VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
Fakulta informatiky a statistiky

DIZERTAČNÁ PRÁCA

Praha, 2006

Ing. Zuzana Fíglová

VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
Fakulta informatiky a statistiky

Analýza modelov diskrétnej voľby a ich aplikácia

DIZERTAČNÁ PRÁCA

Doktorand: Ing. Zuzana Fíglová
Školiteľ: Prof. Ing. Roman Hušek, CSc.
Odbor: Ekonometrie a operační výzkum

Praha, 2006

PREHLÁSENIE K DIZERTAČNEJ PRÁCI

Prehlasujem, že dizertačnú prácu na tému „Analýza modelov diskrétnej voľby a ich aplikácia“ som vypracovala samostatne. Použitú literatúru a podkladové materiály uvádzam v priloženom zozname literatúry.

V Prahe, dňa 14. 4. 2006

....
Ing. Zuzana Fíglová

POĎAKOVANIE

Rada by som sa touto cestou poďakovala svojmu školiteľovi prof. Romanovi Huškovi za cenné rady a pripomienky, ktoré prispeli ku skvalitneniu predkladanej dizertačnej práce. Zároveň by som chcela poďakovať prof. Janovi Pelikánovi a doc. Václave Pánkovej za ich pripomienky a návrhy, predložené pri predbežnej obhajobe práce.

ABSTRACT

Analysis of discrete choice models and their application

Discrete choice models have become popular and widely used in many diverse areas such as marketing research, consumer behavior, psychology, transportation and others. These models are based on the random utility theory, which assume that each consumer maximizes his utility through the consumption of goods and services. We focus on the analysis of linear and nonlinear discrete choice models and their application to the real datasets. We discuss advantages and disadvantages in the application of the linear and nonlinear approach (logit and probit model) to the analysis of consumer behavior. The parameters of the nonlinear models are estimated by the maximum likelihood method. In the first application we analyze the consumer behavior of households in the Czech Republic and its socioeconomic determinants by using binary choice models. The data were obtained from the household survey *Social Situation of Households in the Czech Republic, 2001*. More than 10,000 households participated in the national survey carried out by the Czech Statistical Office. In the estimated models we analyzed and interpreted significant predictors of owning an Internet connection by a household. In the second application we predict whether an applicant for credit/goods will pay back for his liabilities by using a cumulative multinomial logit and probit models on a dataset from a financial institution with more than 15,000 observations included.

Keywords: discrete choice models, random utility theory, consumer behavior, credit scoring

ABSTRAKT

Analyse der Discrete Choice Modellen und ihre Applikation

Discrete Choice Modelle sind sehr populär und weit benützt auf vielen verschiedenen Gebieten, wie zum Beispiel Marketing Forschung, Konsumente Verhaltung, Psychologie, Verfrachtung und andere. Diese Modelle basieren auf der Zufallsnutzentheorie, dass jedes Individuum seinen Nutzen maximiert. Der Fokus dieser Arbeit liegt auf der Analyse der linearen und nichtlinearen Discrete Choice Modellen und ihren Applikationen zum realistischen Datenbestanden. Wir diskutieren über Vorteilen und Nachteilen in der Applikationen lineares und nichtlineares Zugriffs (Logit und Probit Modell) auf der Analyse der Konsumenten Verhaltung. Die Schätzung der Effektstärken im Logit und Probit Modell beruht auf dem sogenannten Maximum-Likelihood-Schätzverfahren. In der ersten Applikation, analysieren wir die Konsumenten Verhaltung der Haushaltungen in der Tschechischen Republik und ihre sozioökonomische Faktoren im Einsatz von binären Choice Modellen. Die Daten erhalten wir aus der Haushaltungen Stichprobensprüfung *Soziale Situation der Haushaltungen in der Tschechischen Republik, 2001.* Mehr als 10 000 Haushaltungen beteiligten sich der nationalen Stichprobensprüfungen, die das Tschechische Statistische Amt realisiert. In den schatzgebühren Modellen analysieren wir und interpretieren signifikante Prediktoren des Haushaltungenaussattungen Internetanschlusses. In der zweiten Applikation vorzusagen wir, ob der Kandidat für die Kredit rückzahlzt ihn, durch den kumulativen multinomialen Logit und Probit Modellen. Die Daten (mit mehr als 15 000 Beobachtungen) haben wir aus einer Finanzinstitution erhalten.

Schlagwörter: Discrete Choice Modelle, Zufallsnutzentheorie, Konsumente Verhaltung, Credit Scoring

OBSAH

1 ÚVOD

Moderná ekonomická teória, reprezentovaná neoklasickou školou, je založená na „homo oeconomicus“, teda na človeku motivovanom vlastným záujmom a schopnom racionálne uvažovať a rozhodovať sa na základe dokonale dostupných informácií. Uvedený model ľudského správania aj napriek mnohým výhradám a problémom tvorí základ ekonomických a sociálnych vied.

Samotné rozhodovanie sa jednotlivca by sme mohli definovať ako rozlíšenie a uprednostnenie jednej možnosti pred druhou. Radí sa medzi najzložitejšie ľudské činnosti. Uskutočnené rozhodnutia môžu nadobúdať krátkodobý či dlhodobý charakter s rôznorodým zameraním. Stretávame sa s nimi pri vykonávaní každodenných bežných činností, napríklad pri výbere oblečenia do práce, pri nákupe v supermarkete, alebo je to komplexné rozhodovanie na makroekonomickej úrovni so strategickým významom pre spoločnosť.

Ľudské rozhodnutia majú charakter výberu z určitej skupiny možností alebo alternatív, pričom rozlišujeme rozhodovacie situácie na základe počtu existujúcich alternatív. Najjednoduchším prípadom rozhodovacej situácie je voľba z množiny obsahujúcej iba dve alternatívy, tzv. **dichotomický** (binomický) výber. Príkladom dichotomickej premennej môže byť vlastníctvo automobilu (vlastním/nevlastním) alebo voľba z dvoch možností pri nákupe určitého produktu. V prípade existencie troch a viac alternatív sa jedná o **multinomický** (viacnásobný) výber a ako príklad slúži voľba dopravného prostriedku do práce (autobus/metro/auto/pešo), či výber študijného programu pri štúdiu na univerzite. V oboch prípadoch je definovaná voľba z určitej množiny alternatív ako závislá endogénna premennú a výber je definovaný pod spoločným názvom ako diskrétna alebo kvalitatívna voľba.

Na ekonometrickú a štatistickú analýzu kvalitatívnych premenných existuje celý rad postupov, medzi ktoré patrí analýza dvojrozmerných a viacrozmerných kontingenčných tabuliek pomocou loglineárnych modelov, testy nezávislosti alebo miery asociácie. Medzi ďalšie nástroje ekonometrickej analýzy, kde závislá premenná nadobúda diskrétny charakter patria **modely diskrétnej** alebo **kvalitatívnej voľby** lineárneho a najmä nelineárneho charakteru. Uvedené modely nachádzajú široké uplatnenie najmä v oblastiach:

- marketing – základná analýza trhu, výskum spotrebiteľov a ich správania sa, výskum produktov a ich značiek, cenové testy, výskum efektivity predaja a distribúcie, apod.,
- sociológia – výskum verejnej mienky a volebných preferencií, sociálny a politický výskum, analýza výsledkov volieb či referenda, apod.,
- zdravotníctvo – najmä v oblasti epidemiológie (modelovanie vplyvu rôznych faktorov na pravdepodobnosť prežitia jedinca), apod.,
- environmentálna politika – analýzy zdrojov znečistenia, analýzy v rámci urbanistického a územného plánovania,
- doprava – plánovanie zavedenia nového dopravného systému a jeho optimálneho využitia,
- ekonómia – analýzy fungovania firiem a spoločností, napríklad efektivity riadenia výrobných procesov, podnikovej stratégie, pri výbere investičných alebo podielových fondov, bankový sektor – poskytovanie komerčných alebo spotrebiteľských úverov, apod.

Vo všetkých vyššie uvedených oblastiach sa jedná o ekonometrické modelovanie rozhodovacích problémov, kde sa potrebné údaje pre tieto mikroekonomické modely získavajú z výberových šetrení, ankiet spotrebiteľského dopytu, z výskumu verejnej mienky, apod. Keďže správanie sa rozhodujúcich subjektov v rozhodovacích situáciách možno logicky popísať iba pravdepodobnostným spôsobom, tak pomocou modelov diskrétnej či kvalitatívnej voľby je skúmaná spravidla závislosť **podmienenej pravdepodobnosti** na daných vysvetľujúcich premenných. Cieľom týchto modelov je zistiť, v náväznosti na kontrolné ekonomické a iné charakteristiky (atribúty) dotazovaných respondentov, skutočné dôvody a faktory, ktoré ovplyvňujú ich rozhodovanie.

Ciele dizertačnej práce

Predkladaná dizertačná práce má tieto ciele:

- objasnenie mikroekonomických východísk teórie spotrebiteľského správania neoklasickej ekonomickej školy a jeho postupný vývoj v kontexte teórie užitočnosti,
- poskytnutie základného náčrtu histórie modelov diskrétnej voľby a ich najvýznamnejšie aplikácie, ktoré podstatným spôsobom prispeli k ich rozvoju a následnému využívaniu,
- uvedenie prehľadu najpoužívanejších modelov diskrétnej voľby, ich matematickú formuláciu, postupy pri ich odhade a prípadné problémy s nimi spojené,
- aplikácia modelov diskrétnej voľby na reálne údaje, zhodnotenie a jej možné využitie v praxi.

Štruktúra predkladanej dizertačnej práce je nasledovná: v úvodnej prvej kapitole sú popísané základné ciele, predpokladaný prínos dizertačnej práce a stručný prehľad jej obsahu. Druhá kapitola je venovaná teoretickým základom a východiskám modelovania ľudského správania v ekonomickej vede. V krátkosti je popísaný jeho historický vývoj, interdisciplinárny presah a príklady praktických aplikácií a súvislostí. Definícia človeka ako jedinca motivovaného jedine sebeckými záujmami a ziskom („homo oeconomicus“), čiže človeka racionálneho, je často vnímaná odlišne a prevažuje prístup k jeho modifikácii. Napriek tomu si neoklasická koncepcia racionality človeka udržuje v ekonomických a ďalších vedách stále dominantnú pozíciu. Dôležitú podkapitolu tvorí teória náhodnej užitočnosti amerického ekonóma McFaddena, ktorý položil jej základy a rozvinul viacnásobný (multinomický) model diskrétnej voľby. V tretej kapitole sa je uvádzané objasnenie pôvodného postavenia modelov diskrétnej voľby a ich raných aplikácií v ďalších spoločenskovedných odboroch okrem ekonómie, hlavne v biológii, chémii či medicíne. Štvrtá a piata kapitola sú zamerané na popis teoretických základov modelov diskrétnej voľby v závislosti na type premenných, kde rozlišujeme modely binárnej (štvrtá kapitola) a multinomickej voľby (piata kapitola). Vzťah medzi premennými v rámci modelu na druhej strane rozdeľuje modely na lineárne, ktorých hlavným predstaviteľom je lineárny pravdepodobnostný model. I keď uvedený model má niekoľko podstatných nedostatkov, jeho nespornou výhodou je jeho všeobecná aplikovateľnosť a jednoduchá interpretácia i spôsob odhadu. Je však dobrým základom na pochopenie podstaty pravdepodobnostných modelov. Medzi nelineárne modely patrí logitový a probitový pravdepodobnostný model. V krátkosti sú uvedené aj ďalšie modely diskrétnych premenných, ako napríklad zmiešaný logitový model či hniezdový logitový model, ktoré kombinujú vlastnosti základného logitového a probitového modelu. Aplikácie uvedených modelov diskrétnej voľby v oblasti vybavenosti domácností predmetmi dlhodobej spotreby a v oblasti kreditného skórovania hodnotenia klientov sú obsiahnuté v záverečnej šiestej kapitole.

Zdrojom poskytnutých dát bol Český statistický úřad, ktorý v roku 2001 uskutočnil výberové šetrenie domácností *Sociální situace domácností 2001*. K dispozícii boli údaje z 10599 domácností, ktoré obsahovali základné socioekonomické charakteristiky o jej členoch, akými boli napríklad vybavenosť domácností predmetmi dlhodobej spotreby, údaje o bývaní, príjme, počte členov domácností, a pod. V tejto aplikácii analyzujeme a predikujeme modely binárnej voľby.

Skórovanie kreditných rizík u individuálneho subjektu je druhou aplikáciou, kde boli použité modely multinomickej voľby aplikované na reálne dáta.

Súčasť dizertačnej práce tvorí zoznam použitej literatúry vrátane odkazov v texte a zdroje dát. V prílohe je uvedený zoznam vedeckých časopisov, ktoré publikujú online v databáze vedeckých časopisov *JSTOR*, ukážka modelu v programe *SAS Enterprise Miner*, zoznam tabuliek a grafov a výstupy odhadnutých modelov z programu *SAS*.

Na ekonometrickú analýzu a odhady modelov diskrétnej voľby boli použité programy *EViews*, *SPSS, SAS, Stata*, *GiveWin* s modulmi *PcGive* a *DCM*.

2 TEÓRIA SPOTREBITEĽSKÉHO SPRÁVANIA

V tejto kapitole je uvedený prehľad teoretických základov modelovania diskrétnej voľby. Sú tu uplatňované tu dva prístupy, prvým je stručný pohľad do histórie ekonomického prúdu neoklasickej ekonómie a následne teórie užitočnosti, ktorá tvorila východisko a významne ovplyvnila formovanie ekonometrických modelov diskrétnej voľby. V druhom prístupe je uvedený nemenej dôležitý vývoj a využitie modelov diskrétnej voľby, najmä logistickej funkcie, v ostatných behaviorálnych vedách.

2.1 Neoklasická ekonómia a teória užitočnosti

Hlavný prúd ekonomickej školy v období od konca 19. storočia až do 30-tych rokov 20. storočia tvorila klasická a neskôr neoklasická ekonómia. Klasická ekonómia anglických ekonómov Smitha, Pettyho a Ricarda tvorí základ moderných optimalizačných modelov. Zaoberala sa predovšetkým makroekonómiou, presadzovala hospodársky liberalizmus a nezasahovanie do ekonomiky zo strany štátu. Charakter spoločnosti videla v prirodzenej povahe človeka. Vo svojom diele z roku 1776 „Pojednanie o podstate a pôvode bohatstva národov“ vychádza Smith z etiky vlastného prospechu, ktorú považuje za najsilnejšiu ľudskú vlastnosť, pričom každý jedinec sa riadi touto vlastnosťou a je automaticky vedený „neviditeľnou rukou trhu“ ku prospechu celej spoločnosti. Za nemenej podstatný prínos klasickej ekonómie sa považuje použitie metódy dedukcie ako hlavnej metódy ekonomickej analýzy, pomocou ktorej sú na základe východiskových predpokladov konštruované mikroekonomické a makroekonomické modely v dnešnej podobe.

Neoklasickou školou sa rozdelila dovtedy jednotná mikroekonómia a makroekonómia, vyčlenila sa jej mikroekonomická časť. Znamenalo to, že neoklasická ekonómia, alebo ekonómia hlavného prúdu, sa na rozdiel od klasickej ekonómie sústredila na stranu dopytu, čiže na stranu spotrebiteľov. Už z jej názvu **neoklasická** vyplýva, že uznáva úlohu trhu klasickej školy, avšak trh rozoberá v podobe trhového modelu. Jej snahou bola matematizácia ekonomických vied, dovtedy dedukovaných iba slovne.

2.1.1 Racionalita a teória užitočnosti

Základným predpokladom neoklasickej ekonómie je predpoklad **racionality** (ratio – úsudok, rozum) ľudského správania sa jednotlivých aktérov (spotrebiteľa, firmy, vlády, a pod.) rozhodovacieho procesu. Racionalitu by sme mohli definovať ako schopnosť jedinca voliť také prostriedky, ktoré vedú k vytýčeným cieľom, napríklad k uspokojovaniu potreby alebo potrieb jedinca. Racionalitu pritom nemožno hodnotiť podľa zvolených cieľov, ale podľa toho, či zvolené prostriedky k týmto cieľom smerujú. Medzi ďalšie aspekty racionality v ekonomickej vede patrí jej konzistentnosť so sústavou pravidiel vymedzujúcich preferencie jednotlivca predstavujúce jeho potreby, logika v jeho správaní, kde jedinec neustále bilancuje a hodnotí pôžitky z uspokojovania potrieb a nákladov na toto uspokojovanie, egoizmus a individualizmus, ako aj dokonalá informovanosť o všetkých variantách alebo alternatívach, ktoré jedinec optimalizuje, či univerzálnosť maximalizačného princípu. Z toho vyplýva, že jedinec, tzv. „homo oeconomicus“ alebo ekonomický človek, predstavuje bytosť správajúcu sa na základe princípov racionality a optimalizácie jeho správania. Samotný pojem „homo oeconomicus“ je pripisovaný anglickému predstaviteľovi neoklasickej ekonómie Jevonsovi.

