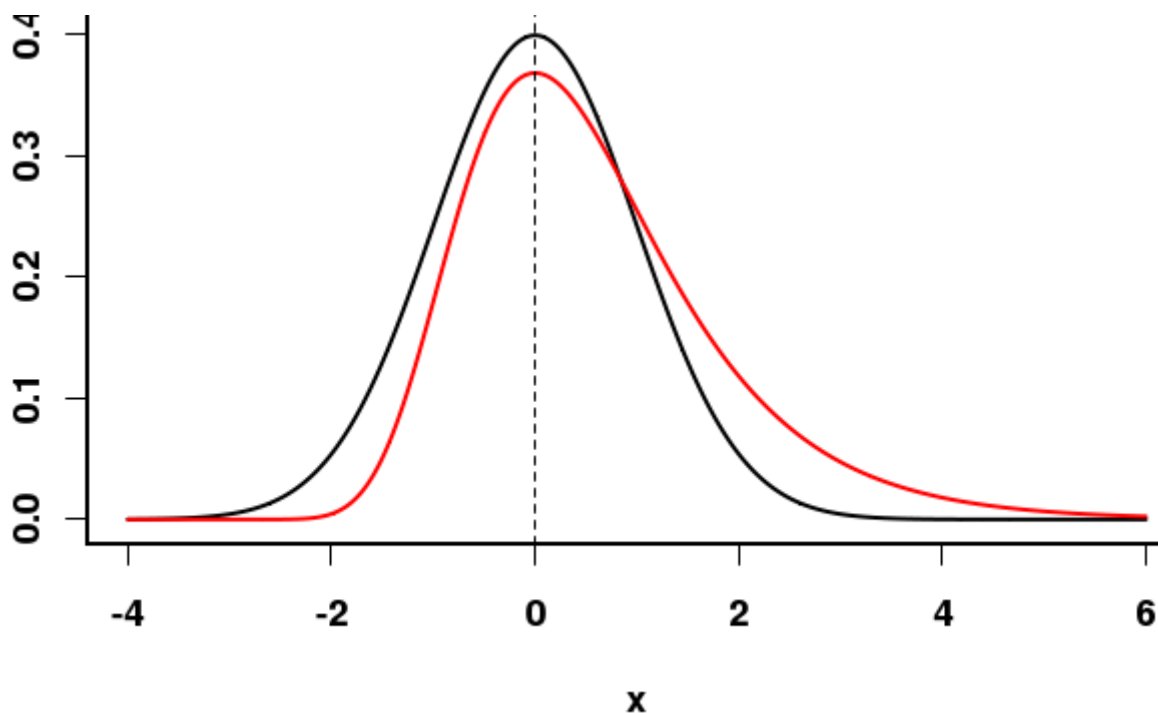
Jak bylo popsáno dříve, každá alternativa je reprezentována deterministickou složkou, která reprezentuje funkci utility pozorovatelných a měřitelných proměnných a složkou náhodnou, která reprezentuje ty složky funkce utility, které nejsou zahrnuty v modelu. Jak už bylo řečeno, chybová složka nám vlastně zahrnuje to, co analytik nebyl schopen správně a plně zachytit. Můžeme zde použít širokou škálu rozdělení, kterými můžeme reprezentovat chybovou složku. Jestliže předpokládáme, že chybová složka pro každou alternativu reprezentuje mnoho chybějících hodnot a každá z nich má relativně malý dopad na každou alternativu, pak centrální limitní věta říká, že suma těchto malých chyb bude normálně rozdělená. Tento předpoklad vede k formulaci Multinomického probitového modelu - pravděpodobnostního modelu volby. Nicméně tento model se v praxi hojně nepoužívá, kvůli jeho složitosti a následnému těžkému odhadování, interpretování a predikci. Existuje zde obdoba – Multinomický logitový model.

2.4 Multinomický logitový model

Matematická forma diskretní volby je založena na předpokladech týkajících se chybové složky funkce utility každé alternativy. Specifické předpoklady vedoucí k Multinomickému logitovému modelu jsou:

- chybová složka má Gumbelovo rozdělení
- chybové složky jsou na sobě nezávislé a jsou stejně rozdělené napříč alternativami
- chybové složky jsou na sobě nezávislé a jsou stejně rozdělené napříč všem pozorováním

Jak už jsme lehce zmínili výše, nejčastěji má chybová složka normální rozdělení. To nás ovšem potom směřuje k MNP, který má několik vlastností, které jsou těžko použitelné v analýze volby. Proto je zde vybráno rozdělení Gumbelovo, které má výpočetní výhody a v kontextu, kde je důležitá maximalizace, blízko aproximuje normální rozdělení a vytváří uzavřenou formu pravděpodobnostního modelu volby.



Zde vidíme ukázkou Gumbelova rozdělení. Myslím si, že je vhodné jej vidět v porovnání s normálním rozdělením v jednom grafu, kde můžeme vidět, že jsou tato rozdělení opravdu velmi podobná. Černá křivka nám znázorňuje normální rozdělení a křivka červená pak rozdělení Gumbelovo.

Gumbelovo rozdělení podle [9] má následující kumulativní distribuční funkci a funkci hustoty:

$$F(\varepsilon) = \exp\{-\exp[-\mu(\varepsilon - \eta)]\}$$

$$f(\varepsilon) = \mu\{\exp[-\mu(\varepsilon - \eta)]\}\exp\{-\exp[-\mu(\varepsilon - \eta)]\}$$

,kde μ je pozitivní škálový parametr, který určuje odchylku rozdělení a η je lokační parametr

Charakteristiky rozdělení:

$$\text{Střední hodnota} = \eta + \frac{\gamma}{\mu}$$

$$\text{Rozptyl} = \frac{\pi^2}{6\mu^2}$$

,kde γ je Euler – Mascheroniho konstanta, $\gamma \approx 0,5772$

Podmínkou je, že musíme uvést umístění a odchylku rozdělení, stejně jako to děláme u normálního rozdělení.

Jestliže máme Gumbelovo rozdělení nadefinováno alespoň takto, pak nás to vede k matematické struktuře, známé jako Multinomický logitový model, který je schopen vypočítat pravděpodobnosti výběru jednotlivých alternativ následovně:

$$P(i) = \frac{\exp(V_i)}{\sum_{j=1}^J \exp(V_j)}$$

kde $P(i)$ je pravděpodobnost, že si rozhodovatel vybere alternativu i , $i = 1, 2, \dots, J$

V_j je deterministická část funkce utility alternativy j

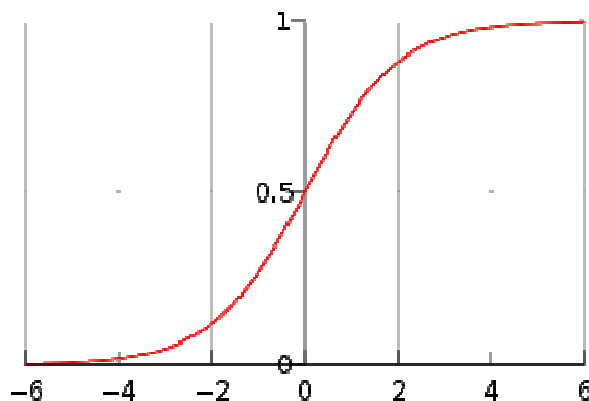
Pro pravděpodobnost $P(i)$ musí platit:

$$0 \leq P(i) \leq 1$$

a

$$\sum_j P(i) = 1$$

MNL oplývá řadou důležitých výhod. První velice důležitou výhodou je, že MNL má velmi jednoduchý výpočet. Ke zpracování není třeba žádného speciálního softwaru, postačí nám obyčejná kalkulačka – proto bývá hojně využíván. Druhou důležitou výhodou je, že pravděpodobnosti vypočtené Multinomickým logitovým modelem tvoří jakousi S křivku, kdy pravděpodobnosti výběru možnosti mohou nabývat pouze hodnot nula až jedna. Nula pro nás samozřejmě znamená, že pravděpodobnost výběru tohoto dopravního prostředku je velmi mizivá a naproti tomu jednička ukazuje, že pravděpodobnost, že si jednotlivec vybere tuto alternativu, je vysoká ve vztahu k alternativním možnostem. U extrémních hodnot pravděpodobností (0,1) má křivka velmi pozvolný sklon. Čím blíže se však blíží ke středu, tím je více strmější. Což pro nás znamená jediné - pokud se bude pravděpodobnost pohybovat někde kolem středu, může mít pro nás i malá změna dramatické následky, pokud se však budeme pohybovat v okolí extrémních hodnot, tak jsou pravděpodobnosti výběru alternativy docela stabilní. Pro ukázkou je níže zobrazena S křivka.



Další výhodou je, že pravděpodobnost volby určité možnosti závisí na rozdílech v systematické složce utility. Což využíváme. Pravděpodobnost volby se nezmění, jestliže přidáme do funkce utility každé alternativy nějakou (stejnou) neznámou hodnotu.

Uvažujme, že máme k dispozici tyto funkce utilit:

$$U_A = \beta_{A,0} + \beta_{A,1} * x_{11} + \beta_{A,2} * x_{12}$$

$$U_{SR} = \beta_{SR,0} + \beta_{SR,1} * x_{21} + \beta_{SR,2} * x_{22}$$

$$U_T = \beta_{T,0} + \beta_{T,1} * x_{31} + \beta_{T,2} * x_{32}$$

Zde jsme si zavedli tři možnosti, ze kterých si může rozhodovatel vybrat. První z nich je možnost jízdu autem jako řidič, druhá možnost je jízda autem jako spolujezdec a třetí je jízda veřejnou dopravou. Nadefinovali jsme si do funkcí utilit konstanty $\beta_{A,0}$, $\beta_{SR,0}$, $\beta_{T,0}$. Takto ale nelze všechny parametry odhadnout. To samé se týká i parametrů $\beta_{,1}$. Protože jestliže přidáme nějakou hodnotu ke všem možnostem, pak to nezpůsobí žádnou změnu v pravděpodobnostech volby dopravního prostředku. Řešení tohoto jevu je, že umístíme do modelu určité omezení na každou sadu parametrů. Nejčastěji nastavujeme parametry pro jednu alternativu, kterou nazveme referenční, aby byly rovny nule. Následně pak interpretujeme ostatní parametry ve vztahu k referenční alternativě. Volba referenční alternativy je libovolná. Takže například:

$$U_A = 0 + \beta_{A,2} * x_{12}$$

$$U_{SR} = \beta_{SR,0} + \beta_{SR,1} * x_{21} + \beta_{SR,2} * x_{22}$$

$$U_T = \beta_{T,0} + \beta_{T,1} * x_{31} + \beta_{T,2} * x_{32}$$

Musíme si však následně dát pozor na interpretaci. Referenční alternativu jsme zvolili první, tudíž volba automobilu jako řidič. Takže až budeme interpretovat výsledky, budeme hovořit o parametrech, které jsou ve vztahu k volbě automobilu jako řidiči.

2.4.1 Nezávislost nepodstatných (irrelevant) vlastností alternativ

Jedním z nejvíce diskutovaných témat u MNL je nezávislost nepodstatných alternativ (IIA). Znamená to, že pro každého jednotlivce poměr pravděpodobností vybraných 2 alternativ je nezávislý na přidání nebo odebrání atributů u ostatních alternativ v sadě možností. Předpoklad je, že ostatní alternativy jsou nepodstatné při výběru mezi dvěma alternativami v páru. Pro ukázkou uvažujte model MNL a vybíráme si mezi třemi dopravními módy – automobil, vlak a autobus. Pravděpodobnosti výběru jednotlivých dopravních módů známe a pojďme se podívat, jak budou vypadat poměry pravděpodobností v konkrétním případě:

$$\frac{P(Auto)}{P(Bus)} = \frac{\exp(V_{Auto})}{\exp(V_{Bus})} = \exp(V_{Auto} - V_{Bus})$$

$$\frac{P(Auto)}{P(Rail)} = \frac{\exp(V_{Auto})}{\exp(V_{Rail})} = \exp(V_{Auto} - V_{Rail})$$

$$\frac{P(Bus)}{P(Rail)} = \frac{\exp(V_{Bus})}{\exp(V_{Rail})} = \exp(V_{Bus} - V_{Rail})$$

Poměry pravděpodobností pro každý pár alternativ závisí pouze na attributech těch alternativ, které se objevují ve vztahu a ne na attributech nějaké třetí alternativy.