Uvedená neoklasická koncepcia ekonomického človeka vychádza z jednej z podôb klasického liberalizmu - z **utilitarizmu** (z lat. utilis = užitočný). Medzi jeho hlavných predstaviteľov patrili Bentham, Mill alebo Sidgwick a ich snahou bolo stanovenie objektívnej zásady, podľa ktorej by bolo určované správne alebo nesprávne správanie sa jedinca v určitých situáciách. Uvedené pravidlo nazvali princípom užitočnosti, kde správanie sa jedinca je správne v prípade, že prináša čo najväčšie šťastie čo najväčšiemu počtu ľudí. Ďalšími filozofickým smermi ovplyvňujúcimi neoklasický prístup sú **hedonizmus** (Epikuros, Diderot), kde hlavným cieľom a motívom človeka má byť dosiahnutie slasti a vyhnutie sa utrpeniu, či **senzualizmus** (Locke, Condillac), podľa ktorého jediným prameňom poznania sú zmyslové javy, pocity a vnemy a úloha rozumu spočíva iba vo formálnom usporiadaní tohto zmyslového materiálu. K rozpracovaniu koncepcie ekonomického človeka významne prispeli aj predstavitelia marginalizmu (Menger, Walras, Pareto), ktorí ho doplnili o psychické zdroje iracionality v podobe reziduí.

V teórii užitočnosti predstavuje významný rozdiel v prístupe chápania užitočnosti z hľadiska jej merateľnosti. Rozlišujeme preto kardinalistický a ordinalistický prístup k užitočnosti. **Kardinalistický prístup** k užitočnosti je starším a pôvodným prístupom, ktorý vznikol v priebehu 19. st. Medzi hlavných predstaviteľov tohto prúdu patrí Menger, Marshall, Walras, Jevons a ďalší. Jeho predpokladom je chápanie užitočnosti ako psychologickej kategórie merateľnej pomocou určitých nástrojov (napr. elektroencefalografom), ktoré umožnia kvantifikovať nielen celkovú, ale dokonca aj hraničnú užitočnosť, ktorú jednotlivec dosiahol. Kardinalistický prístup k užitočnosti je charakterizovaný prívlastkom subjektívny, čiže užitočnosť daného jedinca je merateľná iba v rámci jednej osoby a nie je porovnateľná s užitočnosťou inej osoby. Na druhej strane alternatívny prístup umožňuje interpersonálne porovnanie užitočnosti.

Novším, moderným prístupom, je **ordinalistická teória** užitočnosti, zastúpená menami ako napríklad Pareto, Edgeworth, Slutsky alebo Hicks. Teória predpokladá, že užitočnosť je subjektívna a nemerateľná. Jedinec je schopný určiť iba poradie dôležitosti jednotlivých produktov, čiže vie porovnať ich užitočnosť zo spotreby medzi sebou. Táto teória užitočnosti je nazývaná aj **teóriou preferencií**. V ďalšom texte sa je podrobnejšie rozpracovaná problematika množiny prípustných spotrebných stratégií jednotlivca a jeho preferencií pri výbere z tejto množiny.

Alternatívy ku koncepcii racionality

Mimo ekonómiu hlavného prúdu vznikali pochybnosti o existencii racionality ekonomického subjektu vychádzajúce z empirických skúseností, medzi ktoré patrí princíp satisfakcie a kognitívnej nedokonalosti (napr. Hlaváček a kol., 1999).

Princíp satisfakcie (princíp obmedzenej racionality)

Tento princíp predpokladá, že subjekt v rozhodovacom procese nehľadá optimálnu variantu donekonečna, proces hľadania sa zastaví v momente nájdenia tzv. uspokojujúceho riešenia. Hlavným predstaviteľom tejto myšlienky je americký psychológ Simon, ktorý vo svojom výskume poukázal na to, že predpoklad maximalizácie užitočnosti subjektu pri jeho správaní je v rozpore s faktami, ktoré sú o samotnom ľudskom správaní známe. Pokúsil sa o dôkaz toho, že žiadny subjekt nemôže maximalizovať svoju užitočnosť bez toho, aby ju zároveň neporovnal s ostatnými alternatívami. Znamená to, že ak nebolo uskutočnené toto porovnanie, nemožno tvrdiť o maximalizácii čohokoľvek. Pritom pri rozhodovacom procese používa postup pomocou ktorého hľadá tzv. „uspokojujúce" riešenie alebo rozhodnutie. Tento postup tvorí jadro princípu satisfakcie, ktorý neskôr postupne rozpracovali napríklad autori a tvorcovia behaviorálnej teórie správania firmy Cyert a March.

Koncepcia tzv. kognitívnej nedokonalosti

Vychádza z toho, že racionalita subjektu zlyháva a niektoré subjekty pri zohľadňovaní skúseností z minulosti zahŕňajú do svojich rozhodovacích procesov omyly, chyby a skreslenie systematicky. V podstate táto koncepcia predstavuje popis dôsledkov ľudskej nedokonalosti pri rozhodovaní.

Obe uvedené alternatívy však nepopierajú štandardný koncept ekonomickej racionality ekonomického človeka, ide skôr o jeho zovšeobecnenie a priblíženie k realite.

2.1.2 Preferencie spotrebiteľa

Spotrebiteľ je osoba, ktorá niečo spotrebováva, teda užíva konečné výrobky alebo služby na svoju priamu osobnú spotrebu. Pri výbere alebo voľbe medzi konečnými výrobkami sa spotrebiteľ riadi svojimi preferenciami. **Preferencia** predstavuje uprednostnenie jednej alternatívy alebo varianty pred druhou alternatívou alebo ostatnými variantmi v danej množine. Preferencie sú považované za exogénny parameter správania sa ekonomických subjektov a sú nemenné v zmysle stability voči ekonomickému priestoru rozhodovania a bez ohľadu na ich možné zmeny mimo systém. Na druhej strane ich však napr. Becker chápe ako endogénne a nemenné parametre v zmysle stability vo vzťahu k udalostiam systému, čo spôsobuje problémy pri zvyku, návyku alebo vplyvu reklamy na jednotlivca.

Neoklasická teória predpokladá, že každý jednotlivec je schopný porovnať dve alternatívy (alebo spotrebné koše statkov) A a B z danej množiny alternatív. Medzi vlastnosti preferencií alebo axiómy správania sa racionálneho subjektu patrí:

- **úplnosť –** pre ľubovoľnú dvojicu alternatív A, B platí

  $$A \succ B \quad \cup \quad A \prec B \quad \cup \quad A \approx B,$$

  čiže spotrebiteľ je schopný porovnať ktorékoľvek dva dostupné spotrebné koše statkov. Znaky $\succ$ a $\prec$ predstavujú preferenciu jedného statku pred druhým a znak $\approx$ predstavuje indiferenciu jedného statku ku druhému.

- **reflexívnosť –** pre dva identické spotrebné koše A platí

  $A \geq A,$

čo znamená, že spotrebný kôš *A* má rovnakú alebo vyššiu užitočnosť ako on sám (z hľadiska užitočnosti je indiferentný). Uvedená axióma vlastne predstavuje matematickú podmienku existencie funkcie užitočnosti.

- **tranzitívnosť –** ak existujú tri alternatívy (spotrebné koše) *A*, *B* a *C*, tak platí

  $A\ \mathbf{f}\ B,\quad B\ \mathbf{f}\ C \quad \Rightarrow A\ \mathbf{f}\ C$

  $A \approx B,\quad A \approx C \quad \Rightarrow A \approx C\,,$

  axióma tranzitívnosti predpokladá, že indiferenčné krivky jedného racionálneho spotrebiteľa sa nepretínajú.

Po splnení uvedených troch axióm je možné usporiadať preferencie spotrebiteľa. Avšak bez existencie funkcie užitočnosti spĺňajú uvedené tri vlastnosti tzv. lexikografické preferencie (napr. Soukup, 2001).

- **spojitosť** – spotrebiteľ požaduje zvýšenie spotreby statku *B* pri ľubovoľne malom znížení spotreby statku *A*. Tento predpoklad nám zaistí spojitosť účelovej funkcie užitočnosti.

Po splnení uvedených štyroch axióm je možné matematicky vyjadriť usporiadané preferencie spotrebiteľa pomocou funkcie užitočnosti.

Funkcia užitočnosti reprezentuje usporiadanie jednotlivých spotrebných košov, čiže je to ordinálna funkcia. Fakt, že existuje, však nezaisťuje nájdenie jediného riešenia pri hľadaní jej optima. To nám zabezpečia nasledujúce vlastnosti:

- **lokálna nesaturovateľnosť –** nazývaná aj tzv. **silná monotónnosť –** ak máme dva spotrebné koše, z ktorých každý obsahuje dva statky: $A = (x_0, y_0)$, $B = (x_1, y_1)$, bude $A > B$ ak platí:

  buď: $x_0 > x_1$ a zároveň $y_0 \geq y_1$,

  alebo: $y_0 > y_1$ a zároveň $x_0 \geq x_1$.

  Táto vlastnosť vylučuje existenciu statkov s negatívnymi preferenciami a indiferenčné krivky majú zápornú smernicu.

- **rýdzokonvexnosť** – preferencia priemeru pred extrémami zaisťuje rýdzokonvexný tvar indiferenčných kriviek

  $U[tA + (1 - t)B] > u(A) + u(B)$, pre $t \in (0, 1)$.

2.1.3 Historická koncepcia teórie užitočnosti

Teória racionálnej voľby a s ňou spojená koncepcia teórie užitočnosti, zastúpená okrem iných von Neumannom, Morgensternom, Arrowom a Savageom, sa datuje k 50-tym rokom 20. st. Avšak jej korene siahajú do roku 1738, kedy Bernoulli predložil **teóriu očakávanej užitočnosti** (morálneho očakávania) ako základ rozhodovacieho procesu v podmienkach rizika s použitím logaritmickej funkcie užitočnosti vyriešením problému Petrohradského paradoxu. Podstatou jeho riešenia je nahradenie očakávanej hodnoty výhry hodnotou užitočnosti (napr. Sirůček,1997).

Podľa Benthama, anglického predstaviteľa **utilitarizmu**, jediným meradlom každého ľudského konania je jeho užitočnosť (princíp užitočnosti). Všetci sú povinní sledovať vlastné zámery a uplatnenie jednotlivca tvorí zároveň aj záujem celej spoločnosti. Jeho analýza je však považovaná skôr za kvalitatívnu ako za kvantitatívnu. Napriek tomu jeho výskum podstatne ovplyvnil výskum ďalších utilitaristov, ako napríklad Milla.

Obdobie tzv. **marginalistickej revolúcie** (Walras, Jevons, Menger, Edgeworth) v 70-tych rokoch 19. st. je charakteristické vyjadrením subjektívnej teórie hodnoty pomocou funkcie užitočnosti, ktorá potom vytvorila základ pre analýzu výmeny a produkcie a koncept rovnováhy medzi užitočnosťami sa stal základom pre ekonómov. Cournot bol prvým, ktorý určil dopyt ako funkciu jeho vlastnej ceny.

Na začiatku 20. st. predstavitelia **ordinalistického prístupu** k teórii užitočnosti, Pareto, Fisher a Slutsky, odvodili dopyt spotrebiteľa z maximalizácie funkcie užitočnosti a úplne popísali vlastnosti dopytovej funkcie. Jednalo sa o tzv. axiomatický prístup k teórii užitočnosti za určitosti. Ramsey načrtol obrysy **teórie individuálnej očakávanej užitočnosti** a Finetti položil základy axiomatickej teórie subjektívnej pravdepodobnosti. Významným prínosom k rozvinutiu teórie užitočnosti bola matematizácia teórie spotrebiteľa ekonómov Hicksa a Samuelsona v 30-tych rokoch minulého storočia. Hicksova teória ceny je považovaná za vrchol ordinalistického prístupu v teórii užitočnosti. Na základe maximalizácie užitočnosti spotrebiteľa je možné odvodiť veľa záverov o jeho správaní. Samotná funkcia užitočnosti slúži spotrebiteľovi len na usporiadanie statkov podľa jeho preferencií. Za predpokladu, že pre túto funkciu existuje jedno maximum a je dvakrát diferencovateľná, je možné odvodiť reakciu spotrebiteľa na zmenu cien a jeho správanie vzhľadom k substitučným a komplementárnym statkom.

Prvé pochybnosti do racionálneho modelu ekonomického človeka vniesla teória hier Morgensterna a von Neumanna okolo roku 1940. Autori tejto teórie zdôraznili spoločenské pozadie rozhodovania jednotlivca a túto úvahu ďalej podrobne matematicky rozpracovali. Klasickým príkladom danej teórie je tzv. **väzňova dilema**. Človek sa nerozhoduje len podľa svojich preferencií a svojho rozpočtu, ale berie do úvahy aj vplyv ostatných rozhodujúcich sa subjektov.

Približne o desať rokov neskôr Savage spojil **model očakávanej užitočnosti** Morgensterna a von Neumanna a **model subjektívnej pravdepodobnosti** so základmi Bayesovskej štatistiky a analýzy rozhodovania. V tomto období Arrow a Debreu uverejnili alternatívny prístup v podobe modelovania rozhodovania subjektu v podmienkach neurčitosti alebo rizika. Týmto prístupom bola aplikácia teórie spotrebiteľa na množinu tovarov (predmetov spotreby), podľa ktorého spotrebitelia maximalizujú svoju užitočnosť na množine rozpočtových ohraničení a dopyt po týchto statkoch sa dá kryť jeho ponukou.

V roku 1952 na konferencii o rozhodovaní v podmienkach rizika v Paríži vystúpil Allais a rozdal jej účastníkom dotazníky, kde si mali z každej dvojice peňažných stávok vybrať tú výhodnejšiu. Ak by mal výber určitej stávky maximalizovať užitočnosť, musia byť splnené určité podmienky. Tieto podmienky však boli porušené mnohými účastníkmi konferencie, mimo inými aj Savageom a Arrowom. Svojimi dotazníkmi a príkladmi Allais narušil vtedajšiu dôveru v realistickosť ekonomických modelov a stali sa podnetom rozvoja ďalších nových pohľadov na problematiku rozhodovania. Dokázal, že jedna z axióm správania sa je porušovaná jedincami, ktorí disponujú informáciami. Príklady, ktoré Allais na konferencii uviedol sú známe pod názvom **Allaisov paradox** a poukazujú na význam psychologického aspektu pri rozhodovaní v podmienkach rizika.

V rokoch 1960-70 je naďalej podrobne rozpracovávaná teória subjektívne očakávanej užitočnosti, teória stanovených preferencií a aplikujú sa na problémy Bayesovskej štatistiky, analýzy rozhodovania, rovnováhy na trhu v podmienkach rizika, apod. Okrem iného boli odvodené aj teoretické vlastnosti parametrických funkcií užitočnosti (logaritmická, mocninová, exponenciálna). V tomto období tvorí veľmi významný príspevok do konceptu teórie užitočnosti práca amerického ekonóma McFaddena, ktorého **modelu náhodnej užitočnosti** venujeme osobitnú kapitolu.

V 80-tych rokoch minulého storočia sa množia úvahy týkajúce sa **teórie neočakávanej užitočnosti** v reakcii na trvalé výzvy paradoxov ekonómov Allaisa a Ellsberga a odchýlky v preferenciách v rozhodovacej teórii správania jedinca. Vzniká záujem o oddelenie rizika a časových preferencií, skúmanie vlastností váženej pravdepodobnostnej funkcie a kladú sa nové otázky v problematike oddelenia užitočnosti od pravdepodobnosti.

Začiatkom 21. st. vzniká nová veda, tzv. **neuroekonómia**, ktorá skúma mentálne neurologické procesy na pozadí mikroekonomických rozhodnutí jednotlivca týkajúcich sa napríklad spotreby, úspor či investícií. Vychádza z poznatku, že spotrebitelia väčšinou prijímajú u krátkodobých cieľov iracionálne rozhodnutia a u dlhodobých sa naopak správajú s rozvahou.