IIA vlastnost má několik důležitých důsledků. Dovoluje nám přidat nebo pozměnit alternativy ze sady alternativních možností, aniž by to mělo dopad na strukturu nebo parametry modelu, což má určité výhody. Za prvé, model může být odhadován a aplikován v případech, kde různé počty lidí čelí různým sadám dopravních módů. Za druhé, tato vlastnost zjednodušuje odhady parametrů v MNL modelu. A za třetí, tato vlastnost je výhodná, když aplikujeme model v případě predikce a potřebujeme znát pravděpodobnost výběru zcela nové alternativy.

Na straně druhé, vlastnost IIA nemusí správně odrážet vztahy v chování mezi skupinami alternativ, což znamená, že ostatní alternativy nemusí být nepodstatné v poměru pravděpodobností mezi párem alternativ. A to může vést k chybným predikcím v pravděpodobnostech výběru. Extrémní případ – red bus/blue bus paradox.

Uvažujeme případ, kdy má jednotlivec možnost dojet do práce autem nebo modrým autobusem. Předpokládáme, že atributy auta a modrého autobusu jsou takové, že pravděpodobnost výběru auta bude 2/3 a pro modrý autobus bude pravděpodobnost výběru 1/3. Tudíž poměr pravděpodobností bude 2:1. Nyní předpokládáme, že na trh přijde nový poskytovatel autobusové dopravy, který do dopravy zavede červený autobus. Mezi modrým autobusem a červeným autobusem není vůbec žádný rozdíl. Kromě barvy. Pojede po stejné trase, stejným typem vozidla, bude jezdit podle stejného jízdního řádu a bude zastavovat na stejných zastávkách jako modrý autobus. Logicky se dá uvažovat, že počet lidí, kteří si vyberou auto bude úplně stejný jako předtím a lidé, kteří budou volit veřejnou dopravu se budou rozhodovat mezi červeným a modrým autobusem. Takže zavedení červeného autobusu nemá na jednotlivce, kteří volí osobní automobil žádný nebo skoro žádný vliv. Proto očekáváme, že pravděpodobnosti výběru dopravního módu budou pro auto 2/3, pro modrý autobus 1/6 a pro červený autobus 1/6. Nicméně, díky IIA vlastnosti bude MNL model udržovat relativní pravděpodobnost auta a modrého autobusu 2:1. Jestliže předpokládáme, že lidé jsou lhostejní k barvě prostředku, kterým se dopravují, pak budou mít námi uvažované

autobusy stejnou funkci utility a tedy jejich relativní pravděpodobnosti budou 1:1. Pravděpodobnosti se potom rozdělí následovně:

$$P(\text{Auto})=1/2$$

$$P(\text{Blue bus})=1/4$$

$$P(\text{Red bus})=1/4$$

To znamená, že pravděpodobnost podílu lidí, kteří se rozhodnou pro auto klesne ze 2/3 na 1/2 v důsledku zavedení alternativy, která je identická s již existující alternativou. Tento případ je extrémní, ale paradox nám poskytuje důležitou ukázkou možných důsledků IIA vlastností.

2.4.2 Odhad modelu

Vývoj logitového modelu se skládá z formulace modelu a odhadu číselných hodnot parametrů pro různé atributy, které jsou specifikované ve funkcích utility. Jak jsme již uvedli výše, analytik není schopen postihnout všechny proměnné, které mají vliv na volbu cestovatele. Proto je velice nepravděpodobné, že výsledek bude naprosto přesný. V těchto případech je nezbytné použít metodu odhadu, která získává různé výsledky odhadů, které se týkají toho, jak dobře určit vybranou alternativu. To dobře využívá metoda maximální věrohodnosti, která se skládá z nalezení parametrů modelu, které maximalizují pravděpodobnost pozorovaných možností podmíněného modelu. Což znamená maximalizaci pravděpodobnosti, že vzorek byl generován z modelu s vybranými hodnotami parametrů.

2.4.3 Metoda maximální věrohodnosti

Postup pro získání odhadů metodou maximální věrohodnosti zahrnuje dva kroky:

- 1) Vývoj sdružené pravděpodobnostní funkce hustoty zkoumaného vzorku, která se nazývá pravděpodobnostní funkce
- 2) Odhad hodnot parametrů, které maximalizují pravděpodobnostní funkci

Pravděpodobnostní funkce pro vzorek T jednotlivců a J možností k výběru je následující:

$$L(\beta) = \prod_{t \in T} \prod_{j \in J} (P_{jt}(\beta))^{\delta_{jt}}$$

kde $\delta_{jt} = 1$ je indikátor (=1, jestliže j je vybráno jednotlivcem t , jinak 0)

P_{jt} je pravděpodobnost, že jednotlivec t vybere možnost j

Následně jsou odhadovány hodnoty parametrů, které maximalizují pravděpodobnostní funkci. Ty jsou získány nalezením první derivace pravděpodobnostní funkce a položením nule. Jelikož logaritmus pravděpodobnostní funkce má stejné maximum, je pro nás vhodnější

uvažovat tuto funkci. Do pravděpodobnostní logaritmické funkce následně dosadíme derivovanou funkci P_{jt} a tuto rovnici položíme rovnu nule.

$$P_{jt} = \frac{\exp(X'_{jt}\beta)}{\sum_j \exp(X'_{jt}\beta)}$$

Z rovnice, kdy jsme dosadili derivovanou funkci P_{jt} do první derivace pravděpodobnostní funkce jsme pak schopni získat nejlepší hodnoty, tedy vektor parametrů $\hat{\beta}$. Můžeme si být jisti, že toto je nejlepší řešení v případě, že si ověříme, jestli druhá derivace funkce bude záporně definitní. Pro odhady získané touto metodou potom platí, že mají následující vlastnosti:

- Jsou konzistentní
- Asymptoticky mají normální rozdělení
- Jsou asymptoticky eficientní

2.5 Požadavky na data

Prvním krokem v sestavování modelu volby je shromáždit data o cestovatelově volbě a proměnných, o kterých si myslíme, že budou mít na volbu vliv. Takže by se zde měly určitě objevit data týkající se cestovatele a proměnné spojené s cestou, které ovlivňují úsudek o jednotlivých módech (např. příjem, počet pracujících osob v domácnosti, věk, pohlaví, vlastnictví automobilu, počet automobilů v domácnosti, účel cesty, hustota obyvatelstva...). Další data, která bychom do modelu měli zahrnout, jsou údaje o dostupných možnostech dopravy. Tato data by měla popsat jednotlivé možnosti (např. cestovní čas, čas strávený v dopravním prostředku, čas strávený čekáním, vzdálenost od zastávky, cestovní náklady). A poslední skupinou jsou data, která nám říkají například: jaké jsou cestovní náklady rozdělené podle příjmu domácností, cestovní čas nebo náklady spojené s pohlavím nebo věkovou skupinou cestovatele,...

Pro to, abychom mohli provádět odhady modelů volby, je důležité si data sesbírat. Data obvykle pocházejí z nějakého průzkumu, kdy se dotazujeme určitého vzorku obyvatelstva, který nás zajímá. V praxi se nejčastěji vyskytují tři typy průzkumů. Průzkumy domácností, kdy kontaktujeme respondenta u něj doma, pak průzkumy na pracovišti, kdy se zaměřujeme na cesty, které jsou spojené s prací, a jako poslední průzkumy, kdy se snažíme respondenta zachytit přímo při cestování, ty se nejčastěji užívají při studiích meziměstských cest.

Při sběru dat bychom si měli nejdříve promyslet několik věcí. První ze všeho bych navrhovala si ujasnit, která skupina lidí bude v našem hledáčku, což bude záviset na účelu studie. Poté bychom si měli určit skupinu, kterou bychom rádi do šetření zahrnuli. V našem případě bylo omezením pouze to, zda má respondent alespoň 18 let. To proto, aby měl tu možnost mít k dispozici všechny dopravní módy, včetně auta. Respondenti by měli být jedineční a vzorek respondentů by měl být skutečnou reprezentací celé společnosti. A nakonec bychom se měli zamyslet nad tím, jak velký vzorek populace budeme potřebovat. Samozřejmě čím větší počet dat, tím lépe jsme schopni podchytit realitu. Což se však může bít s cenou průzkumu. Rozhodnutí ohledně velikosti vzorku vyžaduje opatrné zhodnocení, spočívající v tom, abychom měli adekvátní data pro uspokojení cílů studie versus rozpočtové omezení.

2.5.1 Stated preference a revealed preference data

Při tvorbě modelu pravidelně vyvstávaly nějaké problémy. Že máme data zadána revealed - preference technikou namísto stated - preference, bylo zjištěno na základě prvních výsledků modelů, kdy výsledky vůbec neodpovídaly našemu očekávání. Proto bych zde ráda představila tyto dvě techniky získávání dat a také to, jak s jednotlivými případy zacházet. K tomu, abychom vytvořili dobrý model poptávky po dopravě je potřeba postihnout chování systému a celou širokou škálu podmínek, které na něj působí. Dříve se striktně předpokládalo, že modelování dopravy je založeno na datech, která poskytlí respondenti o alternativách, jež měli k dispozici a pro které se opravdu rozhodli. Tomuto získávání dat říkáme revealed – preference data. Získání těchto dat bylo však finančně velice náročné a navíc některé základní informace jsme z modelu prostě nedokázali vyčíst. Dále zde byl problém s nabídkou zcela nové alternativy, kdy bylo potřeba předpovědět zájem o nový dopravní prostředek, který se měl razantně lišit od ostatních, již nabízených.

Na konci 70. let 20 století se objevila stated – preference technika (dříve se již objevila také v oblasti průzkumu trhu). Nabízela zcela novou cestu experimentování v oblasti týkající se dopravy. Řešila problémy, které byly zmiňovány výše. Přitom její princip je velice jednoduchý. Je založen na odhadech, které vznikly z analýzy odpovědí respondentů na hypotetické otázky. Ty se mohou dotazovat na širší škálu atributů a podmínek než systém reálný. Nicméně, tyto techniky bývaly ve velké míře kritizovány, protože i když respondent odpoví, že by tuto alternativu za daných podmínek zvolil, tak není jisté, zda by se tak v reálné situaci opravdu zachoval. Trvalo celé desetiletí, než se situace změnila. Byla jí dána šance na to, aby vyřešila problémy uvedené výše. Což se povedlo a v dnešní době slouží jako výborný

nástroj pro pomoc s modelováním nových alternativ. Klasický případ získávání dat pomocí techniky stated – preference je Swissmetro case. Jedná se o případ, kdy mělo být zavedeno Swissmetro – nový druh dopravy - mezi městy St. Gallen, Ženeva (Švýcarsko).

Pro shrnutí zde nabízím krátký přehled toho, co stated - preference zvládá oproti revealed – preference data:

- zvládá odhadnout poptávku po zcela novém dopravním způsobu nebo nové atributy stávajících způsobů dopravy
- zkoumá, jak se volba dopravního prostředku mění v závislosti na věku, pohlaví, příjmové skupiny
- získávání RP dat může být mnohonásobně dražší, protože při získávání RP dat pozorujeme respondenty v terénu, kdežto u SP dat provádíme anketu, respondent nemusí podnikat žádnou cestu.

Ze shrnutí bychom mohli mít pocit, že RP data jsou pasé, ovšem SP data nejsou také ideální. Popisují pouze hypotetické rozhodnutí v určitém kontextu. Může se také stát, že respondent například nemusí pochopit dotazníkové otázky.

Ze souhrnu výhod plyne, že SP data jsou mnohem výhodnější. Nelze to však tvrdit obecně. Rozhodně bych ale nepodceňovala hodnotu RP dat, protože ta nám pomáhají zachytit svět takový, jaký je, uvažujeme u nich pouze existující možnosti v pozorování, a hlavně mají vysokou hodnotu spolehlivosti.

Záleží jen na nás, jaký typ dat se nám hodí. Měli bychom však mít na paměti, že s každým typem dat zacházíme jinak. U stated – preference techniky máme data v podstatě připravená k analýze, kdežto u revealed – preference techniky si musíme data upravit – nejčastěji zadat atributy možností, které si respondent nevybral k cestování, pokud chceme znát odhady hodnot atributů alternativ, které pro cestu nebyly vybrány.