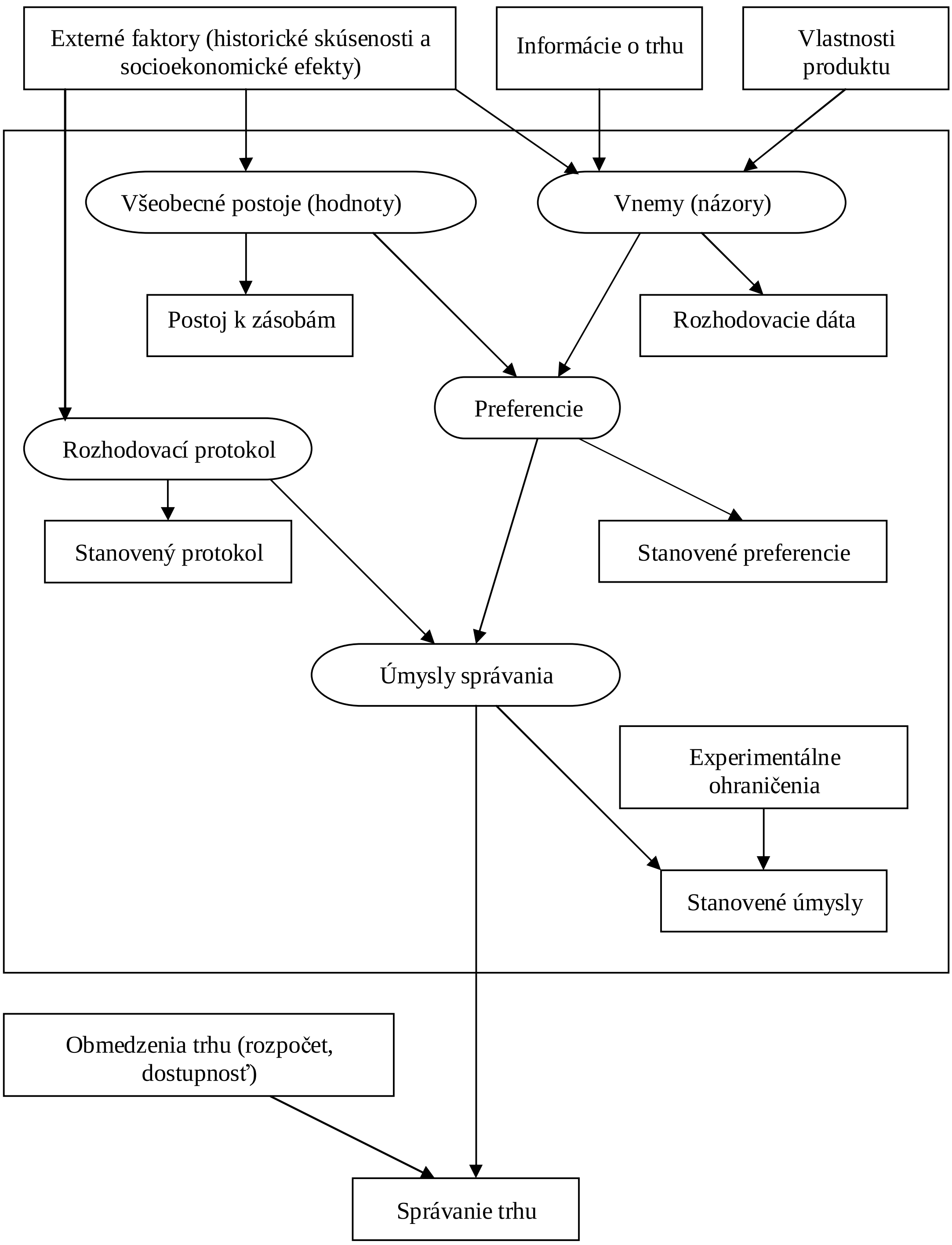
2.2 Model náhodnej užitočnosti

Ako sme uviedli vyššie, dôležitým medzníkom vo vývoji teórie náhodnej užitočnosti je **teória náhodnej užitočnosti** amerického ekonóma a nositeľa Nobelovej ceny za ekonómiu McFaddena, v ktorej rozpracoval a aplikoval pravdepodobnostné modely diskrétnej voľby. Preferencie môžu obsahovať náhodné zložky, ktorými sú očakávania, postoje (stanoviská) a iné nemerateľné faktory. Preferencie sú definované pomocou produktov s celým komplexom merateľných a nemerateľných vlastností. Zvyk a skúsenosti vstupujú do rozhodovacieho procesu jednotlivca z minulosti, čiže z minulých rozhodnutí. Postoje (stanoviská) do rozhodovacieho procesu vstupujú ako dané premenné, avšak za podmienky, že do modelu je tiež zahrnutý spôsob, akým ovplyvňujú trh. Preferencie môžu ďalej modifikovať demografické, ekonomické a sociálne premenné. Táto teória sa stáva použiteľnou prepojením náhodného preferenčného modelu s pravdepodobnosťami odoziev (ohlasov) trhu, ktoré sa potom používajú na podrobnejšie odhady a analýzu. Tento prístup je založený na dvoch podmienkach:

- uvedené pravdepodobnosti odoziev sú podmienené jedine priamo alebo nepriamo merateľnými premennými,
- pravdepodobnosti môžu byť potvrdené voči aktuálnemu správaniu sa trhu.

McFadden rozpracoval komplexný rozhodovací proces spotrebiteľa ako ho popisuje obrázok 2.1. Zvláštnosti v správaní môžu byť kompletne vysvetlené preferenciami, obsahujúce efekty vnemov a postojov jednotlivých spotrebiteľov.

Obrázok 2.1 Grafické znázornenie rozhodovacieho procesu spotrebiteľa (McFadden, 1986)



V uvedenom diagrame McFadden (1986) popisuje komplexný proces rozhodovania sa jednotlivých subjektov (subjektu) na trhu. Výrazy v elipse predstavujú teoretické alebo latentné premenné, výrazy v obdĺžnikoch sú premenné pozorovateľné priamo alebo merané vhodnými experimentmi.

Merateľnými vstupmi v rozhodovacom procese sú vlastnosti produktu, informácie o trhu, historické skúsenosti, socio-ekonomické faktory a ohraničenia trhu, zahrňujúce rozpočet a dostupnosť výrobku. Priamo merateľným výstupom celého procesu je správanie sa trhu, čiže informácie o kúpe daného produktu, zmena značky produktu, apod. Kritickou časťou pri modelovaní kognitívneho rozhodovacieho procesu sú vnemy alebo názory týkajúce sa produktu, všeobecné postoje alebo hodnoty, preferencie medzi produktmi, rozhodovací protokol, ktorý pokrýva preferencie medzi jednotlivými voľbami a úmysly správania sa pri rozhodovaní. Predpokladajme, že kupujúci automobilu má určitú predstavu o trvanlivosti medzi jednotlivými značkami, preferuje špecifické značky alebo modely, charakterizujú ho určité úmysly správania pri výbere špecifickej značky. Jeho vnemy alebo názory sú ovplyvnené vlastnosťami produktu a informáciami o trhu, ďalej vnemy, postoje (stanoviská) a rozhodovací protokol sú ovplyvnené historickými skúsenosťami a socio-ekonomickými faktormi. Všeobecné postoje (hodnoty) a vnemy spolu ovplyvňujú preferencie a preferencie sú transformované pomocou rozhodovacích protokolov do úmyslov správania jednotlivca.

Uvedený pohľad na užitočnosť ako na virtuálne synonymum výberového správania s podmienkami rozmarov, chvíľkových nálad a vnemov jedinca, dominoval v ekonomickej teórii až do 30-tych rokov 20. storočia, kým tento problém podrobne matematicky nerozpracovali Hicks a Samuelson. Spotrebiteľ sa stal optimalizujúcim nástrojom a užitočnosť bola považovaná za fixnú. Ich teória poskytovala základ štúdia ekonomických veličín, akými sú napríklad príjmová elasticita dopytu po tovaroch či efektívna strata spôsobená spotrebnými daňami.

V roku 1965 McFadden, na základe Luceho axiómy, uviedol ekonometrický model, ktorý kombinoval hedonistické ocenenie alternatív a maximalizáciu náhodnej užitočnosti. Tento model nazval **multinomický** (podmienený) **logitový model**. V danom modeli sú prepojené vlastnosti ekonomickej spotrebiteľskej teórie a stochastické prvky, ktoré sú prispôsobené psychofyzikálnemu správaniu spotrebiteľa. Warner (1962) po prvýkrát aplikoval binomický logitový model v oblasti dopravy. Quandt (1968) predstavil model (random parameters travel demand model), ktorý vychádza z teórie náhodnej užitočnosti.

Psychometrické (psychometria – náuka o meraní psychologických javov) metódy poskytujú pohľad na tieto teoretické koncepty. Multidimenzionálne škálovanie rozhodovacích dát prináša indikátory vnemov. Ako príklad nám slúži marketingový výskumník, ktorý chce zistiť aké kritéria (dimenzie) spôsobujú rozdiely vo vnímaní rôznych produktov alebo značiek. Pri svojom výskume vychádza z matice, ktorá obsahuje priemerné rozdiely všetkých dvojíc produktov. Výstup analýzy (multidimenzionálna škálovacia mapa) bude použitý na interpretáciu dimenzií a na identifikáciu medzier na trhu, ktoré nie sú obsadené žiadnym iným existujúcim produktom, ale môžu byť pre spotrebiteľov príťažlivé. **Faktorová analýza** ako výskumná technika, ktorá redukuje veľký počet premenných (napr. 200 otázok v dotazníku) na malý počet premenných, tzv. faktorov (napr. 3) pri malej strate informácií. Predpokladáme, že každú vstupujúcu premennú môžeme vyjadriť ako lineárnu kombináciu malého počtu spoločných skrytých faktorov a jediného chybového faktora. Pomocou tejto techniky sa snažíme vysvetliť závislosť medzi premennými.

Ďalšou z techník je tzv. **„conjoint“** analýza, ktorá rieši problémy experimentálneho párového porovnávania produktov, ich triedenie a zisťovanie, ktoré vlastnosti výrobkov, alebo ktoré ich kombinácie, najviac ovplyvňujú preferencie spotrebiteľov (spotrebitelia každému produktu priradia poradovú hodnotu). Okrem toho sleduje mieru vplyvu premenných popisujúcich produkt, ich jednotlivých kategórií a ich kombinácií na preferencie.

Problémom teda je použitie psychometrických údajov, kvantifikácia premenných z obrázku č. 2.1, simulácia rozhodovacieho procesu spotrebiteľa a predikcia správania sa na trhu spolu s novými vlastnosťami produktu alebo s novými informáciami o trhu.

Psychológia a teória náhodnej užitočnosti

Už v 40. rokoch 20. st. významný americký psychológ, ekonóm a nositeľ Nobelovej ceny za rok 1978 Simon vo svojom výskume poukázal na to, že predpoklad maximalizácie užitočnosti subjektu pri jeho správaní je v rozpore s faktami, ktoré sú o samotnom ľudskom správaní známe. Pokúsil sa o dôkaz toho, že žiadny subjekt nemôže maximalizovať svoju užitočnosť bez toho, aby ju zároveň neporovnal s ostatnými alternatívami. Znamená to, že ak nebolo uskutočnené toto porovnanie, nemožno tvrdiť o maximalizácii čohokoľvek. Pritom pri rozhodovacom procese používa postup pomocou ktorého hľadá tzv. „uspokojujúce“ riešenie alebo rozhodnutie. Tento postup tvorí jadro princípu satisfakcie, ktorý neskôr postupne rozpracovali napríklad autori a tvorcovia behaviorálnej teórie firmy Cyert a March. Podstatný podiel na súčasnom záujme ekonómie o psychológiu je možné pripísať francúzskemu ekonómovi Allaisovi, ktorý v 50. rokoch prispel k vývoju teórie individuálnej voľby v podmienkach rizika.

Psychológovia prejavili ďalší významný záujem o oblasť maximalizácie užitočnosti subjektu až v 50. rokoch, kedy americký psychológ Edwards v roku 1954 publikoval článok o možnostiach, akými ekonómovia vo svojich teóriách a modeloch v tomto období vnímali ľudské myslenie a správanie. Koncom 60. rokov americkí psychológovia Kahneman a Tversky uskutočnili mnoho pokusov a analýz týkajúcich sa uvažovania a rozhodovania ekonomického subjektu v podmienkach neistoty. V nich dokázali jednoduchými prostriedkami navodiť situácie, kde sa dotazovaní subjekti rozhodovali v rozpore so základnými a všeobecne prijímanými zásadami racionálneho správania. V svojom článku (1979) popísali celý rad svojich pokusov a zároveň navrhli nový alternatívny model k vtedajšiemu tradičnému modelu maximalizácie očakávanej užitočnosti. Kahneman za svoj prínos k tejto problematike obdržal v roku 2002 Nobelovu cenu za ekonómiu a to „za integráciu psychologického výskumu do ekonomickej vedy, najmä v súvislosti s ľudským uvažovaním a rozhodovaním v podmienkach neistoty“. Problémom pôsobenia psychológie v ekonómii sa zaoberal aj McFadden, ktorý uverejnil (1999) kognitívne alebo poznávacie efekty rozhodovacieho procesu (tabuľka 2.1.)

Tabuľka 2.1 Kognitívne efekty rozhodovacieho procesu (McFadden, 1999)

Efekt	Popis efektu
KONTEXT	
Ukotvenie	Úsudky sú ovplyvňované kvantitatívnymi podnetmi v otázke rozhodovacieho procesu.
Kontext	Informácie o minulosti a forma prezentácie rozhodovacieho problému zásadne ovplyvňuje konečné rozhodnutie subjektu.
Rámcovosť	Ekvivalentné lotérie prezentované odlišne sú hodnotené odlišne.
Závažnosť	Formát, v ktorom je rozhodovacia úloha stanovená, pôsobí na prikladanie váhy jednotlivým aspektom.
Charakteristika	Subjekty sú nekonzistentné vo výbere a posudzovaní informácií hodnotených ako kľúčové pre rozhodovacie úlohu.
REFERENČNÝ BOD	
Asymetria	Subjekt sa vyznačuje averziou k riziku kvôli zisku, riskuje preferencie kvôli stratám a stratám prikladá väčšiu dôležitosť.
Referenčný bod	Voľba je hodnotená z hľadiska zmien v nadaní alebo v momentálnom stave.
Súčasný stav/talent	Súčasný stav a minulosť podporujú porovnanie alternatív bez skúseností.
DOSTUPNOSŤ	
Dostupnosť	Subjekty nadhodnocujú informácie ľahko vyvolané z pamäti a podhodnocujú informácie o pozadí veci alebo udalosti.
Istota	Istým výsledkom je prikladaná väčšia dôležitosť ako neistým.
Ohnisko	Kvantitatívne informácie sú získané a oznamované v kvalitatívnej forme.
Izolácia	Elementy viacnásobnej alebo viacúrovňovej lotérie sú hodnotené oddelene.
Nadradený/nový	Počiatočné alebo nedávno prežité udalosti sú vyvolané najjednoduchšie.
Regresia	Typické príčiny sú spájané s minulými odchýlkami a vzťah k hlavnému býva podhodnotený.
Reprezentatívnosť	Vysoké podmienené pravdepodobnosti indikujú precenenie nepodmienených pravdepodobností.
Segregácia	Lotérie sú rozdelené na istý výsledok a na risk tohto výsledku.
POVERA	
Dôverčivosť	Dôkaz, ktorý podporuje napodobňovanie alebo ľahkovážne vysvetlenie náhod je príliš jednoducho akceptovateľný.
Oddeľovanie	Subjekty zlyhávajú v uvažovaní alebo akceptovaní logických dôsledkov udalostí.
Povera	Prirodzené štruktúry sa spoliehajú na náhody a kvázimagické sily sú pripisované oponentom.
Podozrievanie	Subjekt neverí výhodným ponukám a spochybňuje oponentove motívy najmä v neznámych situáciách.
PROCES	
Riadenie pravidlom	Správanie je vo väčšej miere vedené princípmi, analógiami a príkladmi, ako užitočnosťou.
Proces	Hodnotenie výsledkov je citlivé na proces a zmenu.
Dočasný	Ignorovanie času je nekonzistentná, krátke oneskorenia sú brané do úvahy relatívne viac ako dlhé.
ODHAD	
Skreslenie	Subjekty môžu skresľovať úsudky o realite alebo o vnímanej strategickej výhode.
Predpoklad	Uvažovanie je zmenené na vnútorne upevnenie alebo predstavené druhým ako sebahodnotenie.

3 HISTORICKÝ VÝVOJ MODELOV DISKRÉTNEJ VOĽBY

Nasledovná kapitola sa zaoberá druhým prístupom k teoretickým východiskám pôvodu modelov diskrétnej voľby a to pôvodom a postupným vývojom logistickej funkcie, v neskoršej úprave logitu a probitu. Logistická funkcia bola pôvodne využívaná v biologických a demografických štúdiách a až oveľa neskôr začala byť všeobecne akceptovaná v ekonomických vedách (Kuznets, 1930).

3.1 Vznik logitu a probitu

Logistická funkcia bola navrhnutá v 19. st. k popisu rastu populácie a postupu autokatalytických alebo reťazových chemických reakcií. V oboch prípadoch je predpoklad vývoja veľkosti populácie $W(t)$ a jej mieru rastu vyjadrenú ako

$$\dot{W}(t) = \mathrm{d}W(t)/\mathrm{d}t. \tag{3.1}$$

Najjednoduchším predpokladom je, že $\dot{W}(t)$ je proporcionálne $W(t)$, nech

$$\dot{W}(t) = \beta W(t), \quad \beta = \dot{W}(t)/W(t), \tag{3.2}$$

kde b je konštantná miera rastu a exponenciálny rast je v tvare

$$W(t) = A \exp bt, \tag{3.3}$$

kde A je možné nahradiť počiatočnou hodnotou $W(0)$. Podľa Malthusa to znamená, že s vývojom veľkosti populácie $W(t)$ sa počet obyvateľstva zvyšuje geometrickým radom. Je to model prijímaný bez akýchkoľvek obmedzení.

Belgický astronóm a štatistik Quetelet, si bol dobre vedomý problémov a prekážok neobmedzeného exponenciálneho rastu počtu obyvateľstva a že to automaticky musí viesť k nezmyselným výsledkom. Sám experimentoval s niekoľkými zmenami (3.2) a požiadal aj svojho študenta Verhulsta, aby sa tým zaoberal.

Matematik Verhulst reagoval na problém pridaním ďalšieho výrazu k (3.2) na vyjadrenie zvyšujúcej sa odolnosti k ďalšiemu rastu a jeho experimentovaniu s rôznymi formami f, t.j.

$$\dot{W}(t) = \beta W(t) - \varphi(W(t)). \tag{3.4}$$

Logistická funkcia vznikne úpravou predchádzajúceho tvaru (3.4)

$$\dot{W}(t) = bW(t)(\Omega - W(t)), \tag{3.5}$$

kde Ω je horná hranica alebo tzv. **hladina nasýtenia** W a jeho asymptota je $t \to \infty$. Rast je teraz proporcionálny nielen k už dosiahnutej veľkosti populácie *W(t)*, ale aj k zvyšnému priestoru pre ďalšiu expanziu v tvare $\Omega - W(t)$.

Ak vyjadríme *W(t)* ako podiel $P(t) = W(t)/\Omega$, tak

$$P(t) = bP(t)\{1 - P(t)\}, \tag{3.6}$$

a riešením tejto diferenciálnej rovnice je výraz

$$P(t) = \frac{\exp(a + bt)}{1 + \exp(a + bt)}, \tag{3.7}$$

ktorý Verhulst nazval **logistickou funkciou.** Veľkosť populácie *W(t)* je tak rovná

$$W(t) = \Omega \frac{\exp(a + bt)}{1 + \exp(a + bt)}. \tag{3.8}$$

Verhulst svoje návrhy publikoval v rokoch 1838 až 1845. V roku 1845 popísal funkciu a jej vlastnosti a pomenoval ju logistickou, bez ďalších vysvetlení. Na jednoduchom grafe znázornil logistickú krivku pozdĺž krivky logaritmickej. Verhulst tiež presne určil tri parametre Ω, α a β z (3.8) vytvorením krivky, ktorá prechádzala cez tri pozorované body.

Keďže pracoval len s údajmi za posledných 20 až 30 rokov, aplikovaná metóda sa nezdala byť vierohodnou, čo bolo potvrdené aj výslednými odhadmi limitovanej veľkosti populácie Ω..

Logistickú funkciu v tvare modelu populačného rastu znovu objavili v roku 1920 Pearl a Reed. O práci Verhulsta nevedeli a nezávisle na ňom objavili logistickú krivku funkcie (3.8). Využili ju pri spracovávaní údajov amerického U.S. Cenzusu a znovu vytvorili krivku prechádzajúcu cez tri body. Tá poskytla vhodný odhad počtu obyvateľstva za obdobie 1790 až 1910. Odhad Ω bol 197 miliónov obyvateľov v USA, jeho súčasný stav je ale približne 280 miliónov. Od začiatku sa Pearl so svojimi spolupracovníkmi snažil aplikovať rastovú logistickú krivku na takmer každú žijúcu populáciu od vínnych mušiek cez počet obyvateľov francúzskych kolónií v severnej Afrike až po rast pižmových melónov.

Ďalšou významnou publikáciou je Yule (1925), a jeho *The growth of population and the factors which control it*, v ktorom venoval Verhulstovej práci celý dodatok svojho článku. Je zároveň aj prvým autorom, ktorý oživil názov logistický.

Ako sme naznačili už v úvode, existuje ďalší pôvod logistickej funkcie v biológii, kde bola s obmenami využitá na popis postupu autokatalytických alebo reťazových chemických reakcií. V nich sa produkt v chemickom procese sám o sebe správa ako katalyzátor, zatiaľ čo zásoby v pôvodnom stave zostávajú nezmenené. To prirodzene vedie k diferenciálnej rovnici (3.6) a teda k logistickej funkcii (3.7).

Kritika aplikácií logistických kriviek v chémii je v prácach Reeda a Berksona (1929), kde zároveň citujú prácu nemeckého profesora chémie Ostwalda.

Základná idea logistického rastu je jednoduchou a efektívnou metódou využívanou pri modelovaní rastu počtu obyvateľstva a pri prenikaní nových produktov a technológií na trh. Ak uvedieme ako príklad produktu na trhu mobilné telefóny, tak ich nástup by sme vlastne tiež mohli definovať ako autokatalytický proces, ktorý je rozširovaný medzi mnoho nových produktov a techník v priemysle.

Vznik probitového modelu je zvyčajne pripisovaný Gaddumovi (1933) a Blissovi (1934), ale pri pohľade na historickú časť práce Finneyho (1971) a referencie Gadduma sa ukazuje, že by to bolo príliš jednoduché. Korene metódy a špeciálne transformáciu frekvencií na ekvivalentné normálne odchýlky už publikoval nemecký vedec Fechner (1860).

Bliss (1934) publikoval dve krátke poznámky v americkom vedeckom časopise *Science*, s uvedením termínu **probit**. Ďalej nasledovala séria článkov vysvetľujúcich maximálny pravdepodobnostný odhad probitovej krivky, v jednom prípade uvedené s pomocou Fishera.

Gaddum a Bliss stanovili kritériá odhadu, čo až do 30-tych rokov minulého storočia bola záležitosť len pre prípad numerickej a grafickej úpravy kriviek kategoriálnych dát. V ich skorých prácach na biologických štúdiách obaja autori pevne dodržiavali klasický model, kde sú stimuly pevne stanovené a reakcie na ne sú náhodné z dôvodu variability individuálnej úrovne tolerancie. Bliss uviedol termín **probit** (skratka „probability unit“) namiesto pôvodnej vhodnej miery pre normálne odchýlky, avšak opustil toto v rámci roka v prospech odlišnej definície, ktorá začala byť odvtedy všeobecne akceptovateľná. Pre každú relatívnu frekvenciu f existuje ekvivalentné normálne odchýlenie $\tilde{Z}$ také, že kumulatívne normálne rozdelenie $\tilde{Z}$ je rovné f a $\tilde{Z}$ je riešením rovnice

$$f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\tilde{Z}} exp\left\{-(1/2)u^2\right\} du, \tag{3.9}$$

riešenie ktorej nájdeme v tabuľkách normálneho rozdelenia. Probit f je ekvivalentné normálne odchýlenie $\tilde{Z}$, alebo spočiatku $\tilde{Z}$ vzrástlo po hodnotách 5. To nám zabezpečí, že hodnota probitu je takmer vždy pozitívna a uľahčuje to výpočet. V probitovej metóde probity relatívnych frekvencií alebo pravdepodobností f boli lineárne závislé na stimuloch, prípadne ich logaritmoch.

K akceptovaniu probitovej metódy pomohli články Blissa, ktorý v tejto oblasti pravidelne publikoval až do 50-tych rokov minulého storočia, Finneyho a ďalších. Úplný rozvoj tejto školy sa pravdepodobne zhoduje spolu s prvým vydaním Finneyho monografie z roku 1947.

Bez základnej teórie v biologických štúdiách, sa probitová analýza začala rýchlo používať pre každý vzťah diskrétneho binárneho výsledku k jednému alebo viacerým determinantom.

Prvé aplikácie v oblasti ekonómie a výskumu trhu sa objavili už v 30-tych rokoch (Kuznets) a neskôr v 50-tych rokov minulého storočia, kedy Farrell (1954) aplikuje probitový model pri skúmaní vlastníctva auta s rozdielnym rokom výroby ako funkciu príjmu domácnosti. Adam (1958) odhadoval lognormálne dopytové krivky ochoty spotrebiteľa kúpiť si cigarety s nízkym obsahom nikotínu za rôzne ceny. Klasická monografia autorov Aitchisona a Browna, zaoberajúca sa lognormálnym rozdelením (1957), predstavila probitovú analýzu širokému publiku ekonómov.

Úvodom do problematiky logistickej funkcie ako alternatívy normálnej pravdepodobnostnej funkcie sa ako prvý zaoberá vo svojej práci Berkson (1929), spoluautor Reeda v článku o autokatalytických funkciách. Berkson v roku 1944 sa zameral na štatistickú metodológiu biologických testov a navrhol použitie logistickej funkcie. Zároveň ako prvý použil termín **logit** analogicky k probitu od Blissa. Inverziou logistickej funkcie (3.7) dostávame **logitovú funkciu** v v jednoduchšom tvare v porovnaní s probitovou funkciou

$$logit(P) = log\frac{P}{1-P} = Z, \tag{3.10}$$

Zásadný problém logit vs. probit bol navyše komplikovaný kritikou Berksona, metódy odhadu maximálnej vierohodnosti a jeho obhajobou metódy minimálneho χ^2 odhadu. V rokoch 1944 až 1980 napísal na obe problematiky veľké množstvo článkov zaoberajúcich sa oboma problémami.

Podobnosť logistickej funkcie a funkcie normálneho rozdelenia zmienil vo svojich prácach Wilson (1925) a Winsor (1932). Wilson bol pravdepodobne prvý, kto publikoval aplikáciu logistickej krivky v biologických testoch v článku Wilson a Worcester (1943), krátky čas pred Berksonom (1944). Avšak bol to Berkson, ktorý logistickú krivku obhajoval ešte niekoľko desaťročí.

Berksonove názory neboli dobre prijaté odborníkmi z oblasti biometrie z niekoľkých dôvodov. Logit bol považovaný za niečo menejcenné a pochybné, pretože na rozdiel od probitu nesúvisel so základným (normálnym) rozdelením na tolerančnej úrovni. Aitchison a Brown (1957) vo svojich tvrdeniach vôbec o logite neuvažovali, pretože mu „chýba dobre rozpoznateľné a zvládnuteľné frekvenčné rozdelenie tolerancie, ktoré probitová krivka obsahuje už vo svojom základe“ (str. 72).

Berkson (1951) si uvedený nedostatok dobre uvedomoval a pokúsil sa ho napraviť modifikáciou autokatalytického argumentu.

Z praktickej stránky bola jednoduchosť výpočtu zreteľnou výhodou oproti probitu, navyše s odhadom metódou maximálnej vierohodnosti. Z tohto pohľadu sa logit rozšíril oveľa rýchlejšie v bežných praktikách než v akademických kruhoch. Do nástupu osobných počítačov a kalkulačiek, boli všetky numerické výpočty uskutočňované ručne, niekedy podporované grafickou kontrolou ručne (opäť) kreslených kriviek. Probitová a logitová analýza skupinových dát alebo triedených početností bol logaritmický papier so špeciálnou sieťou súradníc, na ktorom sa logitová a probitová krivka zdala byť priamou čiarou. Logistickú (autokatalytickú) sieť súradníc predstavil Wilson (1925) a príklady lognormálneho papieru môžeme nájsť u Aitchisona a Browna (1957) a Adama (1958). Berkson osobne navrhol logaritmický grafový papier ako aj niekoľko nomogramov.[1]

3.2 Rozšírenie logitu a probitu v spoločenských vedách

Keď ideová diskusia okolo a logitu a probitu v biologických testoch poľavila, okolo roku 1960, logitová terminológia a logitová transformácia (3.10) boli rýchlo rozšírené a ich pôvod zabudnutý. Celú históriu prevzatia a ďalšieho vývoja logitu by vyžadovala detailnú znalosť niekoľkých ďalších odlišných disciplín. V krátkosti načrtneme dôležité momenty v štatistike, v epidemiológii, v sociálnych vedách a v ekonometrii, avšak bez pokusu o systematickú úpravu.

Po prvýkrát boli aplikácie logitu a probitu zaznamenané v 50-tych rokoch minulého storočia v štatistike a epidemiológii. V štatistike boli skoro prijaté analytické výhody logitovej transformácie ako napríklad stredná hodnota predaja s diskrétnymi binárnymi výsledkami. Cox bol medzi prvými, ktorý využil tieto možnosti v sérii článkov okolo roku 1960, nasledované učebnicou z roku 1969. Logitový model v biologických testoch je jednoduché odvodiť z logistickej regresie, kde binárne výsledky súvisia s počtom determinantov bez špecifického teoretického základu a tento štatistický model sa zdá vhodným nástrojom analýzy ako pôvodne lineárna regresia. Neskôr bolo uznané prepojenie logistického modelu s diskriminačnou analýzou a vo všeobecnosti jeho spojitosť s loglineárnymi modelmi. V epidemiológii prípadové štúdie s využitím logitu a probitu začali ešte skôr, a odvtedy sa priamo dotýkali pravdepodobnosti, pomeru pravdepodobností,

[1] Nomogram je grafické vyjadrenie určitého algoritmu, napr. logaritmické pravítko apod.

logaritmického podielu alebo logitovej transformácie. Prax začala hneď volať po teoretickom zdôvodnení, najmä v prípade použitia výberových dát.

Vzostup výskytu výrazov logit a probit v ekonomických a štatistických časopisoch je pre zaujímavosť prezentovaný v tabuľke 3.1, ako zdroj slúžil elektronický archív vedeckých časopisov JSTOR[2].

Tabuľka 3.1 Počet článkov v časopisoch obsahujúcich slovo „logistic“, „logit“ a „probit“

	ekonómia			štatistika		
	logistic	**logit**	**probit**	**logistic**	**logit**	**probit**
1935 – 39	14	–	–	12	2	6
1940 – 44	7	–	–	9	1	3
1945 – 49	7	–	–	22	6	22
1950 – 54	17	–	2	33	15	50
1955 – 59	19	2	4	39	23	55
1960 – 64	25	1	5	66	27	43
1965 – 69	36	4	4	123	41	44
1970 – 74	60	27	26	182	62	47
1975 – 79	103	114	93	223	75	48
1980 – 84	188	250	241	375	153	99
1985 – 89	203	321	342	545	232	110
1990 – 94	237	373	505	764	327	146
1995 – 99	250	387	526	783	376	160

Až do 80-tych rokov 20. storočia bola výpočtová snaha stále dôležitým a zásadným problémom v diskusiách štatistických techník, avšak nástup počítačovej revolúcie to ukončil. V špecifických otázkach odhadu logitu a probitu sa metóda maximálnej vierohodnosti stala štandardom, keď postup tejto metódy aplikovateľný na individuálne dáta, začal byť zahrňovaný v komerčných štatistických programoch. Táto možnosť bola pravdepodobne po prvýkrát ponúknutá v programe BMDP (BioMedical Data Processing) z roku 1977. Časom publikovali Hosmer a Lemeshow (1989) prvú komplexnú učebnicu s medicínskymi aplikáciami, v ktorej bolo uvedené použitie týchto postupov brané ako samozrejmosť.

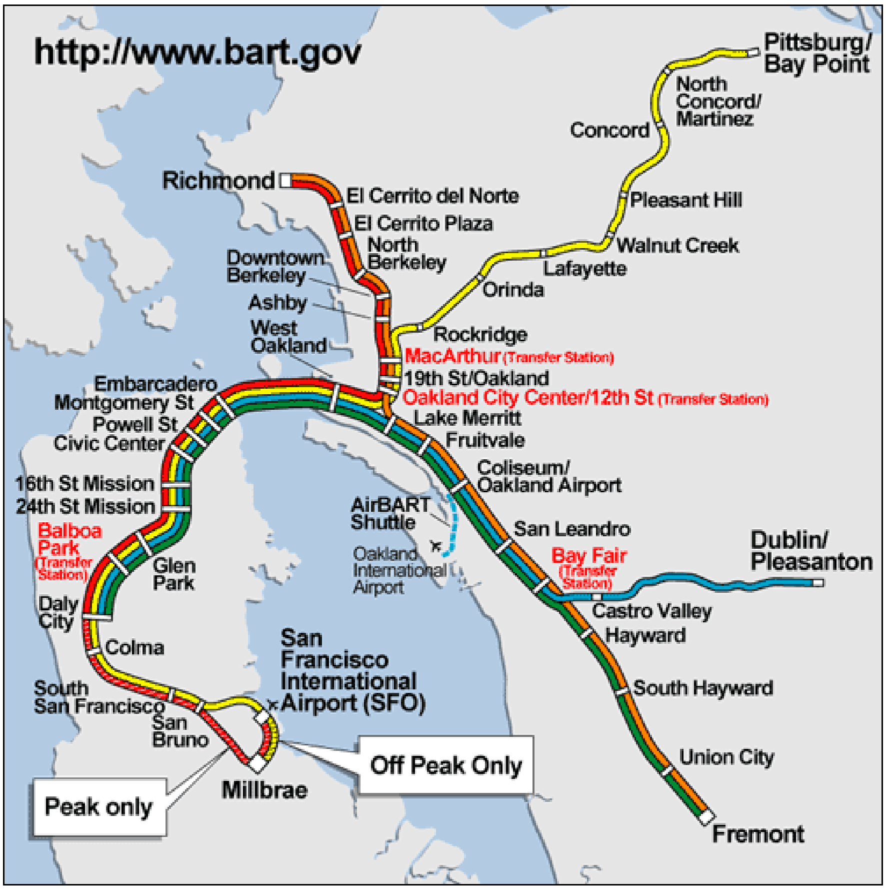
Ako sme naznačili už skôr, probitový model biologických testov bol prevzatý týmito disciplínami. Teoretické ospravedlnenie biologických testov z hľadiska vymedzených stimulov a náhodných prahov bolo prvé upustené v zmene v logistickú regresiu a potom oživené v

[2] JSTOR (The Scholarly Journal Archive). Údaje boli spracované dňa 27. 2. 2004. Keďže väčšina časopisov uverejnená v elektronickej podobe na webových stránkach JSTOR podlieha tzv. „moving wall“ (možnosť uverejnenia po 3 až 5 rokoch), posledné aktuálne údaje sú z roku 1999.

regresnom modeli s latentnými premennými. To sa uskutočnilo pravdepodobne kvôli McKelveymu a Zavoinovi (1975), ktorí to uviedli v usporiadanej probitovej analýze správania sa voličov pri voľbe kongresmanov v USA. Príkladom simultánnych nezávislých objavov je zovšeobecnenie logistickej regresie na prípad multinomický (polytomický), prvýkrát vysvetlené u Gurlanda, Leea a Dahma (1960). O niekoľko rokov neskôr bolo celkom nezávisle predložené štatistikom Cox (1966) a biometrickým štatistikom Mantelom (1966). A znovu o pár rokov neskôr to bolo ešte raz nezávisle znovuobjavené ekonometrom Theilom (1969), ktorý k tomu dospel zo všeobecného pohľadu modelovania akciových trhov.