2.5.2 Metody pro sběr dat alternativních způsobů dopravy

Průzkumy kompletují informace popisující jednotlivce, který vykonal nějakou cestu, a podává o ní informace. Tato data však pro naše účely nejsou kompletní. My potřebujeme získat ještě informace o alternativních módech, které se dají získat z jiných zdrojů. Jsou obvykle generovány z datových souborů, které jsou simulací vlastností (trasu linek, jízdné, cestovní časy v dopravních prostředcích, čekací doby, vzdálenosti mezi začátkem cesty a cílem cesty,...) alternativních módů. Samozřejmě je důležité dodržet určitou strukturu dat. My volíme následující strukturu dat:

CHOICE	C_time	B_time	P_time	W_time	C_nakl	P_nakl
1	10.52818	16.11552	28.21327	89.18369	13.98206	25
1	10.52818	16.11552	28.21327	89.18369	13.98206	25
3	17.80738	26.01954	41.88668	171.8177	28.56573	0
3	17.80738	26.01954	41.88668	171.8177	28.56573	0
1	9.435581	17.14331	37.78629	120.0112	18.54195	25
1	9.435581	17.14331	37.78629	120.0112	18.54195	25
4	17.73565	26.67444	47.52134	159.1168	24.36105	25
1	20.71938	71.74152	53.73421	466.8979	63.27752	90
1	17.73565	26.67444	47.52134	159.1168	24.36105	25
3	13.47817	20.69602	34.299	121.6324	18.84734	0

Máme tady příklad zadávaných dat. Co tato data ale znamenají? V prvním sloupci se vyskytují čísla 1 – 4. To značí, jaký dopravní mód si jednotlivec vybral pro vykonání své cesty. Další čtyři sloupce potom uvádějí cestovní čas, postupně pro auto, kolo, veřejnou dopravu a chůzi. Poslední dva sloupce jsou potom sloupce obsahující cestovní náklady. Jsou specifické pouze pro varianty auto a veřejná doprava, protože u chůze ani u kola nic neplatíme.

Pro naše účely bylo nutné do dat doplnit informace o attributech cestovního času, cestovní vzdálenosti, frekvenci veřejné dopravy a počtu přestupů. Tato data doplňovalo CDV. Území Jihomoravského kraje mají rozděleno do zón a mají k dispozici data o všech cestovních časech a cestovních vzdálenostech mezi všemi zónami a pro všechny dopravní prostředky. Pomocí tohoto souboru potom není problém zjistit, jaký bude cestovní čas a cestovní vzdálenost mezi zónami pro dopravní módy auto, kolo, veřejná doprava a chůze. Tato struktura dat je platná pro software Biogeme, kterým budu data v diplomové práci zpracovávat. Softwaru Biogeme je věnována třetí kapitola.

2.6 Posouzení kvality modelu

V následující kapitole zrekapitulujeme způsoby zhodnocení výsledků modelu. Můžeme jej hodnotit buď na základě neformálních testů. Například znaménko u parametrů – ukazuje, jestli má parametr na utilitu pozitivní nebo negativní vliv. Na základě vlastního úsudku bychom měli být schopni rozlišit, zda je tento výsledek možný či ne. Nebo například rozdíly v parametrech, které jsou specifické pro každou alternativu. Obvykle máme určitou představu o tom, v jakých hodnotách by se měly parametry pohybovat.

Představíme si zde také míru, která popisuje celkovou kvalitu modelu, nazývá se koeficient determinace. Tato míra ukazuje vztah mezi hodnotami:

- $LL(0)$ - log-likelihood, kdy jsou všechny parametry rovny nule
- $LL(c)$ - log-likelihood pouze s konstantami specifickými pro každou alternativu
- $LL(\hat{\beta})$ - log-likelihood pro odhadnutý model metodou maximální věrohodnosti
- $LL(*)=0$ je log-likelihood pro perfektní predikční model

,pak můžeme psát:

$$\rho_0^2 = \frac{LL(\hat{\beta}) - LL(0)}{LL(*) - LL(0)}$$

A jestliže hodnota log-likelihood pro $LL(*)$ je nula, můžeme psát:

$$\rho_0^2 = 1 - \frac{LL(\hat{\beta})}{LL(0)}$$

Potom hodnota koeficientu determinace leží mezi 0 a 1. Hodnota 0 nám značí, že model nebyl odhadnut dobře a jednička naopak, že model bezchybně vysvětluje variabilitu v závislé proměnné a každá volba může být předpovídána. Tato míra bývá hojně využívána. Problém nastává v tom, že my nevíme, která míra je pro nás dobrá natolik, abychom mohli analýzu ukončit. Proto záleží na vlastním úsudku, co budeme považovat za dobré a co ne.

Další měřitelnou mírou je upravený koeficient determinace. Ve svém výpočtu zahrnuje i počet odhadovaných parametrů zahrnutých do modelu:

$$\bar{\rho}_0^2 = \frac{[LL(\hat{\beta}) - K] - LL(0)}{LL(*) - LL(0)} = 1 - \frac{LL(\hat{\beta}) - K}{LL(0)}$$

kde K je počet parametrů použitých v modelu.

Dále se tu objevují statistické testy, které se používají k hodnocení formálních hypotéz. Když testujeme jednotlivé parametry, tak vlastně prozkoumáváme dopad chyb ze vzorkování na jednotlivé parametry. Standardní chyba hraje velice důležitou roli v testování, jestli jednotlivé parametry jsou rovny nějaké hypotetické hodnotě. Statistika používaná pro testování nulové hypotézy, že parametr $\hat{\beta}_k$ je roven hypotetické hodnotě β_k^* je asymptotická t-statistika:

$$t - \text{statistika} = \frac{\hat{\beta}_k - \beta_k^*}{S_k}$$

kde $\hat{\beta}_k$ je odhad pro k-tý parametr

β_k^* je hypotetická hodnota

S_k je standardní chyba odhadu

Pokud je hodnota t-statistiky dostatečně velká, vede to k zamítnutí nulové hypotézy, že parametr se rovná hypotetické hodnotě. Nejčastěji provádíme test, kdy je hypotetická hodnota rovna nule, potom při zamítání nulové hypotézy říkáme, že parametr má významný dopad na model. Pokud nulovou hypotézu nezamítáme, měli bychom zvážit, zda tento parametr do modelu budeme nadále zahrnovat. Záleží pouze na našem úsudku. Pokud si myslíme, že by v modelu měl vystupovat některý parametr, který byl označen za nevýznamný, je možné si ho do modelu zahrnout. Záleží totiž jen na nás, jakou hladinu významnosti zvolíme. Nejčastěji volíme hladinu významnosti 0,05, což znamená, že v 5 případech ze 100 dovolujeme modelu nepřesnosti. Pro t-statistiku to potom znamená: bude-li vypočtená hodnota v absolutní hodnotě větší než 1,96 (tabulková hodnota), lze zamítnout nulovou hypotézu a parametr bude významný.

Pokud nám to přijde zajímavé, můžeme statisticky testovat také dva parametry, zda se sobě rovnají. Například bychom takto mohli testovat, zda není čas v nějakém vztahu s náklady.

Dokonce můžeme testovat parametry celého modelu najednou. Použijeme k tomu test poměrem věrohodností, kdy se v podstatě porovnávají dva modely. Porovnáváme funkci pravděpodobnosti pro neomezenou a omezenou verzi modelu. Omezený model lze získat zavedením omezení (například některé parametry nastavit na nulu, pár alternativ nastavit rovny jiné, atd.) na odhadnutý model. Předpokládejme, že r označuje počet nezávislých omezení, které jsou vztažené na parametry při počítání LL_R . A víme, že platí:

$$LL_U \geq LL_R$$

Testovací statistika je potom:

$$-2 * [LL_R - LL_U]$$

Kde LL_R je hodnota log-likelihood funkce v maximu omezeného modelu

LL_U je hodnota log-likelihood funkce v maximu neomezeného modelu

Tato testová statistika má asymptotické Chí-kvadrát rozdělení s r stupni volnosti. A tak, jak to bylo u testu jednotlivých parametrů, i tady platí, že jestliže je výsledná hodnota dostatečně velká, pak zamítáme nulovou hypotézu. Problém je ovšem v tom, že nulovou hypotézu zamítáme skoro vždycky, proto test není moc nápomocný.

3 Data

Data byla zjišťována Centrem dopravního výzkumu v období od 21. 5. 2013 do 27. 6. 2013 v Jihomoravském kraji. Data byla shromažďována během 23 dní a dotazník byl poměrně rozsáhlý. Tím vznikl poměrně rozsáhlý soubor. Data jsou v této práci velice důležitou komponentou. Cílem další analytické práce je modelovat proces, který generuje variabilitu v závislé proměnné. Proto bych ráda data představila.

První ze všeho, co nás zajímalo, bylo zda data poměrově odpovídají složení mužů a žen v České republice. Informaci o poměrovém složení obyvatelstva podle pohlaví jsme získali od Českého statistického úřadu, který provádí nejrozsáhlejší statistiku – Sčítání lidu, domů a bytů. Sčítání bylo naposledy prováděno v roce 2011 a zjištění byly následující:

ČSÚ:

	Celkem	Muži	Ženy
	10 436 560	5 109 766	5 326 794
%	100 %	48,96 %	51,04 %

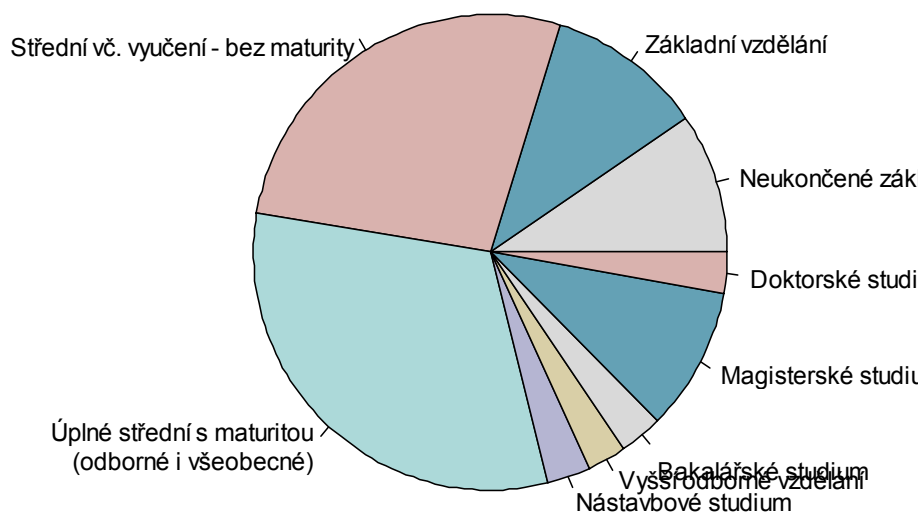
CDV:

	Celkem	Muži	Ženy
	2631	1258	1373
%	100 %	47,81 %	52,19 %

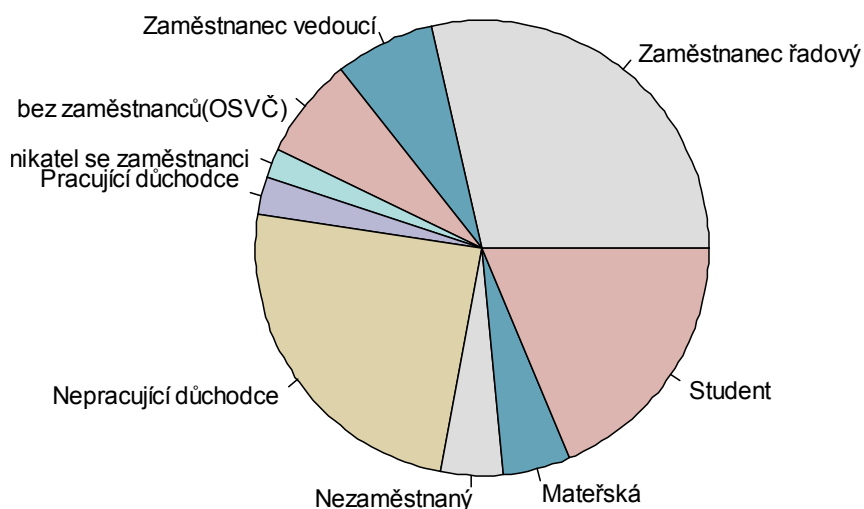
V horní části vidíme tabulku s údaji, které byly zjištěny Českým statistickým úřadem. V části dolní potom data zjištěná Centrem dopravního výzkumu. Již z prvního pohledu vidíme, že poměr mužů a žen je v obou případech v podstatě stejný. Kdybychom chtěli zacházet ještě více do detailu, můžeme na stránkách ČSÚ zjistit i poměrové složení podle pohlaví pouze pro Jihomoravský kraj. Tato data se pohybují ještě o kousek blíže našim datům. Muži jsou zde zastoupeni 48,8% a ženy potom 51,19%.