Dlhý čas boli logistická regresia a modely diskrétnej voľby, či už v binárnom alebo multinomickom kontexte, v prvom rade použité ako technika, bez charakteristickej interpretácie. K zásadnej zmene v ich použití došlo v roku 1973, kedy americký ekonóm McFadden získal grant od spoločnosti *National Science Foundation,* týkajúci sa problematiky aplikácie modelov diskrétnej voľby. Konkrétnou úlohou bola voľba dopravy do práce u jednotlivca. Projekt bol zameraný na využitie dopravného systému v Kalifornii v Bay Area, nazvaného San Francisco Bay Area BART (Bay Area Rapid Transit) uvedenom na obrázku 3.1. Pred uvedením tohto systému do prevádzky boli zozbierané dáta 771 ľudí dochádzajúcich do práce, ako napríklad rodinný príjem, doba trvania cesty do práce autom alebo autobusom, úspora času apod. Modely multinomickej voľby boli potom použité na odhady pravdepodobností, v akej miere bude BART použiteľný. Predikovaný dopyt bol 6,3 percenta, skutočný dopyt, po spustení BART-u bol vypočítaný 6,2 percenta, čo poukazuje na významnosť uplatneného modelu.

Obrázok 3.1 Mapa dopravného systému BART



4 MODELY BINÁRNEJ DISKRÉTNEJ VOĽBY

Štvrtá kapitola je venovaná matematickej formulácii modelov diskrétnej voľby a to najprv modelom s binárnou závislou premennou. Najskôr sú ubjasnené typy premenných používaných v ekonometrickej analýze so zameraním sa na kvalitatívne premenné. Je definovaná množina alternatív spotrebiteľa a na ňu kladené požiadavky. V ďalšej časti sú formulované modely binárnej voľby – lineárny pravdepodobnostný, logitový a probitový model, metódy ich odhadu a problémy interpretácie odhadnutých modelov.

4.1 Typy premenných v ekonometrických modeloch

Pre potreby ekonometrickej analýzy je na začiatku nevyhnutné určiť typ premenných, s ktorými pracujeme. Ich nesprávne určenie vedie k použitiu metódy, ktorá nám spravidla poskytne chybné alebo skreslené výsledky a následne ich nesprávnu interpretáciu.

V literatúre sa stretávame s rozličnými typmi klasifikácie premenných. Podľa jedného z z prístupov, kde je hlavným kritériom **typ vzťahov medzi hodnotami**, sú rozlišované premenné nominálne, ordinálne, intervalové a pomerové (napr. Pecáková a kol., 2004).

- **Nominálna** premenná je klasifikovaná iba kvalitatívne. Znamená to, že jej hodnota buď patrí alebo nepatrí do určitej kategórie, pričom nevieme určiť poradie týchto kategórií. Jednotlivé hodnoty nominálnej premennej sú vyjadrené slovne alebo číselnými kódmi (čo uľahčuje ich ďalšie počítačové spracovanie). Typickou ukážkou je napríklad rodinný stav (slobodný, ženatý, rozvedený, vdovec), miesto narodenia (Bratislava, Nitra, Košice, atď.), národnosť (slovenská, česká, atď.), apod. Špeciálnym prípadom nominálnej premennej je premenná dichotomická, nadobúdajúca iba dve možné hodnoty (pohlavie, fajčiar a nefajčiar, atď.)

- **Ordinálna** (poradová) premenná spĺňa všetky podmienky kladené na nominálnu premennú, a navyše vieme určiť poradie jej hodnôt. Nie je však možné určiť, o koľko je jedna hodnota vyššia alebo nižšia. Zaraďujeme sem napríklad úroveň vzdelania (základné, stredoškolské bez maturity, stredoškolské s maturitou, vysokoškolské), miera spokojnosti zákazníka s určitým výrobkom, apod.

- **Intervalová** (rozdielová) premenná umožňuje vystihnúť nielen poradie, ale kvantifikuje aj rozdiel medzi jej dvoma hodnotami. Neumožňuje však určiť ich pomer, pretože vo svojej škále hodnôt nemá stanovenú tzv. racionálnu nulu. Príkladom je mesačný príjem domácností či hladina cholesterolu v krvi. Ďalším, v literatúre často uvádzaným príkladom, je teplota v stupňoch Celzia (príp. Fahrenheita), kde 0°C neznamená neprítomnosť teploty (teplota 30°C je o 15°C vyššia než teplota 15°C, ale nie je dvakrát vyššia).

- **Pomerová** (podielová) premenná má definovanú racionálnu nulu a preto má zmysel hovoriť o tom, o koľkokrát je jedna hodnota vyššia (nižšia) než druhá. Ak meriame teplotu na Kelvinovej stupnici, tak vieme určiť nielen fakt, že teplota 100 stupňov je o 50 vyššia ako 50 stupňov, ale aj tú skutočnosť, že je to teplota presne dvakrát tak vysoká. Ďalším príkladom môže byť hmotnosť, počet členov domácnosti, vek, apod.

Nominálne a ordinálne premenné sú súhrnne označované ako **kvalitatívne**, intervalové a pomerové premenné sú označované ako **kvantitatívne** (kardinálne, numerické). Kvantitatívne premenné ďalej môžu nadobúdať hodnoty *diskrétne*, čiže iba celočíselné, alebo hodnoty *spojité*, tj. ľubovoľné hodnoty z určitého intervalu.

Pod pojem **kategoriálne** premenné sa zahŕňajú zahrnúť nominálne, ordinálne a kvantitatívne diskrétne premenné (kategória značí obmenu tejto premennej). Podľa počtu obmien možno rozdeliť kategoriálne premenné na dichotomické a multinomické. **Dichotomická** (alternatívna, binárna, nula-jednotková) premenná nadobúda iba dve hodnoty. Najčastejšie sú tieto hodnoty kódované hodnotami 0 a 1 (odtiaľ názov nula-jednotková premenná). Pri posudzovaní dôležitosti týchto dvoch hodnôt, rozlišujeme premenné *symetrické* s rovnakou dôležitosťou kategórií (muž, žena), a *asymetrické* premenné, kde je jedna kategória dôležitejšia (uzdravenie pacienta). **Multinomická** (polytomická, viackategoriálna, množná) premenná nadobúda viac než dve kategórie. Ako príklad môžeme uviesť rodinný stav, typ školy, apod.

Premenné určitého typu je možné previesť na iný typ. Napríklad premenná vek je kvantitatívna spojitá premenná. V prípade jej modifikácie na viacero premenných v tvare vekových kategórií tým vznikajú premenné ordinálne. Tieto ordinálne premenné ďalej môžu byť dichotomické (pri ich rozdelení na mladších a starších) alebo nominálne množné (mládež, seniori, ostatní).

Podľa typu závislých a nezávislých premenných v regresnom modeli Powers (2000) uvádza typológiu regresných modelov a príslušnú metódu analýzy:

Tabuľka 3.2 Typológia regresných modelov

Závislá premenná	**Nezávislá premenná**	**Metóda analýzy**
Spojitá	Spojitá	Regresná analýza, korelačná analýza
Spojitá	Kategoriálna	Regresná analýza, analýza rozptylu
Binárna	Kategoriálna	Logitový/probitový model, loglineárna analýza
Binárna	Spojitá	Logitový/probitový model
Neusporiadaná multinomická	Kategoriálna	Loglineárna analýza, multinomický logitový model
Neusporiadaná multinomická	Spojitá	Multinomický logitový model
Usporiadaná multinomická	Kategoriálna	Usporiadaný logitový/probitový model, loglineárna analýza
Usporiadaná multinomická	Spojitá	Usporiadaný logitový/probitový model
Cross-classified data	Kategoriálna	Loglineárna analýza
Obmedzená	Kategoriálna, spojitá	Loglineárna analýza, Logitový/doplnkový dvojlogaritmický model

Množina alternatív

Výberová situácia, v ktorej sa spotrebiteľ (domácnosť, firma) rozhoduje, je charakterizovaná **množinou alternatív**, ktorá musí spĺňať nasledovné kritériá (McFadden, 1974):

- množina alternatív obsahuje **konečný počet** alternatív,
- alternatívy sú navzájom **nezlúčiteľné** (nekompatibilné) – spotrebiteľova voľba jednej alternatívy implikuje, že si nezvolí žiadnu ďalšiu alternatívu z množiny alternatív,
- množina alternatív je **vyčerpávajúca** (úplná) – sú v nej zahrnuté všetky možné alternatívy a spotrebiteľ si nevyhnutne zvolí len jednu alternatívu z celej množiny.

Prvé dve vlastnosti množiny alternatív nie sú tak reštriktívne, pretože je možné zabezpečiť ich splenenie vhodným definovaním jednotlivých alternatív. Pri predpoklade napríklad dvoch alternatív označených ako A a B, ktoré nie sú navzájom nezlúčiteľné, si subjekt môže zvoliť obe alternatívy. Definovaním voľby: jedine A, jedine B a obe A aj B, túto vlastnosť však zabezpečíme. Podobne množina alternatív nemusí byť vyčerpávajúca alebo úplná, pretože subjekt si nemusí chcieť zvoliť ani jednu z alternatív A a B, takže treba pridať voľbu „žiadna z alternatív A a B“, takže množina alternatív je v tomto prípade vyčerpávajúca.

Okrem nevyhnutného definovania množiny alternatív je potrebné definovať toho, kto sa bude v danej rozhodovacej situácii rozhodovať. Medzi **subjekty rozhodovacieho procesu** môžeme napr. zaradiť:

- jednotlivec (spotrebiteľ) – voľba povolania, študijného odboru na vysokej škole, a pod.
- domácnosť – nákup dovolenky, predmetu dlhodobej spotreby, apod.,
- firma – uzavrieť/neuzavrieť obchod, aký výrobok uviesť na trh, nákup strojového zariadenia.

V ekonometrickej analýze rozhodovania sa subjektu ekonóma zaujíma otázka prečo sa subjekt rozhodol pre danú alternatívu a nie pre inú a aké sú faktory ovplyvňujúce jeho rozhodnutie a celý proces rozhodovania.

4.2 Modely binárnej voľby

Najjednoduchším prípadom modelov diskrétnej voľby sú modely s binárnou závislou premennou. Modely binárnej voľby predpokladajú, že spotrebitelia sa rozhodujú medzi dvoma alternatívami, napríklad či vlastní alebo nevlastní predmet dlhodobej spotreby, plánuje alebo neplánuje dovolenku, je fajčiar alebo nefajčiar apod. O tomto type závislej premennej hovoríme, že je binárna (alternatívna, nula-jednotková, dichotomická). V ekonometrických modeloch túto vlastnosť kódujeme hodnotami 0 a 1.

Pri špecifikácii modelu budeme brať do úvahy prípady, kedy pravdepodobnosť zvolenia konkrétnej alternatívy je buď lineárnou alebo nelineárnou funkciou vysvetľujúcich premenných. Podrobnejšie uvedieme tri základné prístupy modelovania binárnej diskrétnej voľby a to lineárny pravdepodobnostný, nelineárny logitový a probitový model.

V literatúre môžeme nájsť tri prístupy k formulácii modelov diskrétnej voľby:

- prístup z hľadiska **teórie náhodnej užitočnosti** (McFadden, 1974) – prístup založený na maximalizácií užitočnosti jednotlivca,
- **transformačný** (štatistický) prístup (Powers, 2000) – kde uvažujeme so skupinovými dátami, ich frekvenčnými početnosťami a proporciami,
- prístup, založený na prítomnosti tzv. **latentnej premennej** – predpokladá existenciu indexu užitočnosti v podobe latentnej (nemerateľnej) náhodnej premennej.

4.3 Lineárny pravdepodobnostný model

Lineárny pravdepodobnostný model (LPM), ktorý predstavuje najjednoduchší model binárnej voľby, umožňuje pomerne jednoducho aproximovať vysvetlenie a prognózu správania sa ekonomických subjektov. Sú to rozhodovacie situácie, kedy je zisťovaná pravdepodobnosť, že spotrebiteľ, domácnosť, firma alebo všeobecní zastupitelia príjmu či odmietnu konkrétne ekonomické (prípadne iné) alternatívne rozhodnutia. Ako vysvetľujúce faktory, ktoré zvolenie konkrétnej alternatívy z daných dvoch možností ovplyvňujú, sú volené ekonomické, sociálne, demografické a ďalšie charakteristiky.

Prostredníctvom jednoduchého príkladu jednorovnicového lineárneho pravdepodobnostného modelu je vyjadrený regresný vzťah závislosti podmienenej pravdepodobnosti, že domácnosť vlastní/zakúpi alebo nevlastní/nezakúpi predmet dlhodobej spotreby (napríklad dom) v závislosti na výške jej príjmu (Gujarati, 1995). Binárna vysvetľovaná premenná Y_i je charakterizovaná nasledovne

$$Y_i = \begin{cases} 1 & \text{v prípade vlastníctva domu,} \\ 0 & \text{v opačnom prípade,} \end{cases} \qquad i = 1,2,\ldots,n.$$

Prvým prístupom k formulácii lineárneho pravdepodobnostného modelu je prístup z hľadiska **teórie náhodnej užitočnosti** (McFadden, 1974) u individuálnej domácnosti.

Domácnosť maximalizuje svoju užitočnosť spôsobom, že si zvolí alternatívu 1 v tom prípade, ak jej prinesie vyššiu užitočnosť ako alternatíva 0. Ďalej užitočnosť i-tej domácnosti, ktorá si zvolí alternatívu j je označná ako U_{ij}. Potom U_{i1} predstavuje užitočnosť i-tej domácnosti, ktorá vlastní dom a U_{i0} je jej užitočnosť v opačnom prípade

$$Y_i = \begin{cases} 1 & \text{ak } U_{i1} > U_{i0}, \\ 0 & \text{ak } U_{i1} \le U_{i0}. \end{cases} \qquad i = 1,2,\ldots,n. \tag{4.1}$$

Označme hodnoty náhodnej premennej Y_i ako $j = 0$ alebo 1 a pravdepodobnosti binárnej voľby pomocou $P(y_i = j)$. Hodnota y_i predstavuje napozorovanú alebo **výberovú** hodnotu náhodnej premennej Y_i.

Výberová pravdepodobnosť, že diskrétna náhodná premenná Y_i nadobúda hodnotu 1 je rovná $p_i = P(y_i = 1)$ a $1 - p_i$ je pravdepodobnosť opačná, náhodná premenná Y_i je rovná hodnote 0.

Funkcia hustoty pravdepodobnosti diskrétnej náhodnej premennej Y_i má **Bernoulliho rozdelenie** v tvare

$$Y_i \approx f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}, \qquad y_i = 0, 1, \qquad p_i \in \langle 0,1 \rangle. \tag{4.2}$$

Z (4.2) vyplýva, že pravdepodobnosť vlastníctva domu je

$$f(1) = P(y_i = 1) = p_i$$

a opačná pravdepodobnosť v prípade nevlastnenia domu je rovná

$$f(0) = P(y_i = 0) = 1 - p_i.$$

Výsledný tvar jednorovnicového lineárneho stochastického regresného modelu dvoch premenných je zapísaný v tvare

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1,2,...,n, \tag{4.3}$$

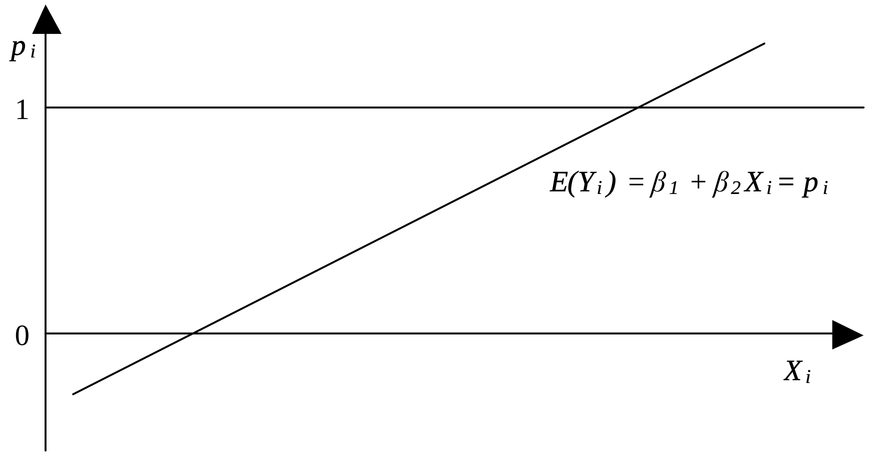
kde

X_i – výška príjmu i-tej domácnosti,

u_i – náhodná zložka modelu.