Nyní už něco k samotným datům – začneme od dat, která se týkala jednotek osob. Celkově bylo dotázáno 2631 osob. Z toho se do výzkumu zapojilo 1258 mužů a 1373 žen. Průměrný věk zjištěný v rámci našeho dotazníkového šetření je 44 let. Když jsem populaci rozdělila na muže a ženy, tak průměrné věky obou skupin byly překvapivě blízko sebe. U mužů je to lehce pod 44 let (43,933) a u žen lehce nad 44 let (44,068). Jediným omezením, co se věku týče bylo splnění věkové hranice 18 let. Je to proto, aby mohl respondent uvažovat všechny módy (měl k dispozici řidičský průkaz).

Při sledování dat získaných od respondentů ohledně vzdělání vidíme, že se výzkum opravdu snažil postihnout široké spektrum dotazovaných lidí, co se vzdělání týče. Z následujícího grafu vidíme, že se v průzkumu objevili lidé jak s neukončeným vzděláním, tak lidé, kteří mají vysokou školu. Pokud nebude uvedeno jinak, tak se grafy týkají všech 2631 respondentů.



Můžeme sledovat i poměr ekonomicky aktivních lidí:



Zde vidíme, že největší díl zde zaujímají řadoví zaměstnanci, dále nepracující důchodci a studenti. Také si můžeme všimnout, že poměr ekonomicky aktivních lidí vůči ekonomicky neaktivním je zde menší. Z 1253 ekonomicky aktivních lidí pracuje 80% ve standardní pracovní době, mezi kterou řadím 30 až 50 odpracovaných hodin týdně.

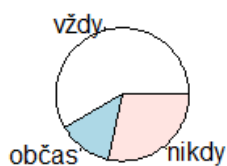
V dotazníkovém šetření se objevil docela vysoký počet lidí, kteří mají flexibilní pracovní dobu. V průzkumu se objevilo necelých 40% respondentů, kteří nemají fixní pracovní dobu. A dokonce celých 26% může svou práci vykonávat doma. Při zkoumání dat jsem narazila na skutečnost, že mužů, kteří mohou vykonávat svou práci doma je více než žen.

To je podloženo i další analýzou, kdy jsme si vypočetli, jaký je podíl mužů a žen ve vedoucích pozicích v našem průzkumu. Vychází nám 38% pro ženy a 62% pro muže.

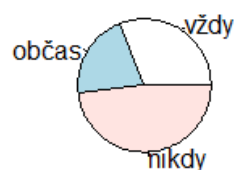
Data obsahují také informaci o tom, že 35% dotazovaných respondentů v dotazníkovém šetření nevlastní řidičský průkaz, což by odpovídalo průzkumu ministerstva dopravy, kdy v roce 2006 vlastnilo řidičský průkaz v Jihomoravském kraji pouze 63,81 % obyvatelstva. Zajímavé byly i odpovědi na otázku, jak často mají respondenti k dispozici auto. 44% odpovědělo, že vždy, 17% má k dispozici auto jen občas, 39% nemá k dispozici auto vůbec. Pokud bychom si tato data přepočítali na pohlaví, zjišťujeme, že celých 58% mužů

má auto k dispozici vždy, 13% pouze občas a 29% autem vůbec nedisponuje. Co se žen týče, ty jsou na tom mnohem hůře. Pouhých 31% žen disponuje autem vždy, což je skoro o polovinu méně než muži, 20% disponuje autem občas a 48% vůbec auto k dispozici nemá. Zde vidíme koláčové grafy týkající se dispozice auta. Koláčový graf pro muže obsahuje 1258 pozorování, a graf pro ženy obsahuje 1373 pozorování.

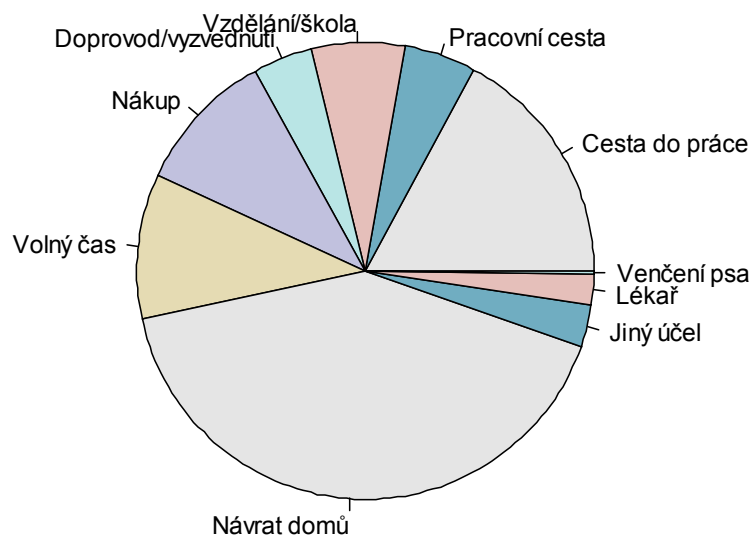
Muži



Ženy

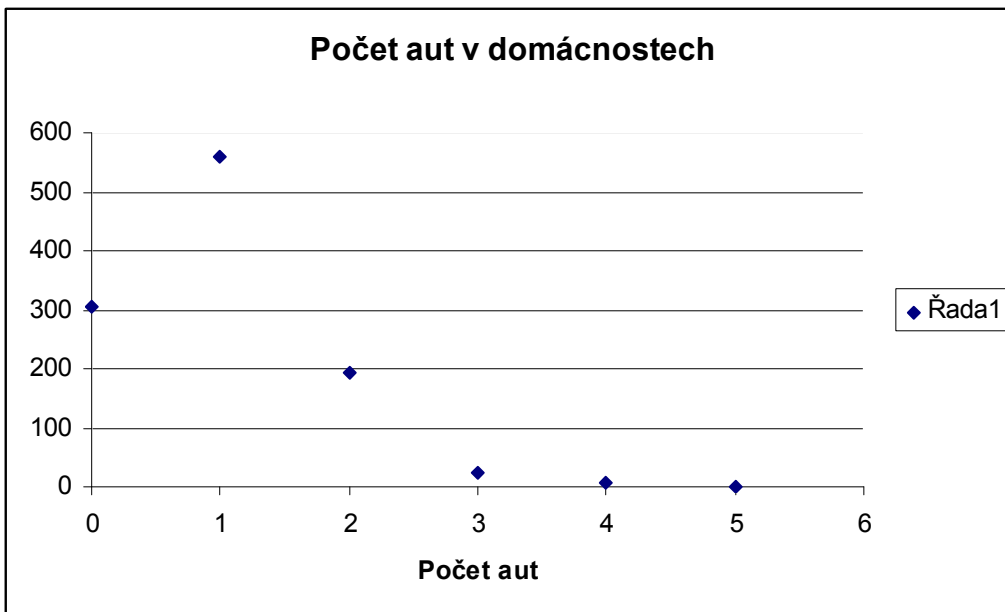


Pojďme si ukázat také něco o datech vázaných k jednotce cesty. Za dobu výzkumu bylo zaznamenáno 5772 cest. To znamená, že každý respondent v průměru provedl 2,19 cesty. Samozřejmě, že tu byli respondenti, kteří neuskutečnili ani jednu cestu a na straně druhé tu byli respondenti, kteří zvládli za jeden den uskutečnit až 7 cest. Nejčastěji se ale objevovaly výpovědi, že respondenti zvládli 1-2 cesty denně. Co se účelu, za kterým lidé cestovali, týče:



Můžeme zmínit i hodinu, ve které se uskutečnilo nejvíce cest. Byla to 7. hodina ranní. Ovšem když se podíváme na předchozí graf, vidíme, že největší část grafu zabírá cesta do práce a cesta domů, můžeme tedy vyvodit, že cesty se budou nejvíce uskutečňovat mezi 6 – 8 ranní a 14 – 16 odpolední, což také odpovídá našim datům.

A konečně data týkající se jednotky domácností. Z dotazníků jsme schopni vyčíst i nějaké informace o domácnostech. Celkově jsme měli k dispozici informace z 1092 domácností. Průměrný počet osob v domácnostech je 2,65 osoby. Průměrná vzdálenost na zastávku je 5,5 minuty. A poslední informaci o datech znázorním v grafu. Jedná se o počet aut, kterými disponuje domácnost. Celkově je zde 1052 aut a jsou rozděleny následovně:



4 Popis statistického softwaru Biogeme

Pro zpracování dat budeme používat software zvaný Biogeme. Bielaire Optimization toolbox for GEV Model Estimation (BIOGEME) je víceúčelový nástroj pro výpočet odhadů modelů diskrétní volby. Tato aplikace byla v první řadě vyvinuta pro posun ve výzkumu modelů diskrétní volby. Tato aplikace je volně stažitelná na stránkách

<http://biogeme.epfl.ch>,

kde jsem také čerpala literaturu.

První verze Biogeme se objevila v roce 2001. Aplikace spatřila světlo světa v roce 2003 a byla vytvořena Michaelem Bielairem. Biogeme si dokáže poradit s mnoha modely, zde je výčet:

- Logit
- Binary probit
- Nested logit
- Cross-nested logit
- Multivariate Extreme Value models
- Diskrétní a spojitě obměny Multivariate Extreme Value models
- Modely s nelineární funkcí utility
- Modely uzpůsobené pro panelová data
- Heteroskedastické modely

4.1 Jak Biogeme funguje

Biogeme potřebuje k fungování, odhadování následující dvě složky:

- specifikaci modelu (.mod)
- složku s daty (.dat)

Data do souboru zapisujeme tak, aby v prvním řádku byly obsaženy názvy proměnných a každý další řádek potom obsahuje právě jedno pozorování. Řádky musí obsahovat stejný počet číselných údajů. Za oddělovače údajů můžeme považovat tabulátory nebo mezery.

Specifikace modelu je organizovaná do určitých sekcí.

[Model Description]

Zde máme možnost uvést popis modelu, který se nadepíše do výsledného souboru. V podstatě slouží pro lepší orientaci v souborech. Text vždy ohraničíme uvozovkami.

[Choice]

Říká programu, kde najde v souboru závislé proměnné.

Zde bychom měli poznamenat, že si musíme dávat pozor na velká a malá písmena. Příkazy pak dostávají jiný smysl.

[Beta]

Každý odhadovaný parametr musí být zmíněn v této sekci. U každého parametru bychom měli uvést následující: jméno parametru, standardní hodnotu, dolní mez, horní mez a podle toho, zda má být parametr odhadován (popřípadě nemá být odhadován) zadat 0 (popřípadě 1).

[LaTeX]

Biogeme generuje výstupní soubor také ve formátu LaTeX. Zde si můžeme pojmenovat proměnné tak, aby vyhovovaly LaTeXové syntaxi.

[Utilities]

Zde definujeme funkci utility. Specifikace pro každou alternativu musí začínat na novém řádku. Jsou zde definovány tyto předpoklady:

1. Identifikátor alternativy – číslování je v souladu se sekci [Choice]
2. Jméno alternativy
3. Stav dostupnosti
4. Tvar lineární funkce utility – parametry, které se objeví zde ve funkci utility se musí objevit i v sekci [Beta]

[Expressions]

Některé atributy nejsou přímo dostupné ze souboru dat. V této sekci programu Biogeme popíšeme, jak si dopočítá potřebné atributy. Buď můžeme hodnoty porovnávat a ponechávat/vylučovat nebo je můžeme přepočítat, například na jiné jednotky.

[Exclude]

Zde si nadefinujeme, jak už sám název sekce napovídá, které složky ze vzorku vyřadíme. Vyhodnotí každé pozorování z datového souboru a rozhodne o jeho zařazení či vyloučení. Zde tedy nastavíme kritéria vyloučení.

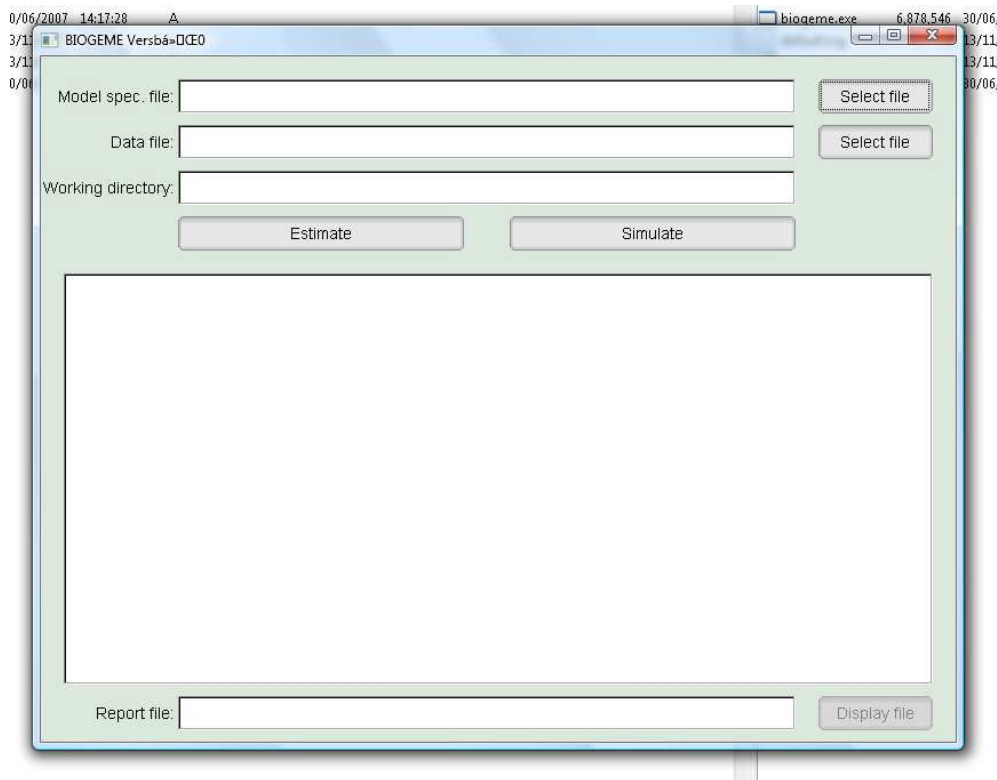
[Model]

Zde uvedeme, který typ modelu je třeba odhadnout. Můžeme si na tomto místě uvést příklady:

- \$BP pro binární probitový model
- \$MNL pro logitový model
- \$NL pro nested logitový model
- \$CNL pro cross-nested logitový model

Uvedli jsme zde jen pár ukázek sekcí pro představu. Sekcí existuje mnoho a užíváme je v závislosti na tom, co od modelu požadujeme.

Pro soubor s daty platí přísná pravidla struktury, která jsme si ukazovali výše. Jestliže máme vytvořen datový soubor a soubor se specifikací souboru, vložíme je jednoduše do následujícího okna.



4.2 Analýza pomocí programu Biogeme

Biogeme zkontroluje, zda existuje soubor s příponou .par, který obsahuje různé parametry. Pokud neexistuje a neexistuje ani soubor default.par, tak si ho Biogeme vytvoří a nastaví výchozí hodnoty parametrů. Jako další krok začne Biogeme načítat data a námi

vytvořenou specifikaci modelu. Nakonec software odhaduje neznámé parametry metodou maximální věrohodnosti.

Biogeme také zobrazuje čas, za který byl schopen všechny tyto operace provést. Nakonec vypíše všechny soubory, které byly vytvořeny. Tyto soubory obsahují výsledky odhadování a další informace.

Výstupní soubory:

- log
- html a rep – výsledky modelu
- res – konečné odhady užití pro koeficienty
- sta – statistika dat
- tex – výstupy v LaTeXovém formátu
- summary.html – přehledový, souhrnný soubor

V dalším textu uvedeme názorné ukázky zdrojových kódů modelů, které jsou modelovány přímo na naše data.

5 Sestavení modelu volby dopravního prostředku

5.1 Základní model

Nejdříve bych se vyjádřila k datům. Získat data z průzkumu je jedna věc, ale dokázat z nich dostat nějaké informace, to je věc druhá. Každý analytik přistupuje k testování dat jinak, proto je možné, že dva různí analytici dostanou z totožného datového souboru jiné informace.

Jak už bylo řečeno v dotazníkovém šetření bylo provedeno 5772 cest. Bylo na výběr 9 módů – *chůze, kolo, autobus, tramvaj/metro, vlak, moped, auto jako řidič, auto jako spolujezdec a jiný dopravní prostředek*. Provádět analýzu dat pro jednotlivé módy je zbytečné, protože například u *moped* jako dopravní prostředek zvolilo pouze 22 respondentů, a z tak nízkého počtu pozorování nelze věrohodně odhadnout hodnoty parametrů užitkových funkcí, proto je dopravní mód *moped* z modelu vyřazen. *Autobus, tramvaj/metro* a *vlak* jsou sloučeny do jedné skupiny, protože jsou to všechno prostředky veřejné dopravy a je zbytečné je analyzovat zvlášť. Nakonec jsou analyzovány následující čtyři módy:

- Auto jako řidič
- Kolo
- Veřejná doprava
- Chůze

Abychom odhadli parametry, které zachycují efekt atributů *cestovního času* a *cestovních nákladů*, bylo třeba strojově do datového souboru dodat vlastnosti způsobů těch dopravních módů, které pro cestování nebyly zvoleny. Cestovatel nám totiž udal, že si pro cestování vybral určitý dopravní mód a k tomuto módu dodal informace o *cestovním času* a *cestovních nákladech*. My jsme dostali pouze informaci o tom, že respondent zvolil například *veřejnou dopravu* jako dopravní mód pro cestování, jaký byl *cestovní čas* pro tento mód a potřebovali jsme pro sestavení modelu znát i informace o tom, jaký *cestovní čas* by byl pro *auto, kolo* a *chůzi*. V rámci doplňování dat bylo území, na kterém se cestovalo, rozděleno do zón a na základě toho, jaká byla výchozí zóna a kde cílová zóna se určoval celkový *cestovní čas* specifickým dopravním módem. Občas jsme narazili na případ, kdy byla cesta vykonána intrazonálně, což znamenalo to, že celkový čas cesty byl 0. Tato pozorování jsme z datového souboru vyloučili.

Stejně tak jako s *cestovním časem* jsme zacházeli i s *délkou cesty*. Data byla opět strojově doplněna. Protože by byla tato data (*cestovní čas, cestovní vzdálenost*) silně

korelovaná, rozhodli jsme se, že do modelu zahrneme odvozenou veličinu – *náklady na cestu*, k čemuž jsme přistupovali následovně. Uvažovali jsme nad tím, že za *chůzi* a *kolo* neplatíme žádné náklady. *Náklady* pro auto byly stanoveny jako *délka cesty*, násobená průměrnou cenou benzínu. Nakonec jsme délku cesty násobili konstantou 2,2 Kč. Tato částka byla zjištěna na základě průměrné spotřeby automobilů a průměrné ceny benzínu v poslední době. U veřejné dopravy se *náklady* stanovovaly obtížněji. Museli jsme brát v úvahu, že respondent může být vlastníkem časové jízdenky, takže jeho náklady na cestu jsou nulové. Dále jsme vzali v úvahu mapu Jihomoravského kraje, která byla členěna do zón a v kombinaci s nalezením ceníku pro stanovení jízdného na základě času a počtu projetých zón jsme stanovili *cestovní náklady*.

Cestovní čas a *cestovní vzdálenost* nám slouží jako vysvětlující proměnné modelu. Každý řádek v datovém souboru značí jedno pozorování, samozřejmě kromě řádku prvního, který obsahuje názvy sloupců. V našem datovém souboru je v prvním sloupci *volba* dopravního prostředku respondenta. Další čtyři sloupce potom určují dostupnost dopravních módů a následují čtyři sloupce *cestovních časů* postupně pro *auto*, *kolo*, *veřejnou dopravu* a *chůzi* a nakonec datový soubor obsahuje *cestovní náklady* pro *auto* a *veřejnou dopravu*. Pro lepší představu je zde dodán vzorek dat.

CHOICE	av1	av2	av3	av4	C_time	B_time	P_time	W_time	C_nakl	P_nakl
1	1	1	1	1	10.52818	16.11552	28.21327	89.18369	13.98206	20
1	1	1	1	1	10.52818	16.11552	28.21327	89.18369	13.98206	20
3	0	0	1	1	17.80738	26.01954	41.88668	171.8177	28.56573	0
3	0	0	1	1	17.80738	26.01954	41.88668	171.8177	28.56573	0
1	1	1	1	1	9.435581	17.14331	37.78629	120.0112	18.54195	27
1	1	1	1	1	9.435581	17.14331	37.78629	120.0112	18.54195	27
1	1	1	1	1	17.73565	26.67444	47.52134	159.1168	24.36105	25
1	1	1	1	1	20.71938	71.74152	53.73421	466.8979	63.27752	34
1	1	1	1	1	17.73565	26.67444	47.52134	159.1168	24.36105	25

Nyní bych se ráda vyjádřila k tomu, jak model zadáváme do programu Biogeme. Strukturu Biogeme jsme si ukazovali již výše, zde již k našemu konkrétnímu modelu. Sekce *Choice* nám definuje, které sloupce má Biogeme brát jako identifikátor vybraných možností. My máme pro jednoduchost sloupec alternativ nazván *CHOICE*. V sekci *Beta* definujeme parametry, které jsou zahrnuty do funkcí utilit. V ukázce modelu, který je uveden níže vidíme, že do multinomického modelu jsme zahrnuli čtyři konstanty specifické pro všechny dopravní módy. Dále jsme uvažovali v modelu parametry pro *cestovní čas*. Ke každému dopravnímu módu jsme vložili parametr, který je specifický pro určitý dopravní mód, což je z toho důvodu, že *cestovní čas* ve veřejné dopravě není vnímán stejně, jako *cestovní čas* strávený za volantem. Ve veřejné dopravě se můžeme věnovat jiným aktivitám – čtení knihy, vyřizování emailů, dělání domácích úkolů, což za volantem nemůžeme (Ettema & Verschuren 2007,

Multitasking and Value of Travel Time Savings). Proto jsme zvolili specifické parametry. A poslední parametry zahrnuté do modelu jsou *cestovní náklady*. O těch jsme hovořili už u tvorby dat, takže jen v rychlosti, máme dvě specifické alternativy pro náklady. Budou vystupovat u módů *auto* a *veřejné doprava*. U každého parametru v Biogeme musíme nadefinovat:

- standardní hodnotu, která bude použita jako výchozí bod pro odhad, obvykle volíme 0
- dolní a horní hranice, slouží jako hranice pro algoritmus
- status, který může nabývat hodnoty 0, v případě že má být parametr odhadován a 1 v případě, kdy má zůstat na zvolené standardní hodnotě – tzv. referenční hodnota

V sekci Utilities potřebujeme nadefinovat:

- identifikátor alternativy, který musí být v souladu s identifikátorem daným v sekci *Choice*
- jméno alternativních módů
- dostupnost módu
- specifikaci deterministické funkce utility, která je lineární v parametrech. Jména proměnných a parametrů musí být stejná, jako jsme si je nadefinovali v datovém souboru a *Beta* sekci

Ještě něco málo k dostupnosti módu. Pokud si s touto částí nechceme lámat hlavu, stačí zadat všude u proměnných *av1*, *av2*, *av3*, *av4* jedničky a o dostupnosti dopravního módu nebudou pochyby, což jsme v podstatě provedli u dopravních módů *chůze* a *veřejná doprava*. U *chůze* není pochyby, ale u *veřejné dopravy* by mohl někdo namítat, že je důležitá vzdálenost zastávky a že čím vyšší bude vzdálenost od zastávky, tím menší bude pravděpodobnost, že si tento dopravní mód respondent vybere. To je pravda. Avšak po důkladném shlédnutí dat jsem zjistila, že většina respondentů (93%) má zastávku veřejné dopravy vzdálenou do 10 minut chůze, proto jsem usoudila, že zde není potřeba omezení. Na dostupnost dopravních módů *auta* a *kola* se dotazoval průzkum, takže zde byla použita tato data.

Otázkou je, zda je zadávání dostupnosti dopravního prostředku důležité. Odpověď zní, že je to důležité, protože kdybychom nevyloučili neexistující alternativu, bude jí přiřazena nenulová pravděpodobnost, že bude zvolena. Jestliže tedy respondent odpoví, že zvolil dopravní prostředek auto, a zároveň toto pozorování bude obsahovat informaci, že daný mód nemá respondent k dispozici, tak bude pozorování softwarem vyloučeno.

V sekci *Exclude* jsme pouze vyřadili ta pozorování, u kterých nebyl zvolen žádný dopravní mód (0) – nepodnikli žádnou cestu.