Priebeh funkcie LPM znázorňuje obrázok 4.1.

Obrázok 4.1 Lineárny pravdepodobnostný model



Strednú hodnotu diskrétnej náhodnej premennej Y_i, čiže podmienené očakávanie alebo **podmienenú pravdepodobnosť** vlastníctva domu v závislosti na výške príjmu ***i***-tej domácnosti, môžeme za predpokladu, že $E(u_i) = 0$, zapísať v tvare

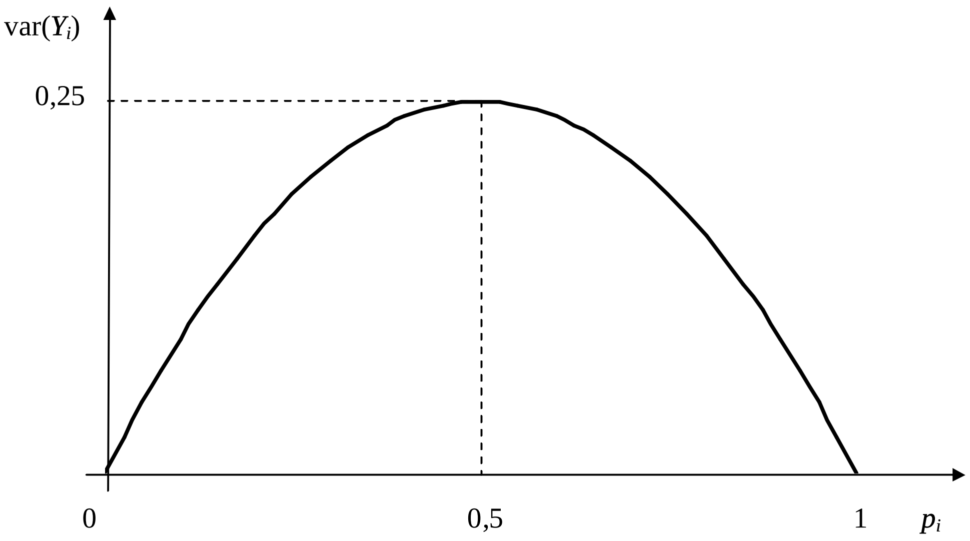
$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$$

Podmienený rozptyl diskrétnej náhodnej premennej Y_i vzhľadom ku konkrétnej kombinácií hodnôt vysvetľujúcich premenných má tvar

$$\text{var}(Y_i) = P(y_i = 1)[1 - P(y_i = 1)] = p_i(1 - p_i)$$

a jej priebeh je znázornený na obrázku 4.2. Je zrejmé, že rozptyl sa mení, pričom najvyšší rozptyl o veľkosti 0,25 dosahuje veličina Y_i pri pravdepodobnosti $p_i = 0{,}5$. Naopak, pri extrémnych hodnotách pravdepodobností $p_i = 0$ alebo $p_i = 1$ je rozptyl nulový.

Obrázok 4.2 Rozptyl veličiny Y_i v LPM



Formulácia LPM v prípade viacnásobnej regresie

Pri predpoklade viacnásobnej regresie v LPM, ako $\mathbf{x}_i$ je označený vektor K vysvetľujúcich premenných. Početnosť tohto vektora v každom z M opakovaní výberového experimentu tvorí veľmi malý rozsah n_i pozorovaní. Potom hodnota Y_i predstavuje počet výskytov skúmanej alternatívy v n_i pozorovaniach, ktoré zodpovedajú jednotlivým hodnotám vektoru vysvetľujúcich premenných $\mathbf{x}_i$. Výberová relatívna početnosť skúmanej alternatívy v i-tom výberovom experimente p_i je potom označená výrazom $p_i = Y_i/n_i$. Ďalej pre skutočný podiel P_i a výberový podiel p_i platí nasledovný vzťah

$$p_i = P_i + u_i \,, \qquad i = 1,2,..., M, \qquad (4.4)$$

alebo

$$E(p_i) = P_i \,.$$

Pri predpoklade lineárnej závislosti P_i na vektore K vysvetľujúcich premenných $\mathbf{x}_i$ môžeme pre základný súbor zapísať

$$P_i = \mathbf{x}_i^T \boldsymbol{\beta}, \qquad i = 1,2,..., M, \tag{4.5}$$

kde **β** je vektor neznámych parametrov o rozmere K×1 a vzhľadom k (4.4)

$$p_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \qquad i = 1,2,..., M. \tag{4.6}$$

Rozptyl má tvar

$$\mathrm{var}(y_i) = n_i p_i (1 - p_i)$$

Maticový zápis zovšeobecneného lineárneho pravdepodobnostného modelu binárnej voľby možno pre prípad viacnásobnej regresie vyjadriť vzťahom

$$\mathbf{p} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\,, \tag{4.7}$$

kde

- **p** – vektor výberových podielov skúmanej alternatívy rozmeru M×1,
- **X** – matica M×K pozorovaní vysvetľujúcich premenných,
- **u** – vektor náhodných zložiek rozmeru M×1.

V uvedenom modeli v dôsledku **heteroskedasticity** náhodných zložiek u_i je získaný odhad parametrického vektoru **β** metódou **zovšeobecnených najmenších štvorcov** pomocou odhadovej funkcie v tvare

$$\mathbf{b}^* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{p}, \tag{4.8}$$

kde diagonálna kovariančná matica binomicky rozdelených náhodných zložiek **V** rozmeru M×M je rovná

$$\mathbf{V} = \begin{bmatrix} \frac{P_1(1-P_1)}{n_1} & . & \dots & 0 \\ . & \frac{P_2(1-P_2)}{n_2} & & . \\ . & . & & . \\ . & . & & . \\ 0 & . & \dots & \frac{P_M(1-P_M)}{n_M} \end{bmatrix}.$$

V prípade, ak sú skutočné podiely P_i neznáme, je možné použiť modifikovanú odhadovú funkciu metódy zovšeobecnených najmenších štvorcov v nasledovnom tvare

$$\hat{\mathbf{b}}^* = (\mathbf{X}^\mathbf{T}\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\hat{\mathbf{V}}^{-1}\mathbf{p}\,, \tag{4.9}$$

kde diagonálne prvky konzistentného odhadu $\hat{V}^{-1}$ sú rovné $\frac{\hat{P}_i(1-\hat{P}_i)}{n_i}$ a hodnoty $\hat{P}_i$ sú variantne získané konzistentné odhady skutočných P_i. Najčastejšie $\hat{P}_i = p_i$, alebo $\hat{P}_i = \mathbf{x}_i^T\mathbf{b}$, pričom odhad parametrov vektoru $\boldsymbol{\beta}$ MNŠ získame ako $\mathbf{b} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{p}$. Je možné použiť aj vzťah $\hat{\boldsymbol{P}}_i = \boldsymbol{x}_i^T\boldsymbol{b}^*$, ktorý je založený na odhade parametrov vektoru $\boldsymbol{\beta}$ metódou zovšeobecnených najmenších štvorcov.

Ďalšími metódami odhadu matice **V** je tzv. dvojstupňová MNŠ alebo Newton-Raphsonova iteračná metóda.

Dvojstupňová MNŠ

Metóda je založená na opakovanej aplikácii klasickej MNŠ (napr. Hušek, 2003). V prvej fázi odhadneme regresné parametre (označme ich $\mathbf{b_1}$)

$$\mathbf{b_1} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{p}\,,$$

ktoré budú neskreslené a pomocou nich potom je uskutočnený odhad pravdepodobnosti p_i a vypočítaná matica V, ktorá je použitá v druhej fáze tejto metódy na upresnenie hľadaných parametrov

$$\mathbf{b_2} = (\mathbf{X^T \hat{V}^{-1} X})^{-1} \mathbf{X^T \hat{V}^{-1} p},$$

kde $\mathbf{\hat{V}^{-1}}$ je odhad matice, získaný z regresných parametrov v prvom stupni.

Newton-Raphsonova iteračná metóda

Táto metóda je založená na kvadratickej aproximácii extremalizovanej kriteriálnej funkcie prvými troma členmi Taylorovho rozvoja (napr. Hušek, 1998). Je bežne používaná na odhad parametrov zovšeobecneného lineárneho modelu. Východiskom sú určité počiatočné odhady vektoru parametov $\boldsymbol{\beta}$, ktoré si označíme ako $\mathbf{b_0}$, nutných na výpočet ostatných štatistík modelu, vrátane matice $\mathbf{V}$. Tieto odhady sa potom neustále upresňujú dovtedy, kým rozdiely medzi jednotlivými iteráciami neklesnú pod určitú vopred zadanú úroveň (napr. 0,0001). Podrobnejší postup je uvedený u nelineárneho logitového modelu.

U LPM uvedieme výsledný iteračný vzorec. Odhad vektoru parametra $\boldsymbol{\beta}$ v s+1 iterácií je vektor

$$\mathbf{b_{s+1}} = \mathbf{b_s} + [\mathbf{H}(\mathbf{b_s})]^{-1}\mathbf{g}(\mathbf{b_s}),$$

kde $\mathbf{H}(\mathbf{b_s})$ – Hessova matica o rozmere $n \times n$,

$\mathbf{g}(\mathbf{b_s})$ – vektor o rozmere $n \times 1$.

Pre vektor $\mathbf{g}(\mathbf{b_s})$ (nazývaný aj gradient) platí

$$\mathbf{g}(\boldsymbol{\beta}) = \frac{\partial\, ln\, L(\mathbf{y}|\boldsymbol{\beta})}{\partial\, \boldsymbol{\beta}}. \tag{4.10}$$

Výsledný tvar (4.10) je uvedený pri metóde maximálnej vierohodnosti v časti 4.4.1 resp. 4.4.2.

Pre Hessovu maticu platí

$$
\begin{aligned}
\mathbf{H}(\boldsymbol{\beta}) &= \frac{\partial^2 \ln L(\mathbf{y}|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}} = \frac{\partial}{\partial \boldsymbol{\beta}_{j'}} \sum_{i=1}^{n} X_{ij} \frac{1}{p_i(1-p_i)}(y_i - p_i) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}_{j'}} \sum_{i=1}^{n} X_{ij}(p_i - p_i^2)^{-1} y_i - \frac{\partial}{\partial \beta_{j'}} \sum_{i=1}^{n} X_{ij}(1-p_i)^{-1} \\
&= \sum_{i=1}^{n} \frac{(-1)X_{ij}(X_{ij'} - 2p_i X_{ij'})y_i}{p_i^2(1-p_i)^2} - \sum_{i=1}^{n} \frac{X_{ij}X_{ij'}p_i^2}{p_i^2(1-p_i)^2} \\
&= -\sum_{i=1}^{n} X_{ij}X_{ij'} \frac{p_i^2 + (1-2p_i)y_i}{p_i^2(1-p_i)^2}
\end{aligned}
\tag{4.11}
$$

Výraz (4.11) môžeme zapísať v tvare

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^{\mathbf{T}}\mathbf{V}\mathbf{X}, \tag{4.12}$$

kde $\mathbf{V}$ je diagonálna matica s prvkami $v_i = \dfrac{p_i^2 + (1-2p_i)y_i}{p_i^2(1-p_i)^2}$.

Pre odhady regresných parametrov Newton-Raphsonovou interačnou metódou platí

$$\mathbf{b_{s+1}} = \boldsymbol{\beta}_\mathbf{s} + [\mathbf{H}(\mathbf{b_s})]^{-1}\mathbf{g}(\mathbf{b_s}) = \mathbf{b_s} + [\mathbf{X}^\mathbf{T}\hat{\mathbf{V}}_\mathbf{s}^{-1}\mathbf{X}]^{-1}\mathbf{X}^\mathbf{T}\hat{\mathbf{V}}_\mathbf{s}^{-1}\mathbf{p}\,. \tag{4.13}$$

V predchádzajúcej podkapitole boli uvedené štatistické vlastnosti LPM a ukázané možné odhady regresných parametrov tohto modelu od klasickej MNŠ až po zložitejšiu Newton-Raphsonovu iteračnú metódu. Odhad LPM metódou vážených najmenších štvorcov uvedieme nižšie. LPM boli formulované na základe transformačného alebo štatistického prístupu a v prípade existencie viacnásobnej regresie aj z toho odvodených početností. Tretí prístup k formulácii LPM, ktorý spája problematiku voľby u jednotlivca s teóriou náhodnej užitočnosti, je uvedený nižšie.

4.3.1 Alternatívny prístup k formulácii LPM

Ďalším z alternatívnych prístupov k špecifikácii ekonometrického modelu diskrétnej voľby je prístup pomocou tzv. **latentnej** alebo **nemerateľnej náhodnej premennej**, ktorý umožňuje spojiť problém voľby s teóriou náhodnej užitočnosti (Hušek, 2003).

Tento prístup predpokladá existenciu **indexu užitočnosti** I_i v tvare latentnej alebo nemerateľnej spojitej náhodnej premennej, ktorá existuje pre ľubovoľný ***i***-ty individuálny subjekt v súlade s teóriou racionálnej voľby. Je definovaná nasledovným lineárnym regresným vzťahom (jeho linearita však nie je podmienkou)

$$I_i = \beta_0 + \beta_1 X_i \,,$$

alebo jeho všeobecný tvar je

$$I_i = \mathbf{x}_i^T \boldsymbol{\beta}\,, \qquad -\infty < I_i < \infty, \tag{4.14}$$

kde

$\mathbf{x}_i^T$ – vektor vysvetľujúcich premenných o rozmere $1\times K$,

$\boldsymbol{\beta}$ – vektor neznámych parametrov o rozmere $K\times 1$,

čo znamená, že so zmenou vektoru $\mathbf{x}_i^T$ sa v obore reálnych čísel zmení aj hodnota indexu I_i.

Za uvedených predpokladov môžeme lineárny pravdepodobnostný model tvaru (4.3) vyjadriť ako

$$Y_i = I_i(Y_i^*), \tag{4.15}$$

kde

$$Y_i^* = \mathbf{x}_i^T \boldsymbol{\beta} + u_i\,, \tag{4.16}$$

a Y_i^* je latentná (nemerateľná) premenná. Pozorovania sú však k dispozícii iba pre jej dichotomickú realizáciu, čiže pre náhodnú premennú Y_i, pre ktorú platí

$$Y_i = \begin{cases} 1 & \text{ak } Y_i^* > 0, \\ 0 & \text{ak } Y_i^* \le 0. \end{cases} \tag{4.17}$$

Latentná premenná Y_i^* vyjadruje súhrnný vplyv všetkých vysvetľujúcich premenných na dichotomické hodnoty Y_i a to vo forme reálneho skalárneho indexu užitočnosti I_i. Ak napríklad znamenajú napozorované hodnoty merateľnej dichotomickej premennej Y_i prípad, či subjekt vlastní alebo nevlastní určitý predmet dlhodobej spotreby, potom možno latentnú premennú Y_i^*, vyjadrenú indexom charakterizovať ako **kritickú (prahovú) hodnotu**, mieru želania alebo schopnosti (sklonu) tento predmet vlastniť.

Pri predpoklade, že sa individuálne ekonomické subjekty chovajú tak, aby maximalizovali svoju užitočnosť, zvolí ***i***-ty subjekt alternatívu indexu označenú ako $Y_i = 1$ v tom prípade, ak latentná náhodná premenná $Y_i^* = U_{i1} - U_{i0}$ nadobúda také hodnoty, že $y_i^* > 0$. Potom hodnoty merateľnej dichotomickej premennej Y_i určené pomocou vzťahu (Mittelhammer, et al. 2000)

$$y_i = \begin{cases} 1 \quad \text{ak } y_i^* > 0 \quad \text{alebo } U_{i1} > U_{i0}, \\ 0 \quad \text{ak } y_i^* \le 0 \quad \text{alebo } U_{i1} \le U_{i0}, \end{cases} \tag{4.18}$$

Na základe (4.15) a (4.16) môžeme písať (napr. Maddala, 2001)

$$\begin{aligned} p_i &= P(y_i = 1) = P(y_i^* > 0) = P(u_i > -\mathbf{x}_i^T \boldsymbol{\beta}) \\ &= 1 - F(-\mathbf{x}_i^T \boldsymbol{\beta}), \end{aligned} \tag{4.19}$$

kde F je kumulatívna distribučná funkcia (KDF) náhodnej zložky u_i. Pomocou rovnice (4.19) je definovaný vzťah medzi výberovou pravdepodobnosťou p_i a systematickou zložkou modelu binárnej diskrétnej voľby $\mathbf{x}_i^T \boldsymbol{\beta}$ na základe funkčného tvaru KDF náhodnej zložky modelu. Ak je zvolená KDF symetrická okolo nulovej strednej hodnoty (čo je častý prípad), čiže v tvare

$$1 - F(-Z) = F(Z)$$

potom môžeme vzťah (4.15) prepísať v tvare

$$p_i = F(\mathbf{x}_i^T \boldsymbol{\beta}) = F(I_i), \tag{4.20}$$

a tento funkčný tvar závisí na charaktere rozdelenia náhodnej zložky u_i.