Zde si konečně můžete prohlédnout ukázkou modelu zadávaného do Biogeme:

```
[ModelDescription]
```

```
"MNL-základní model"
```

```
[Choice]
```

```
CHOICE
```

```
[Beta]
```

```
// Name Value LowerBound UpperBound status (0=variable, 1=fixed)
```

```
ASC1 0 -10000 10000 1
```

```
ASC2 0 -10000 10000 0
```

```
ASC3 0 -10000 10000 0
```

```
ASC4 0 -10000 10000 0
```

```
B_TT_C 0 -10000 10000 0
```

```
B_TT_B 0 -10000 10000 0
```

```
B_TT_P 0 -10000 10000 0
```

```
B_TT_W 0 -10000 10000 0
```

```
B_nakl_C 0 -10000 10000 0
```

```
B_nakl_P 0 -10000 10000 0
```

```
[Utilities]
```

```
// Id Name Avail linear-in-parameter expression (beta1*x1 + beta2*x2 +  
... )
```

```
1 CAR av1 ASC1 * one + B_TT_C * C_time + B_nakl_C * C_nakl
```

```
2 BIKE av2 ASC2 * one + B_TT_B * B_time
```

```
3 PUBLIC av3 ASC3 * one + B_TT_P * P_time + B_nakl_P * P_nakl
```

```
4 WALK av4 ASC4 * one + B_TT_W * W_time
```

```
[Expressions]
```

```
one = 1
```

```
[Model]
```

```
// Currently, only $BP (Binary probit), $MNL (multinomial logit), $NL
```

```
//(nested logit), $CNL
```

```
// (cross-nested logit) and $NGEV (Network GEV model) are valid //keywords
```

```
$MNL
```

```
[Exclude]
```

```
CHOICE = 0
```

Po dosazení tohoto modelu spolu s upravenými daty Jihomoravského kraje do konzoly Biogeme získáme následující výsledky.

```

Model: Multinomial Logit
Number of estimated parameters: 9
    Number of observations: 3022
      Number of individuals: 3022
        Null log-likelihood: -3712.889
        Init log-likelihood: -3712.889
        Final log-likelihood: -2735.307
    Likelihood ratio test: 1955.164
      Rho-square: 0.263
    Adjusted rho-square: 0.261
    Final gradient norm: +8.724e-005
      Diagnostic: Convergence reached...
      Iterations: 13
      Run time: 00:05
    Variance-covariance: from analytical hessian
      Sample file: Data_základ.dat

```

Utility parameters

Name	Value	Std err	t-test	p-val	Rob. std err	Rob. t-test	Rob. p-val
ASC1	0.00	--fixed--					
ASC2	-2.93	0.150	-19.62	0.00	0.158	-18.62	0.00
ASC3	-0.401	0.110	-3.64	0.00	0.114	-3.52	0.00
ASC4	-1.09	0.0768	-14.17	0.00	0.0795	-13.69	0.00
B_TT_B	-0.0337	0.00617	-5.46	0.00	0.00685	-4.92	0.00
B_TT_C	-0.0227	0.0142	-1.60	0.11 *	0.0152	-1.50	0.13 *
B_TT_P	-0.0271	0.00522	-5.19	0.00	0.00550	-4.93	0.00
B_TT_W	-0.00538	0.000776	-6.92	0.00	0.000874	-6.15	0.00
B_nakl_C	-0.0176	0.00674	-2.61	0.01	0.00692	-2.54	0.01
B_nakl_P	-0.0332	0.00271	-12.24	0.00	0.00315	-10.55	0.00

Utility functions

```

2   BIKE      av2   ASC2 * one + B_TT_B * B_time
1   CAR       av1   ASC1 * one + B_TT_C * C_time + B_nakl_C * C_nakl
3   PUBLIC    av3   ASC3 * one + B_TT_P * P_time + B_nakl_P * P_nakl
4   WALK      av4   ASC4 * one + B_TT_W * W_time

```

Vidíme, že výsledný soubor obsahuje spoustu informací. Hned na začátku máme údaj o počtu odhadnutých parametrů a o počtu pozorování zahrnutých do modelu. Dále je zde údaj o koeficientu determinace, ten nám v tomto případě vychází 0,263. Potom se tu objevují i hodnoty log-likelihood funkcí. Jako první se objevuje hodnota log-likelihood funkce, kdy jsou všechny parametry rovny nule, potom funkce pouze s konstantami specifickými pro každou alternativu a nakonec hodnota log-likelihood funkce pro model s odhadnutými parametry metodou maximální věrohodnosti. Vidíme i vypočtenou hodnotu testu poměrem věrohodností. Dále je ve výsledném datovém souboru uveden počet iterací, které byly potřeba k dosažení konvergence. Co se odhadnutých parametrů týče, nacházejí se v sekci *Utility parameters*, kde najdeme informace jako název parametru, odhad parametru, směrodatná odchylka odhadu parametru a t-statistika korespondující ke každému parametru.

Při první letném pohledu na odhadnuté parametry můžeme zkontrolovat, že všechny parametry týkající se *cestovního času a nákladů* jsou záporné. Což je hodnota, kterou jsme očekávali. Záporná hodnota parametrů *dopravního času* se dá interpretovat tak, že nemáme radost z toho, že musíme svůj volný čas trávit v dopravním prostředku. To samé platí u *cestovních nákladů*. Nedělá nám dobře, když musíme za dopravu platit. Nic s tím neuděláme, ale radost nám to rozhodně nepřinese. Také si můžeme uvědomit, že se zvyšujícím se *cestovním časem* a se zvyšujícími se *náklady* utilita klesá.

Teď už k jednotlivým hodnotám parametrů. Hodnoty *ASC* jsou specifické pro všechny dopravní módy. Mohou nabývat jak kladných, tak záporných hodnot. Pro účely odhadu jsme zvolili dopravní mód *auto* jako referenční. Zbývající odhadnuté hodnoty nám říkají, že pokud by byly zbývající části všech alternativních utilit stejné, tak by byl preferován dopravní mód *auto*, protože hodnoty ostatních parametrů mají záporné znaménko. Po preferované možnosti *auta* by následoval mód *veřejné dopravy*, *chůze* a nakonec bychom s nejmenší pravděpodobností zvolili *kolo*.

Cestovní čas nám nejméně vadí u *chůze*, dále u *auta*, *veřejné dopravy* a nakonec je zase nejhorší u dopravního módu *kola*. Je důležité si ale všimnout, že *cestovní čas* u *auta* je statisticky nevýznamný. Předpokládáme, že je to způsobeno tím, že *cestovní čas* silně koreluje s *náklady na cestu*.

Náklady musíme interpretovat vzhledem k nemotorovým alternativám v modelu (*chůze*, *kolo*). A z odhadnutých parametrů vyplývá, že oproti módům, kde nemusím platit nic, je pro nás méně bolestnější sponzorovat *auto*.

5.2 Podrobnější model

Do modelu jsme se snažili dostat také socioekonomickou složku. Nejdříve bychom si měli představit data, která jsme do modelu vkládali a co jsme s nimi prováděli. O datech *cestovního času a cestovních nákladech* jsme již diskutovali výše. Další vysvětlující proměnnou je v našem modelu *pohlaví*. Zde nebylo třeba velkých úprav, *muž* je označen 0 a *žena* je potom kódována 1. U *typu domácnosti* jsem postupovala následovně. Zde byla vysvětlující proměnná reprezentována počtem osob v domácnosti. Nejdříve jsem v modelu nechala data původní, což znamenalo, že se zde vyskytovaly domácnosti s jedním až osmi členy. Posléze jsem data agregovala tak, že jsem přemýšlela nad tím, kde by u počtu osob mohl nastat zvrát. Nakonec jsem data rozdělila do dvou kategorií - na *malou domácnost*, která čítala až tři členy a byla označena 0 a *velká domácnost*, která potom obsahovala 4 a více členy. V datovém souboru byla tato proměnná označena 1. To z toho důvodu, že máme omezený počet dopravních prostředků a u více členné domácnosti by mohla nastat situace, kdy už některý dopravní prostředek (*auto, kolo*) nebude k dispozici. V datech se vyskytla proměnná, která se respondentů dotazovala na to, zda mají k dispozici osobní automobil. Odpovědi zněly *vždy, občas a nikdy*. Odpovědi *vždy* a *nikdy* jsme do modelu zahrnuli pomocí nadefinování dostupnosti dopravního módu. Zbývá data týkající se odpovědi *občas* byly do modelu zahrnuty pomocí samostatného parametru. Dále se zde objevují data, týkající se otázky, zda má respondent doma k dispozici *místo na parkování*, jestliže je odpověď *ano*, pak je označena 1. *Věk* respondenta jsme ponechali jako spojitou proměnnou. Dále jsme do modelu zahrnuli *práci*. Chtěli jsme zahrnout i *vzdělání*, ale tyto dva parametry byly silně korelované, proto jsme se rozhodli pro zařazení pouze *práce*. Vysvětlující proměnná *práce* je kategorická a vyskytují se zde tři kategorie – respondent *pracuje, nepracuje, studuje*. Kategorii *pracuje* jsme zvolili jako referenční. Jako poslední dvě vysvětlující proměnné jsou strojově vytvořená data pomocí softwaru R. Tato data mi vytvořilo CDV a značí *počet přestupů* mezi počáteční a cílovou destinací a *frekvence služeb*, která je definována jako počet spojů za den.

Další data jsme do analýzy buď nezařadili nebo nám při odhadech parametrů vyšla jako nevýznamná.

Při sestavování funkcí utilit jsem postupovala následovně. Základní myšlenka při sestavování modelu byla, že než jsem vysvětlující proměnnou vložila do určité funkce utility, pořádně jsem se nad ní zamyslela a vytvořila si hypotézu, která by podle mě mohla platit. Tuto hypotézu jsem si potom dosazením do modelu buď potvrdila nebo vyvrátila.

Model vypadá následovně, protože jsme se co nejvíce snažili, abychom co nejlépe vysvětlili variabilitu modelu, ale zároveň jsme požadovali to, aby nebyl moc složitý. Šlo nám o určitý kompromis.

```

1  CAR    av1  ASC1 * one + B_TT_C * C_time + B_nakl_C * C_nakl + B_gen1
      * P_gen + B_str1 * H_person_n + B_obcas * P_car_obc +
      B_hpark * P_hparking

2  BIKE   av2  ASC2 * one + B_TT_B * B_time + B_vek2 * P_age + B_gen2 *
      P_gen + B_workN * Nepr + B_workS * Stud + B_str2 *
      H_person_n

3  PUBLIC av3  ASC3 * one + B_TT_P * P_time + B_nakl_P * P_nakl + B_workN
      * Nepr + B_workS * Stud + B_freq * SerFreq + B_trans *
      NoTrans

4  WALK   av4  ASC4 * one + B_TT_W * W_time + B_vek4 * P_age + B_workN *
      Nepr + B_workS * Stud

```

A výsledek :

Model: Multinomial Logit

Number of estimated parameters: 21

Number of observations: 3022

Number of individuals: 3022

Null log-likelihood: -3712.889

Init log-likelihood: -3712.889

Final log-likelihood: -2488.760

Likelihood ratio test: 2448.259

Rho-square: 0.330

Adjusted rho-square: 0.324

Final gradient norm: +7.088e-003

Diagnostic: Convergence reached...

Iterations: 13

Run time: 00:08

Variance-covariance: from analytical hessian

Sample file: Data.dat

Utility parameters

Name	Value	Std err	t-test	p-val	Rob. std err	Rob. t-test	Rob. p-val

-----	-----	-----	-----	-----	-----	-----	-----
ASC1	0.00	--fixed--					
ASC2	-3.65	0.383	-9.55	0.00	0.406	-8.99	0.00
ASC3	-0.922	0.183	-5.04	0.00	0.214	-4.32	0.00
ASC4	-0.927	0.175	-5.29	0.00	0.183	-5.08	0.00
B_TT_B	-0.0250	0.00607	-4.12	0.00	0.00688	-3.64	0.00
B_TT_C	-0.0423	0.0153	-2.76	0.01	0.0162	-2.61	0.01
B_TT_P	-0.0162	0.00625	-2.59	0.01	0.00783	-2.07	0.04
B_TT_W	-0.00390	0.000764	-5.11	0.00	0.000914	-4.26	0.00
B_freq	0.000868	0.000319	2.73	0.01	0.000341	2.54	0.01
B_gen1	-0.387	0.107	-3.62	0.00	0.105	-3.71	0.00
B_gen2	-1.08	0.238	-4.53	0.00	0.236	-4.57	0.00
B_hpark	0.563	0.111	5.07	0.00	0.109	5.17	0.00
B_nakl_C	-0.00234	0.00734	-0.32	0.75*	0.00812	-0.29	0.77*
B_nakl_P	-0.0409	0.00436	-9.38	0.00	0.00498	-8.21	0.00
B_obcas	-1.76	0.126	-14.01	0.00	0.123	-14.29	0.00
B_str1	0.397	0.107	3.69	0.00	0.104	3.82	0.00
B_str2	0.401	0.247	1.62	0.10*	0.253	1.58	0.11*
B_trans	-0.256	0.0970	-2.64	0.01	0.102	-2.51	0.01
B_vek2	0.0193	0.00642	3.00	0.00	0.00625	3.08	0.00
B_vek4	-0.0119	0.00273	-4.37	0.00	0.00299	-3.99	0.00
B_workN	0.370	0.130	2.85	0.00	0.130	2.84	0.00
B_workS	0.916	0.210	4.37	0.00	0.227	4.04	0.00

Můžete si hned povšimnout, že i přes agregaci dat a vyjmutí nevýznamných parametrů se tu dva nevýznamné parametry (na hladině 0,05) přece jen objevily. V modelu jsem ponechala nevýznamný parametr týkající se *typu domácnosti*, protože by byl významný na hladině 0,1 a parametr pro *náklady* u auta jsem v modelu ponechala, protože byl už v základním modelu a bylo by nelogické, kdyby se objevily *náklady* u *veřejné dopravy* a u *auta* ne.