Štandardnou vlastnosťou KDF je, že $F(Z) \in \langle 0,1 \rangle$, tak aj pravdepodobnosti p_i nadobúdajú hodnoty iba z intervalu $\langle 0,1 \rangle$.

Ako príklad lineárneho pravdepodobnostného modelu s nesymetrickým rozdelením slúži tzv. **Weibullov** model v tvare (Greene, 2003)

$$p_i = (Y_i = 1 | \mathbf{x}_i) = \exp[-\exp(\mathbf{x}_i^T \boldsymbol{\beta})],$$

prípadne **doplnkový dvojlogaritmický model**

$$p_i = (Y_i = 1 | \mathbf{x}_i) = 1 - \exp[-\exp(\mathbf{x}_i^T \boldsymbol{\beta})]$$

K požiadavkám kladených na model binárnej diskrétnej voľby patrí, aby vysvetľovaná premenná modelu, ktorá predstavuje pravdepodobnosť zvolenia skúmanej alternatívy nadobúdala hodnoty v intervale $\langle 0,1 \rangle$. Z toho vyplýva, že $F(.)$ je spojitá funkcia nelineárna aspoň v krajných hraniciach intervalu $\langle 0,1 \rangle$.

Z tohto dôvodu sa symetrická KDF zdá ako zodpovedajúci spôsob transformácie pôvodného lineárneho pravdepodobnostného modelu binárnej voľby. Symetrická KDF tak predstavuje pravdepodobnostnú transformáciu indexu I_i a pravdepodobnosti p_i sa nachádzajú v intervale $\langle 0,1 \rangle$. Okrem toho, táto KDF predstavuje monotónnu závislosť medzi p_i a I_i, čo znamená, že s rastúcou hodnotou indexu I_i má i-ty individuálny ekonomický subjekt vyššiu užitočnosť pri zvolení alternatívy $Y_i = 1$.

Výhodou KDF využívanej na analýzu a predikciu modelov diskrétnej voľby je aj to, že pri rovnakých zmenách vysvetľujúcich premenných vo vektore $\mathbf{x}_i^T$ vykazujú KDF tím menšie prírastky pravdepodobnosti p_i, čím viac sa ich hodnoty blížia k jednej z hraníc intervalu $\langle 0,1 \rangle$.

Zvolenie nelineárneho tvaru modelu diskrétnej voľby podporuje aj skutočnosť, že pri výskyte väčšieho počtu vysvetľujúcich premenných v modeli nemožno očakávať ich aditívne pôsobenie na vysvetľovanú premennú. Medzi vysvetľujúcimi premennými môže byť určitý vzťah alebo **interakcia**, takže hodnota pravdepodobnosti p_i vyvolaná zmenou iba konkrétnej jedinej vysvetľujúcej premennej závisí aj na hodnotách niektorých ďalších vysvetľujúcich premenných modelu vo väčšej alebo menšej miere. Vplyv tejto inerakcie nemožno postihnúť pomocou LPM.

4.3.2 Nedostatky u LPM a jeho interpretácia

Lineárny pravdepodobnostný model binárnej voľby napriek svojej výpočtovej jednoduchosti a ľahkej interpretácii získaných výsledkov má niekoľko zásadných nedostatkov, medzi ktoré okrem iného patrí aj porušenie Gauss-Markovových predpokladov lineárneho regresného modelu. Nedostatky sú považované za pomerne významné, a je im venovaná táto podkapitola.

Typ rozdelenia náhodnej zložky

Dichotomická náhodná zložka u_i nemá normálne rozdelenie, pretože rovnako ako premenná Y_i, aj náhodná zložka u_i nadobúda iba dve hodnoty, čiže má Bernoulliho rozdelenie (4.2).

Ak je model (4.3) zapísaný v tvare

$$u_i = Y_i - \beta_0 - \beta_1 X_i \tag{4.21}$$

tak, v prípade, ak diskrétna náhodná premenná $Y_i = 1$, náhodná zložka je v tvare

$$u_i = 1 - \beta_0 - \beta_1 X_i \tag{4.22}$$

a v prípade pre hodnotu $Y_i = 0$

$$u_i = - \beta_0 - \beta_1 X_i \ . \tag{4.23}$$

V prípade problému s normálnym rozdelením náhodnej zložky u_i nastávajú komplikácie pri súbore s malým počtom pozorovaní, kde síce bodový odhad parametrov modelu je platný, ale problémom je testovanie hypotézy o významnosti modelu ako celku a konštrukcia intervalo spoľahlivosti. Existuje riešenie v prípade, že veľkosť sledovaného súboru rastie donekonečna, takže binomické rozdelenie náhodnej zložky aproximujeme normálnym rozdelením.

Heteroskedastický rozptyl náhodnej zložky LPM

V modeli je predpokladané splnenie nasledovných požiadavkov kladených na klasický (štandardný) lineárny regresný model:

- $E(u_i) = 0$, čiže náhodné zložky majú vo všetkých výberoch indentické rozdelenie s nulovou strednou hodnotou,
- $E(u_i u_j) = 0$ pre i ¹ j, čiže neexistuje sériová korelácia, čo však neznamená, že náhodná zložka u_i je homoskedastická.

Rozdelenie náhodných zložiek u_i podľa (4.22) a (4.23) je P_i a $1 - P_i$.

Rozptyl náhodnej zložky u_i je rovný

$$\mathrm{var}(u_i) = E[u_i - E(u_i)]^2 = E(u_i^2) \qquad \text{pre } E(u_i) = 0$$

Z toho vyplýva

$$\begin{aligned} \mathrm{var}(u_i) = E(u_i^2) &= (-\beta_0 - \beta_1 X_i)^2 (1 - p_i) + (1 - \beta_0 - \beta_1 X_i)^2 (p_i) \\ &= (-\beta_0 - \beta_1 X_i)^2 (1 - \beta_0 - \beta_1 X_i) + (1 - \beta_0 - \beta_1 X_i)^2 (\beta_0 + \beta_1 X_i) \\ &= (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i) \end{aligned} \tag{4.24}$$

alebo

$$\mathrm{var}(u_i) = E(Y_i \mid X_i)[1 - E(Y_i \mid X_i)] = p_i(1 - p_i) \tag{4.25}$$

kde toto použitie je založené na fakte, že stredná hodnota $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i = p_i$.

Z rovnice (4.25) je zrejmé, že rozptyl náhodnej zložky u_i je heteroskedastický z dôvodu jeho závislosti na premenných $i = 1,2,...,M$. Pri existencii heteroskedasticity náhodnej zložky, nie je odhad modelu MNŠ efektívny, čiže neexistuje minimálny rozptyl.

Avšak ani tento problém nie je neprekonateľný, problém heteroskedasticity u náhodnej zložky LPM je odstraniteľný pomocou **metódy vážených najmenších štvorcov**. Keďže rozptyl náhodnej zložky u_i závisí na očakávaných hodnotách diskrétnej náhodnej premennej Y_i (a tá závisí na hodnotách premennej X), ako vidíme z (4.24), upravíme dáta vydelením oboch strán modelu (4.3) váhami w_i a získame (Gujarati, 1998)

$$\sqrt{E(Y_i|X_i)[1-E(Y_i|X_i)]} = \sqrt{p_i(1-p_i)} = \sqrt{w_i}$$

$$\frac{Y_i}{\sqrt{w_i}} = \frac{b_0}{\sqrt{w_i}} + b_1\frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \tag{4.26}$$

Náhodná zložka v (4.26) je po tejto transformácii homoskedastická a odhad modelu môžeme uskutočniť metódou najmenších štvorcov (MNŠ). Keďže nepoznáme skutočnú strednú hodnotu $E(Y_i|X_i)$ a ani váhy w_i, avšak váhy možno odhadnúť nasledujúcim spôsobom. Napriek prítomnosti homoskedasticity v modeli (4.3) odhadneme model metódou najmenších štvorcov a získame odhad $\hat{Y}_i$, ktorý je rovný $E(Y_i|X_i)$. Potom odhad váh w_i je rovný $\hat{w}_i = \hat{Y}_i(1-\hat{Y}_i)$. V ďalšom kroku sú odhadnutými váhami $\hat{w}_i$ upravené premenné ako v (4.26) a model je opäť odhadnutý MNŠ, čím sú získané konzistentné, ale vychýlené odhady parametrov LPM.

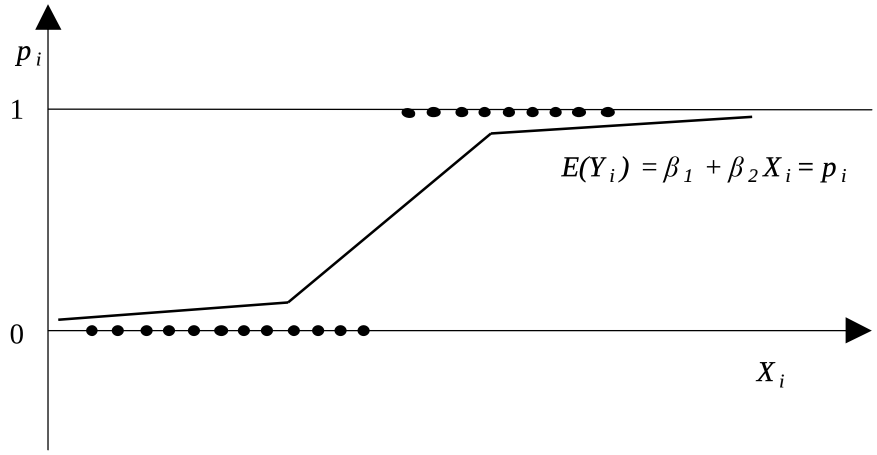
Pravdepodobnosti ležia mimo interval á0,1ñ

Pri modelovaní pravdepodobnosti, je potrebné, aby hodnoty p_i ležali v rámci intervalu $\langle 0,1\rangle$, čo nespĺňajú krajné hodnoty pozorovaní. Existujú dva spôsoby, ako zistiť, či odhadnuté hodnoty $\hat{Y}_i$ ležia v intervale $\langle 0,1\rangle$. Jedným z nich je odhadnúť model MNŠ a zistiť, či odhadnuté hodnoty $\hat{Y}_i$ patria do tohto intervalu. Ak sú niektoré hodnoty $\hat{Y}_i$ záporné, tak im priradíme hodnotu nula a hodnotám $\hat{Y}_i$ väčším ako jedna priradíme hodnotu jedna, čiže

$$p_i = \begin{cases} 0 & \text{pre} \quad b_1 + b_2X_i < 0 \\ b_1 + b_2X_i & \text{pre} \quad 0 \le b_1 + b_2X_i \le 1 \\ 1 & \text{pre} \quad b_1 + b_2X_i > 1 \end{cases} \tag{4.27}$$

V tomto prípade bude mať graf LPM tvar

Obrázok 4.3 LPM s upravenými p_i podľa (4.27)



Ďalším spôsobom je navrhnúť takú metódu odhadu, ktorá zaručí, aby odhadnuté pravdepodobnosti $\hat{Y}_i$ ležali v intervale $\langle 0,1 \rangle$. Tými metódami sú nelineárny logitový a probitový pravdepodobnostný model.

Testy vypovedacej schopnosti modelu

Pri testovaní celkovej zhody LPM s dátami nie je vhodné použitie klasického koeficientu viacnásobnej determinácie R^2. Jeho hodnoty sú totiž i v prípade vysokej vypovedacej schopnosti odhadnutého modelu podstatne menšie ako jedna (Pindyck a Rubinfeld, 1998).

Linearita modelu

Vyššie uvedené problémy sa dajú viac-menej upraviť. Najpodstatnejším problémom lineárneho pravdepodobnostného modelu je jeho vlastná linearita, ktorá navyše vedie ku konštantnému marginálnemu efektu vysvetľujúcich premenných, čo spravidla neplatí. Znamená to, že potrebujeme model, ktorý bude spĺňať nasledovné požiadavky:

- so zvyšovaním hodnôt X_i, bude rásť pravdepodobnosť $p_i = E(Y = 1|X)$, avšak nebude v žiadnom prípade ležať mimo intervalu $\langle 0,1 \rangle$,
- vzťah medzi p_i a X_i bude nelineárny, čo znamená, že pre krajné hodnoty premennej X sú prírastky (úbytky) pravdepodobnosti P_i menšie a menšie.

Interpretácia odhadnutého LPM

Výhodou LPM je jeho jednoduchá aplikovateľnosť a interpretácia parametrov $\boldsymbol{\beta}$ odhadnutého modelu. Kladné/záporné znamienko odhadnutého parametru β_k zvyšuje/znižuje skutočnú podmienenú pravdepodobnosť p_k. Vypočítané pravdepodobnosti p_i vyjadrujú skutočné podmienené pravdepodobnosti zvolenia konkrétnej alternatívy individuálnym ekonomickým subjektom.

Po odhade parametrov $\boldsymbol{\beta}$ modelu, je zaujímavé zistenie, akým spôsobom sa odrazí zmena (nárast alebo pokles) v niektorej z vysvetľujúcich premenných na odhadnutej pravdepodobnosti p_i, čiže tzv. **marginálny efekt** príslušnej k-tej vysvetľujúcej premennej X_{ik} na pravdepodobnosti p_i. Výpočet marginálneho efektu uskutočníme pomocou parciálnych derivácií vo forme

$$\frac{\partial p_i}{\partial X_{ik}} = b_k . \tag{4.28}$$

Zo (4.28) je zrejmé, že marginálny efekt v prípade LPM je konštantný a priamo určený parametrom β_k.

4.4 Nelineárne pravdepodobnostné modely

K najčastejšie používaným KDF pri modelovaní problémov diskrétnej voľby patrí KDF normálneho a logistického rozdelenia. Tieto rozdelenia sú symetrické okolo nulových stredných hodnôt a ich KDF majú tvar tzv. S-kriviek. Rozptyl štandardného normálneho rozdelenia je rovný jednej, rozptyl logistického rozdelenia je rovný $\pi^2/3$.

Na základe špecifikácie parametrov nelineárneho modelu diskrétnej voľby, ktorá vychádza zo zvolenej konkrétnej KDF vyjadrujúcej funkčný vzťah medzi p_i a $I_i = \mathbf{x}_i^T \boldsymbol{\beta}$, možno vzhľadom k (4.2) a (4.19) špecifikovať vierohodnostnú funkciu parametrov pravdepodobnostného modelu v tvare

$$\begin{aligned} L(\boldsymbol{\beta},\mathbf{y}) &= \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{y_i} \\ &= \prod_{i=1}^{n} \left[1-F(\mathbf{x}_i^T \boldsymbol{\beta})\right]^{y_i} F(-\mathbf{x}_i^T \boldsymbol{\beta})^{1-y_i} , \end{aligned} \tag{4.29}$$

a po použití prirodzených logaritmov

$$
\begin{aligned}
ln\left[L(\boldsymbol{\beta},\mathbf{y})\right] &= \sum_{i=1}^{n}\left[y_i\, ln(p_i) + (1-y_i)\, ln(1-p_i)\right] \\
&= \sum_{i=1}^{n}\left[y_i\, ln(1-F(\mathbf{x}_i^T\boldsymbol{\beta})) + (1-y_i)\, ln\, F(-\mathbf{x}_i^T\boldsymbol{\beta})\right]
\end{aligned}
\tag{4.30}
$$

Princíp odhadu LPM spočíva v nájdení takého vektora **β**, ktorý bude maximalizovať vierohodnostnú funkciu (4.29).

Po jej parciálnom zderivovaní vyjadríme podmienky prvého rádu v tvare

$$
\frac{\partial\, ln\left[L(\boldsymbol{\beta},\mathbf{y})\right]}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\left[y_i \frac{f(\mathbf{x}_i^T\boldsymbol{\beta})}{F(\mathbf{x}_i^T\boldsymbol{\beta})} + (1-y_i)\frac{-f(\mathbf{x}_i^T\boldsymbol{\beta})}{1-F(\mathbf{x}_i^T\boldsymbol{\beta})}\right]\mathbf{x}_i = 0\,,
\tag{4.31}
$$

kde $f(z) = \partial F(z)/\partial z.$

4.4.1 Logitový model

Logistické regresné modely patria medzi zovšeobecnené lineárne modely (generalized linear models), ktoré sa používajú na predikciu závislej premennej Y s diskrétnym rozdelením (môže nadobúdať iba obmedzený počet hodnôt) a/alebo premenná Y má nelineárny vzťah k prediktorom X.