Co se konstant specifických pro všechny dopravní módy týče, interpretace je podobná jako v modelu předchozím. Kdybychom uvažovali pouze tyto parametry a zbytky utilit by si byly pro všechny dopravní módy rovny, tak by byl opět nejvíce preferovaný mód *auto*, následován *veřejnou dopravou*, *chůzí* a poslední by byl mód *kolo*. Parametry *cestovního času* jsou opět záporné, jak se dalo očekávat. Nejméně by nám, co se *cestovního času* týče vadila cesta *pěšky*, poté cesta *veřejnou dopravou*, *kolo* a nakonec *auto*. Trochu mě překvapuje parametr *cestovního času* u *auta*. Myslela jsem si, že pro nás bude příjemnější strávit čas ve vlastním automobilu, komfortu, čistotě než například ve veřejné dopravě. Zdůvodňuji si to tak, že pokud jdeme pěšky, dýcháme čerstvý vzduch, děláme něco pro své tělo. Ve veřejné

dopravě můžeme zase vyřídit emaily, a při jízdě na kole platí to samé, co u chůze. Takže se zdá, že tento případ není zas až tak nepochopitelný a nevysvětlitelný. Parametr *pohlaví* interpretujeme vzhledem k dopravním módům *chůze* a *veřejná doprava*. Do těchto utilit jsem tento parametr nezařadila, protože si myslím, že na tyto módy nemá *pohlaví* vliv. U parametru *pohlaví* nám skutečnost, že respondent je *žena* snižuje pravděpodobnost výběru dopravního módu *auta*. A dále, pokud je respondent *žena*, tak nám klesá i pravděpodobnost výběru *kola* jako dopravního prostředku oproti ostatním módům. Jestliže respondent udal, že auto má k dispozici pouze *občas*, tak pak pravděpodobnost jeho volby *auta* pro uskutečnění cesty klesá a volí raději alternativní dopravní módy. *Auto* pak používá pouze v případě nutnosti, flexibility. Naopak skutečnost, že má respondent doma k dispozici *parkovací místo* zvyšuje pravděpodobnost výběru *auta* jako dopravního prostředku. Na základě parametru týkajícího se *věku* respondentů jsme zjistili, že s rostoucím věkem je pravděpodobnost volby *kola* vyšší než volba *auta* a *veřejné dopravy*. Můžeme si to zdůvodnit například tím, že starší lidé se bojí jezdit *autem*, proto raději zvolí *kolo*. Dokonce zde může hrát roli i fakt, že starší lidé spíše volí *kolo* než aby šli *pěšky* – na kole si mohou odvézt například nákup. Druhý parametr týkající se věku (*B_vek4*) nám říká, že s rostoucím věkem se snižuje pravděpodobnost, že člověk půjde *pěšky*, oproti dopravním módům, že pojede *autem* nebo *veřejnou dopravou*. Parametr týkající se *typu domácnosti* jsem zahrнула do utilit u *auta* a u *kola*, protože v případě větší domácnosti může docházet k „pranici“ o *auto* nebo *kolo*. Podle výsledku můžeme soudit, že pravděpodobnost volby *auta* jako dopravního prostředku se s *větší domácností* zvýší oproti módům *veřejná doprava* a *chůze*, což může být způsobeno tím, že když jede jeden člen domácnosti například do práce, může s sebou vzít i některé ostatní členy domácnosti v podstatě v rámci spolujízdy, tato hypotéza by však potřebovala hlubší analýzu, aby mohla být s jistotou potvrzena. Druhý parametr týkající se domácnosti (*B_str2*) zvyšuje pravděpodobnost volby *kola* na úkor *veřejné dopravy* a *chůze*. Tento parametr není statisticky významný. Parametr týkající se *práce* nám říká, že *nepracující* lidé a lidé *studující* budou s větší pravděpodobností volit všechny ostatní módy raději než *auto* i přes to, že jej mají k dispozici. Poslední parametry nám zcela logicky říkají, že pokud se bude zvyšovat *počet přestupů*, tak se bude utilita snižovat a pokud se bude *frekvence spojů* zvyšovat, tak se bude nepatrně zvyšovat i pravděpodobnost výběru *veřejné dopravy*.

5.3 Modely pro jednotlivé účely

Teď už zbývá do modelů pouze přidat účel cesty, aby byla naše analýza kompletní. Naším předpokladem je, že se budou parametry mezi různými účely lišit. I když jsme původně dotazovali osoby na deset různých účelů, ale protože například na účel venčení psa odpovědělo pouze 13 respondentů, rozhodli jsme se toto a další podobné případy nezahrnovat do analýzy. Nakonec jsme vybrali čtyři nejčastější účely cest. Cesta do *práce*, do *školy*, na *nákup*, *cesta za volno – časovými aktivitami* a *cesta domů*.

5.3.1 Cesta do práce

Do modelu, kde byl cíl cesty *cesta do práce* jsem zařadila tyto vysvětlující proměnné: *cestovní čas* specifický pro všechny dopravní módy, *cestovní náklady*, parametr týkající se *vlastnictví auta pouze občas*, parametr vlastního *parkovacího místa doma* a parametr vlastního *parkovacího místa v práci*. Dále parametr *frekvence spojů* a parametr *počtu přestupů*. Data jsou vytvořena stejně jako tomu bylo v minulém případě, proto si myslím, že není potřeba rozebírat jednotlivé vysvětlující proměnné.

Jediné, co je zde navíc oproti minulému modelu, je informace o *soukromém parkovacím místě v práci*. Naopak přibylo parametrů, které ztratily na významnosti a byly z modelu vyloučeny. Například parametr *práce*, který je zde nevýznamný. Vyřadili jsme také parametr *věk* a *typ domácnosti*, protože také ztratily svůj význam.

Snažila jsem se zařadit do modelu i jiné vysvětlující proměnné, jako je například *vzdálenost na zastávku* nebo *sleva* na veřejnou dopravu, ale žádný z parametrů jako významný nevyšel. Model pro účel *cesty do práce* pak vypadá takto:

```
1  CAR    av1    ASC1 * one + B_TT_C * C_time + B_nakl_C * C_nakl + B_obcas
      * P_car_obc + B_hpark * P_hparking + B_wpark * P_w_parking
      + B_gen1 * P_gen

2  BIKE   av2    ASC2 * one + B_TT_B * B_time + B_gen2 * P_gen

3  PUBLIC av3    ASC1 * one + B_TT_P * P_time + B_nakl_P * P_nakl + B_freq
      * SerFreq + B_trans * NoTrans

4  WALK   av4    ASC4 * one + B_TT_W * W_time
```


Výsledky modelu pro účel *cesty do práce*:

Model: Multinomial Logit

Number of estimated parameters: 14

Number of observations: 647

Number of individuals: 647

Null log-likelihood: -817.345

Init log-likelihood: -817.345

Final log-likelihood: -521.775

Likelihood ratio test: 591.138

Rho-square: 0.362

Adjusted rho-square: 0.344

Final gradient norm: +4.937e-003

Diagnostic: Convergence reached...

Iterations: 12

Run time: 00:01

Variance-covariance: from analytical hessian

Sample file: Cesta_do_prace+park.dat

Utility parameters

Name	Value	Std err	t-test	p-val	Rob. std err	Rob. t-test	Rob. p-val
ASC1	0.00	--fixed--					
ASC2	-2.24	0.393	-5.71	0.00	0.444	-5.05	0.00
ASC4	-1.40	0.224	-6.26	0.00	0.237	-5.90	0.00
B_TT_B	-0.0264	0.0128	-2.06	0.04	0.0191	-1.38	0.17 *
B_TT_C	-0.0520	0.0219	-2.37	0.02	0.0246	-2.12	0.03
B_TT_P	-0.0280	0.0118	-2.37	0.02	0.0126	-2.23	0.03
B_TT_W	-0.00344	0.00148	-2.32	0.02	0.00172	-1.99	0.05
B_freq	0.000510	0.000637	0.80	0.42	* 0.000637	0.80	0.42 *
B_gen1	-0.633	0.210	-3.02	0.00	0.194	-3.26	0.00
B_gen2	-0.972	0.480	-2.02	0.04	0.479	-2.03	0.04
B_hpark	0.739	0.209	3.54	0.00	0.199	3.71	0.00
B_nakl_P	-0.0637	0.00911	-7.00	0.00	0.00956	-6.66	0.00
B_obcas	-1.71	0.261	-6.58	0.00	0.251	-6.82	0.00
B_trans	-0.170	0.197	-0.86	0.39	* 0.185	-0.92	0.36 *
B_wpark	0.576	0.231	2.50	0.01	0.224	2.57	0.01

Z výsledných odhadů parametrů vidíme, že jsou zde dva parametry nevýznamné, a to *frekvence spojů a přestupů*. Vysvětluji si to tak, že pokud člověk každodenně rutinně dojíždí do práce, ví přesně, kdy mu spoj jede a i u méně frekventovaných spojů lze předpokládat, že jsou soustředěny do období ranní a odpolední špičky. Protože odhadnuté parametry *cestovního času a cestovních nákladů u auta* vycházely nevýznamné, zhodnotili jsme, že to bude nejspíše silnou korelací. Po odstranění je situace mnohem lepší.

Jinak jsou ale odhadnuté parametry velice podobné těm, které byly odhadnuty v případě minulém. Koeficient determinace nám poskočil ještě o kousek výše, a to na hodnotu 0.362, možná je to proto, že už se zabýváme určitým účelem, data jsou více homogenní, a proto můžeme data lépe analyzovat. Musíme však podotknout, že koeficient determinace závisí na počtu pozorování, proto jej nemůžeme porovnat s předchozím koeficientem.

5.3.2 Cesta do školy/za vzděláním

Další sestavené modely už nejsou tak košaté jako byly předcházející dva modely. Mnoho odhadů parametrů vyšlo nevýznamně, proto jsme byli nuceni některé vysvětlující proměnné z modelu vyloučit.

Data jsou již známa, ale jednu změnu v datovém souboru jsem přece jen provedla. Tato změna bude platit pro všechny tři následující modely. Vysvětlující proměnnou *práce* jsem agregovala a pro následné analýzy modelu jsem vytvořila pouze dvě kategorie proměnných. Uvažovanými kategoriemi jsou, zda se jedná o respondenta *pracujícího* nebo *nepracujícího*. Funkce užitku tedy vypadají takto:

```

1  CAR    av1    ASC1 * one + B_str1 * H_person_n + B_obcas * P_car_obc +
      B_hpark * P_hparking

2  BIKE   av2    ASC1 * one + B_work * P_work

3  PUBLIC av3    ASC1 * one + B_TT_P * P_time + B_trans * NoTrans + B_work
      * P_work + B_sleva * P_sleva

4  WALK   av4    ASC1 * one + B_TT_W * W_time + B_vek4 * P_age + B_work *
      P_work

```

Model: Multinomial Logit

Number of estimated parameters: 9

Number of observations: 213

Number of individuals: 213

```

Null log-likelihood: -237.271
Init log-likelihood: -237.271
Final log-likelihood: -170.960
Likelihood ratio test: 132.623
      Rho-square: 0.279
Adjusted rho-square: 0.242
Final gradient norm: +6.812e-004
      Diagnostic: Convergence reached...
      Iterations: 18
      Run time: 00:00
Variance-covariance: from analytical hessian
Sample file: Cesta_do_školy+data.dat

```

Utility parameters

Name	Value	Std err	t-test	p-val	Rob. std err	Rob. t-test	Rob. p-val
ASC1	0.00	--fixed--					
B_TT_P	0.0689	0.0145	4.76	0.00	0.0181	3.81	0.00
B_TT_W	0.00470	0.00162	2.91	0.00	0.00221	2.13	0.03
B_hpark	2.11	0.980	2.15	0.03	1.41	1.50	0.13 *
B_obcas	-1.39	0.875	-1.59	0.11	* 0.881	-1.58	0.11 *
B_sleva	0.898	0.330	2.72	0.01	0.383	2.35	0.02
B_str1	1.51	0.987	1.53	0.13	* 1.36	1.11	0.27 *
B_trans	-0.983	0.303	-3.25	0.00	0.315	-3.12	0.00
B_vek4	0.101	0.0167	6.06	0.00	0.0163	6.21	0.00
B_work	-2.75	1.29	-2.12	0.03	1.47	-1.87	0.06 *

V modelu vyšly všechny odhadované konstanty specifické pro jednotlivé alternativní možnosti nevýznamné, proto byly z modelu vyloučeny. Odhadnuté parametry *cestovního času* máme pouze pro dopravní módy *veřejná doprava* a *chůze*, ostatní parametry vyšly nevýznamné. Hodnoty parametrů mají kladnou hodnotu, což jsme neočekávali. Parametr, který zahrnuje možnost soukromého parkovacího místa u bydliště (*B_hpark*) říká, že pravděpodobnost volby dopravního prostředku *auta* se zvyšuje, jestliže má respondent parkovací místo k dispozici. Pravděpodobnost volby dopravního prostředku se bude naopak snižovat s faktem, má-li respondent k dispozici auto pouze *občas*. Bude tedy ve větší míře dávat přednost prostředkům *veřejné dopravy, kola a chůze*. Jestliže respondent vlastní nějakou kartu, díky které má právo na *slevu* ve veřejné dopravě, je zde potom větší pravděpodobnost,

že tento cestující zvolí dopravní mód *veřejné dopravy*. Pokud respondent pochází z větší *rodiny*, můžeme tvrdit, že je pravděpodobnější, že zvolí dopravní prostředek *auto*. Tento parametr však vychází nevýznamně. Pokud na respondenta čeká na cestě velké množství *přestupů* ve *veřejné dopravě*, je vypočteno, že to bude důvod pro to, aby cestující volil jiný dopravní prostředek než *veřejnou dopravu*. Pokud je respondent *pracující* člověk, je zde malá pravděpodobnost, že při cestě za vzděláním zvolí *kolo, chůzi nebo veřejnou dopravu*.

5.3.3 Cesta za volno-časovými aktivitami

Zde už vidíme sestavené funkce utilit pro cesty za *volno-časovými aktivitami*:

- 1 CAR av1 ASC1 * one + B_obcas * P_car_obc + B_hpark * P_hparking
- 2 BIKE av2 ASC1 * one + B_TT_B * B_time + B_gen * P_gen + Beta_work * P_work
- 3 PUBLIC av3 ASC3 * one + B_nakl_P * P_nakl + B_freq * SerFreq + Beta_work * P_work
- 4 WALK av4 ASC1 * one + B_TT_W * W_time + Beta_work * P_work

Model: Multinomial Logit

```

Number of estimated parameters: 9
  Number of observations: 244
    Number of individuals: 244
      Null log-likelihood: -290.236
        Init log-likelihood: -290.236
          Final log-likelihood: -226.223
            Likelihood ratio test: 128.025
              Rho-square: 0.221
                Adjusted rho-square: 0.190
                  Final gradient norm: +1.101e-003
                    Diagnostic: Convergence reached...
                      Iterations: 12
                        Run time: 00:00
                          Variance-covariance: from analytical hessian
                            Sample file: Volný_čas+data.dat

```

Utility parameters

Name	Value	Std err	t-test	p-val	Rob. std err	Rob. t-test	Rob. p-val
ASC1	0.00	--fixed--					
ASC3	-1.01	0.317	-3.18	0.00	0.304	-3.31	0.00
B_TT_B	-0.0213	0.00881	-2.41	0.02	0.00740	-2.87	0.00
B_TT_W	-0.00271	0.000897	-3.02	0.00	0.00101	-2.68	0.01
B_freq	0.00393	0.000849	4.63	0.00	0.000859	4.57	0.00
B_gen	-2.21	0.527	-4.20	0.00	0.520	-4.25	0.00
B_hpark	0.568	0.288	1.97	0.05	0.276	2.05	0.04
B_nakl_P	-0.0308	0.0137	-2.25	0.02	0.0128	-2.40	0.02
B_obcas	-1.20	0.377	-3.19	0.00	0.376	-3.19	0.00
Beta_work	-0.756	0.291	-2.60	0.01	0.259	-2.92	0.00

U výsledků modelu pro cesty konané za *volno-časovými aktivitami* vidíme, že jsme byli opět nuceni redukovat konstanty specifické pro alternativní dopravní módy. Významná je pouze konstanta specifická pro veřejnou dopravu. Z té potom můžeme vyčíst, že kdybychom hodnotili dopravní módy pouze na základě konstant nebo by byly zbylé části všech utilit stejné, respondent bude preferovat dopravní módy *auto, kolo a chůzi* před *veřejnou dopravou*. Co se *cestovního času* týče, zde už nám vychází hodnoty, které jsme očekávali. Navíc motorové dopravní módy (*auto, veřejná doprava*) jsou preferovány před *kolem a chůzí*. Odhadnutý parametr pro *frekvenci spojů* veřejné dopravy nám pravděpodobnost volby *veřejné dopravy* mírně zvyšuje vzhledem k ostatním dopravním prostředkům. Jestliže je respondent *žena*, tak je menší pravděpodobnost, že bude volit dopravní mód *kolo*. Raději zvolí ostatní dopravní prostředky. Pokud máme k dispozici soukromé *parkovací místo*, je i zde větší pravděpodobnost, že bude přicházet více v úvahu *auto*. A opět platí, že jestliže má respondent *auto* k dispozici pouze *občas*, tak bude větší pravděpodobnost, že zvolí *veřejnou dopravu, auto a kolo*. *Náklady* na veřejnou dopravu opět snižují pravděpodobnost volby *veřejné dopravy* vzhledem k ostatním dopravním módům. Poslední parametr nám říká, že *pracující* člověk raději volí *auto* než jiné dopravní prostředky.

5.3.4 Cesta na nákup

V tomto případě jsme do modelu zahrnuli velice malé množství vysvětlujících proměnných, jelikož všechny ostatní proměnné vycházely nevýznamné. Úprava dat je opět stejná jako v případech minulých. Funkce utility pro *cestu na nákup* tedy jsou:

```

1  CAR    av1    ASC1 * one + B_obcas * P_car_obc + B_hpark * P_hparking
2  BIKE   av2    ASC2 * one + B_work * P_work
3  PUBLIC av3    ASC3 * one + B_work * P_work
4  WALK   av4    ASC4 * one + B_work * P_work

```

Model: Multinomial Logit

Number of estimated parameters: 6

Number of observations: 207

Number of individuals: 207

Null log-likelihood: -250.661

Init log-likelihood: -250.661

Final log-likelihood: -155.960

Likelihood ratio test: 189.401

Rho-square: 0.378

Adjusted rho-square: 0.354

Final gradient norm: +4.186e-004

Diagnostic: Convergence reached...

Iterations: 6

Run time: 00:00

Variance-covariance: from analytical hessian

Sample file: Nakup+jen work.dat

Utility parameters

Name	Value	Std err	t-test	p-val	Rob. std err	Rob. t-test	Rob. p-val
ASC1	0.00	--fixed--					
ASC2	-2.88	0.661	-4.36	0.00	0.676	-4.26	0.00
ASC3	-1.37	0.459	-3.00	0.00	0.447	-3.07	0.00
ASC4	-0.634	0.440	-1.44	0.15	* 0.426	-1.49	0.14 *
B_hpark	0.850	0.438	1.94	0.05	* 0.421	2.02	0.04
B_obcas	-2.36	0.491	-4.80	0.00	0.503	-4.69	0.00
B_work	-1.04	0.415	-2.50	0.01	0.422	-2.46	0.01

Tento model obsahuje nejméně odhadnutých parametrů. Tentokrát jsem v modelu zachovala všechny konstanty specifické pro alternativní možnosti. Nutno podotknout, že konstanta odpovídající dopravnímu módu *chůze* je nevýznamná. Konstanty nám o modelu sestaveném

pro cestu na nákup říkají, že pokud by byly zbylé části utilit stejné, nejvíce preferovanou možností by bylo *auto*. Nejméně pravděpodobný scénář by byl, že bychom jeli na nákup na *kole*. Soukromé *parkovací místo* u bydliště nám opět zvyšuje pravděpodobnost volby *auta* jako dopravního prostředku vzhledem k ostatním dopravním módům. Skutečnost, že má respondent auto pouze *občas* nám opět pravděpodobnost volby *auta* snižuje. A pokud je respondent *pracující* člověk, opět nejraději ze všech dopravních prostředků volí *auto*.

Závěr

Jsem velice ráda za to, že jsem si vybrala téma *Diskrétní analýza volby dopravního prostředku* jako téma diplomové práce. Umožnilo mi to spolupracovat s Centrem dopravního výzkumu, kde jsem měla možnost alespoň trochu nahlédnout pod pokličku tak váženého institutu a sledovat vývoj projektů od začátku až do konce. Také si velice cením vážených rad, kterých se mi zde dostalo. Tato práce mi pomohla využít teoretické poznatky, které jsem načerpala ve škole, a zužitkovat je do něčeho smysluplného, celistvého.

Velkým přínosem mi bude i rozšíření mých znalostí, co se softwaru R týče a doufám, že jednou budu moci využít své znalosti, které jsem sesbírala užíváním aplikace Biogeme. V práci jsme dostali všem cílům, které jsme si v úvodní části stanovili. Byl postupně sestaven model, který obsahoval pouze atributy týkající se jednotlivých alternativních možností (cestovní čas a cestovní náklady). Ten jsme dále rozvinuli v model obsahující sociodemografii respondentů. Nakonec jsme se zabývali specifikací modelů pro jednotlivé účely cest. Analyzovali jsme cesty, které byly uskutečněny za účelem cesty do práce, cesty do školy/za vzděláním, cesty za volno-časovými aktivitami, cesty na nákup.

Můžeme si všimnout, když už jsme seznámeni se všemi modely, že parametry se v modelu někdy i drasticky mění. Například v podrobnějším modelu, kde byla zahrnuta sociodemografie se objevilo 21 odhadnutých parametrů, naproti tomu v posledním modelu, kdy byl účel cesty nákup, se v modelu objevilo pouze 6 odhadnutých parametrů. Při sestavování modelů jsem postupovala podobně, ovšem problém byl v tom, že většina parametrů byla nevýznamná (na hladině 0,05).

Literatura

- [1] Ben-Akiva, M. and S.R. Lerman (1985) Discrete Choice Analysis: Theory and Application to Travel Demand, The MIT Press, Cambridge
- [2] Ortuzar, J. de D. and L.G. Willumsen (1997), Modelling Transport, John Wiley & Sons, New York, NY.
- [3] Bierlaire, M. (2003), Biogeme: A free package for estimation of discrete choice models, Proceedings of the 3rd Swiss transportation research conference, Ascona, Switzerland.
- [4] Braun Kohlová, M. (2010), Cesty městem, O racionalitě každodenního cestování, Sociologické nakladatelství, Praha
- [5] Horowitz, J.L., F.S. Koppelman and S.R. Lerman (1986) A self-instructing course in disaggregate mode choice modeling, Final report, prepared for the U.S. Department of Transportation, University Research and Training Program, Washington, D.C.
- [6] Matějka, F. (2011). Rational inattention to discrete choices: A new foundation for the multinomial logit model. CERGE-EI Working Paper Series 442
- [7] Börsch-Supan, A. (1987) Econometric analysis of discrete choice, Lecture notes in economics and mathematical system, Springer – Verlag, Berlin
- [8] Koppelman, F. S., Bhat, Ch (2006) A self-instructing course in mode choice modeling: Multinomial and Nested logit models, prepared for U.S. Department of Transportation, Federal Transit Administration, Washington, DC
- [9] http://en.wikipedia.org/wiki/Gumbel_distribution
- [10] http://en.wikipedia.org/wiki/Transportation_forecasting
- [11] http://en.wikipedia.org/wiki/Mode_choice
- [12] http://www.fsutmsonline.net/images/uploads/reports/FDOT_BC415rpta.pdf

[13] http://eprints.qut.edu.au/16500/1/Omer_Khan_Thesis.pdf