Základ pravdepodobnostného logitového modelu tvorí KDF logistickej náhodnej premennej pre rôzne hodnoty x_i, ktorej funkčný predpis je (Maddala, 2001)

$$
p_i = F_L(I_i) = \frac{e^{I_i}}{1+e^{I_i}} = \frac{1}{1+e^{-I_i}}\,, \quad -\infty < I_i < \infty,
\tag{4.32}
$$

kde $F_L(I_i)$ je KDF logistickej náhodnej premennej u_i. Výhodnou vlastnosťou klesajúcej logistickej funkcie hustoty pravdepodobnosti je, že je to hladká krivka, pre $I_i < 0$ je konvexná a pre $I_i > 0$ je konkávna a v bode $I_i = 0$ má inflexný bod.

Pre limity logistickej funkcie platí

$$\lim_{I_i \to -\infty} f(I_i) = \lim_{I_i \to -\infty} \frac{1}{1+e^{-I_i}} = 0$$

a

$$\lim_{I_i \to \infty} f(I_i) = \lim_{I_i \to \infty} \frac{1}{1+e^{-I_i}} = 1.$$

Z toho vyplýva, že ľavý koniec logistickej krivky sa blíži k hodnote nula a pravý k hodnote jedna.

Logistické rozdelenie je možné veľmi dobre aproximovať pomocou *t* rozdelenia s malým počtom stupňov voľnosti (napr. sedem).

Ak prevedieme výraz (4.32) na tvar

$$F_L(I_i) = \frac{e^{I_i}}{1+e^{I_i}}, \tag{4.33}$$

tak po úprave dostaneme

$$\begin{aligned} F_L(I_i) + F_L(I_i)e^{I_i} &= e^{I_i} \\ F_L(I_i) &= e^{I_i} - F_L(I_i)e^{I_i} \\ F_L(I_i) &= e^{I_i}\left[1 - F_L(I_i)\right] \\ \frac{F_L(I_i)}{1-F_L(I_i)} &= e^{I_i} \end{aligned} \tag{4.34}$$

Po zlogaritmovaní predchádzajúceho výrazu (4.34) dostaneme

$$ln\left[\frac{F_L(I_i)}{1-F_L(I_i)}\right] = I_i, \tag{4.35}$$

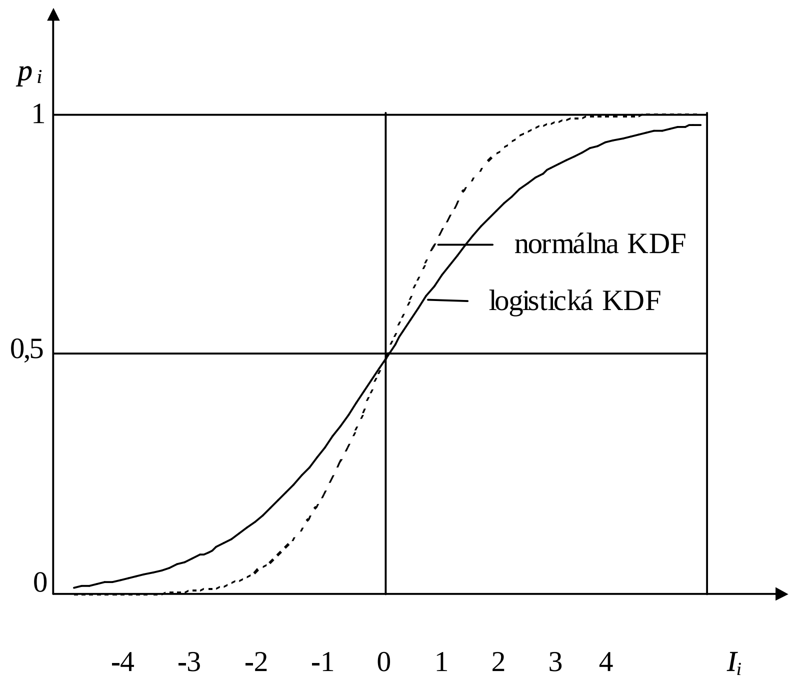
a po substitúcii (4.20) za $F_L(I_i)$ zapíšeme

$$L_i = ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{4.36}$$

Výraz predstavuje logaritmus podielu pravdepodobností (odds ratio) oboch možných alternatív a je označovaný názvom **logit**, v tomto prípade to je výberový logit (Berkson, 1944). Zo vzorca vidíme, že závislosť logitov na vektore vysvetľujúcich premenných $\mathbf{x}_i^T$ je lineárna a jej hodnoty nie sú nijako obmedzené. Z toho vyplýva, že na odhad vektora parametrov $\boldsymbol{\beta}$ v (4.36) možno využiť aj odhadové postupy založené na kritériu najmenších štvorcov, pričom je najprv potrebné vhodným spôsobom eliminovať heteroskedasticitu (napr. Gujarati, 1988 alebo Hušek, 1995).

Znamená to, že funkcia logit je vlastne inverznou fukciou k logistickej krivke a pomocou nej môžeme odvodiť vlastnosti logitu. Funkcia logitu je klesajúcou funkciou, v nule sa blíži k $+\infty$ a v jednotke k $-\infty$. Logistická distribučná funkcia se používa z dôvodu dobrej aproximácie normálnej distribučnej funkcie (s maximálnou odchýlkou 0,11). Funkcia vykazuje zrýchľujúci sa rast až do inflexného bodu, potom sa rast spomaľuje a funkcia sa asymptoticky blíži k hodnote 1.

Obrázok 4.4 KDF logistického a štandardizovaného normálneho rozdelenia



Logitový model binárnej diskrétnej voľby dosahuje maximálnu hodnotu pravdepodobnosti $f(\mathbf{x}_i^T\boldsymbol{\beta}) = 0{,}5$ pre hodnotu indexu užitočnosti $I_i = 0$. Po vyššie uvedenej transformácii bude pravdepodobnosť zvolenia alternatívy p_i ležať v intervale $\langle 0,1 \rangle$ a v blízkosti hraničných hodnôt tohto intervalu budú prírastky $f(\mathbf{x}_i^T\boldsymbol{\beta})$ klesať.

Metódy odhadu parametrov logitového modelu

Okrem LPM sú najčastejšie parametre modelov binárnej voľby odhadnuté MMV u individuálnych dát, ktoré sú náhodným výberom z Bernoulliho rozdelenia o rozsahu *n*.

Logaritmická funkcia maximálnej vierohodnosti pre logitový model diskrétnej voľby má tvar (Greene, 2003)

$$\begin{aligned} ln\left[L(\boldsymbol{\beta},\mathbf{y})\right] &= \sum_{i=1}^{n}\left[y_i\, ln\left(\frac{e^{I_i}}{1+e^{I_i}}\right) + (1-y_i)\, ln\left(\frac{1}{1+e^{I_i}}\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i I_i - ln(1+e^{I_i})\right] \end{aligned} \tag{4.37}$$

Podmienky prvého rádu pre maximum funkcie (4.37) majú podobu

$$\frac{\partial\, ln\left[L(\boldsymbol{\beta},\mathbf{y})\right]}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left[y_i\,\mathbf{x}_i - \frac{e^{I_i}}{1+e^{I_i}}\mathbf{x}_i\right] = \sum_{i=1}^{n}\left[y_i - p_i(\boldsymbol{\beta})\right]\mathbf{x}_i = 0, \tag{4.38}$$

kde

$$p_i(\boldsymbol{\beta}) = F_L(I_i) = \frac{e^{I_i}}{1+e^{I_i}} = P(y_i = 1)\,. \tag{4.39}$$

Sústava normálnych rovníc odhadovej funkcie (4.38) je nelineárna v $\boldsymbol{\beta}$. Matica druhých parciálnych derivácií je daná výrazom

$$\frac{\partial^2\, ln\left[L(\boldsymbol{\beta},\mathbf{y})\right]}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n}\frac{e^{I_i}}{[1+e^{I_i}]^2}\mathbf{x}_i\mathbf{x}_i^T \tag{4.40}$$

Jediné riešenie sústavy (4.38) pomocou MMV pre vektor **β** existuje v prípade, že Hessova matica je negatívne definitná. Odhadová funkcia MMV pre vektor **β** logitového modelu má pre veľké výbery asymptotické vlastnosti, čiže je konzistentná, asymptoticky výdatná a normálne rozdelená (Mittelhammer, et al., 2000).

Predikcia u logitového modelu

Znamienko odhadnutých parametrov **β** logitového modelu, podobne ako v prípade LPM, zvyšuje (kladné) alebo znižuje (záporné) pravdepodobnosť p_i.

Podiel pravdepodobností je vlastne tvar pred zlogaritmovaním $\frac{p_i}{1-p_i}$ vyjadruje **šancu zvolenej varianty**, tj. koľkokrát je pravdepodobnejšie, že hodnota premennej y_i = 1, než že y_i = 0. Napríklad ak hodnota pravdepodobnosti p_i = 0,8, a teda 1 – p_i = 0,2, je podiel pravdepodobností rovný $\frac{p_i}{1-p_i}$ = 4. Zvolenie určitej alternatívy (napr. nákup predmetu dlhodobej spotreby) je teda štyrikrát pravdepodobnejšie ako jej nezvolenie, prípadne zvolenie alternatívy opačnej.

Absolútny člen β_0 je hodnotou logaritmu zvolenej alternatívy, ak sú všetky vysvetľujúce premenné nulové.

Marginálny efekt u logitového modelu

Marginálny efekt, teda zmena v niektorej z vysvetľujúcich premenných na odhadnutej pravdepodobnosti p_i, je daná vzťahom (Maddala, 1999)

$$\frac{\partial p_i}{\partial X_{ik}} = b_k \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{[1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}]} = b_k\, p_i (1-p_i), \tag{4.41}$$

pretože vzhľadom k (4.39) platí

$$1-p_i = \frac{1}{1+e^{I_i}}$$

Vzťah (4.41) však platí iba pre prípad, kedy vysvetľujúca premenná je **spojitého charakteru**. Pre prípad **umelej nula-jednotkovej vysvetľujúcej** premennej však marginálny efekt počítame ako rozdiel medzi odhadnutou pravdepodobnosťou p_i, kedy k-ta umelá nula-jednotková vysvetľujúca premenná je rovná nule a p_i s umelou premennou rovnou hodnote jedna v tvare

$$[p_i = P(y_i = 1 \mid X_{ik} = 0)] - [p_i = P(y_i = 1 \mid X_{ik} = 1)], \tag{4.42}$$

kde za hodnoty ostatných vysvetľujúcich premenných v logitovom modeli sú dosadené ich priemery.

Pri predikovaní vplyvu napr. predpokladanej zmeny vysvetľujúcej premennej X_{ik} na očakávanú hodnotu logitu L_i, jeho veľkosť je meraná priamo pomocou parametru β_k, pretože pre logitový model platí

$$\frac{\partial L_i}{\partial X_{ik}} = \frac{\partial}{\partial X_{ik}}\left[ln\left(\frac{p_i}{1-p_i} \right)\right] = b_k . \tag{4.43}$$

Uvedený výraz je interpretovaný tak, že pri jednotkovom zvýšení vysvetľujúcej premennej X_{ik} a konštantnej úrovni ostatných vysvetľujúcich premenných sa logit L_i zmení v priemere o β_k jednotiek.

4.4.2 Probitový model

Nelineárny pravdepodobnostný probitový (v literatúre známy aj ako *normitový*) model, podobne ako logitový model, predpokladá nelineárnu závislosť vysvetľovanej pravdepodobnosti p_i zvolenia konkrétnej alternatívy na vysvetľujúcich faktoroch. Je založený na KDF štandardne normálne rozdelenej náhodnej zložky u_i. Funkčná závislosť medzi p_i a $I_i = \mathbf{x}_i^T \boldsymbol{\beta}$ je daná nasledovným vzťahom

$$p_i = P(t \le I_i) = F_N(I_i) = \int_{-\infty}^{I_i} \frac{1}{\sqrt{2p}} e^{\frac{-t^2}{2}} dt, \tag{4.44}$$

kde $F_N(I_i)$ je KDF štandardného normálneho rozdelenia a $t \sim N(0,1)$.

Pravdepodobnosť p_i, že endogénna premenná $Y_i = 1$ nadobúda hodnoty z intervalu $\langle 0,1 \rangle$, pretože to je pravdepodobnosť, že štandardizovaná normálna náhodná premenná t je maximálne rovná indexu užitočnosti $I_i = \mathbf{x}_i^T \boldsymbol{\beta}$. S nárastom I_i od $-\infty$ do $+\infty$ sa tak pravdepodobnosť p_i zvyšuje monotónnym spôsobom.

Vlastnosť neklesajúcej KDF probitového modelu binárnej voľby spôsobuje, že s rastúcou hodnotou indexu užitočnosti sa zvyšuje pravdepodobnosť, že individuálny ekonomický subjekt zvolí alternatívu 1. Indexu užitočnosti $I_i = 0$ zodpovedá pravdepodobnosť $p_i = 0{,}5$, čiže pravdepodobnosti voľby sú pre obe alternatívy rovnaké. S rastúcou hodnotou I_i sa pravdepodobnosť zvolenia alternatívy 1 zvyšuje, ale prírastky pravdepodobnosti p_i nie sú pre rovnako veľké zmeny indexu užitočnosti konštantné v porovnaní s LPM binárnej voľby.

Zmeny vysvetľujúcich premenných vo vektore $\mathbf{x}_i^T$, ovplyvňujúce voľbu jednotlivých alternatív individuálnymi subjektami, majú najväčšiu účinnosť v inflexnom bode KDF, keď $I_i = 0$ (funkcia hustoty pravdepodobnosti $f(I_i)$ nadobúda maximum) a zodpovedajúca pravdepodobnosť $p_i = F(I_i = 0) = 0{,}5$. Čím viac sa pravdepodobnosť blíži ku krajným hraniciam intervalu $\langle 0,1 \rangle$, hodnota $f(I_i)$ klesá, čiže zmeny vektoru $\mathbf{x}_i^T$ majú stále menší efekt.

Postupy pri odhade parametrov probitového modelu binárnej voľby sú analogické ako v prípade logitového modelu. Tvar odhadovej funkcie MMV je

$$ln[L(\boldsymbol{\beta},\mathbf{y})] = \sum_{i=1}^{n} [y_i ln(1 - F_N(-I_i)) + (1 - y_i) ln(F_N(-I_i))]. \tag{4.45}$$

Podmienky prvého rádu pre maximalizáciu vierohodnostnej funkcie v (4.45) majú tvar

$$\frac{\partial\, ln[L(\boldsymbol{\beta},\mathbf{y})]}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} [y_i - p_i(\boldsymbol{\beta})] f(I_i) \mathbf{x}_i = 0, \tag{4.46}$$

kde

$$p_i(\boldsymbol{\beta}) = F_N(I_i), \tag{4.47}$$

$$f(I_i) = \frac{f_N(I_i)}{F_N(I_i)\left[1 - F_N(I_i)\right]}. \tag{4.48}$$

Keďže sústava normálnych rovníc v (4.46) je nelineárna v parametroch $\boldsymbol{\beta}$, východisko pri jej riešení je niektorá z metód numerickej optimalizácie (napr. Hušek, 1986, 1998).

Po výpočte druhých parciálnych derivácií odhadovej funkcie MMV v (4.45) podľa $\boldsymbol{\beta}$, vyjadríme pomocou vzťahov

$$\frac{\partial^2 \ln\left[L(\boldsymbol{\beta},\mathbf{y})\right]}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \lambda_i(I_i)\left[\lambda_i(I_i) + I_i\right]\mathbf{x}_i\, \mathbf{x}_i^T, \tag{4.49}$$

kde

$$l_i(I_i) = y_i\left[\frac{f_N(I_i)}{F_N(I_i)}\right] + (1 - y_i)\left[\frac{-f_N(I_i)}{1 - F_N(I_i)}\right]. \tag{4.50}$$

Možno dokázať, že ak je Hessova matica negatívne definitná, existuje pre voľbu parametrov $\boldsymbol{\beta}$ v (4.46) jediné riešenie (Amemiya, 1985), ktoré je optimálnou odhadovou funkciou MMV pre $\boldsymbol{\beta}$ v tom zmysle, že má najmenšiu asymptotickú kovariančnú maticu. Podobne ako v prípade logitového modelu, pri dostatočne veľkom výbere pozorovaní majú odhady parametrov $\boldsymbol{\beta}$ získané metódou MMV asymptotickej normality a konzistencie (Mittelhammer, et al., 2000).

Marginálny efekt u probitového modelu

Marginálny efekt niektorej zo **spojitých** vysvetľujúcich premenných X_{ik} na odhadnutej pravdepodobnosti p_i probitového modelu vyjadríme vzťahom

$$\frac{\partial p_i}{\partial X_{ik}} = b_k \frac{1}{\sqrt{2p}} exp\left[-1/2(\mathbf{x}_i^T\boldsymbol{\beta})^2\right] = b_k\ f(\mathbf{x}_i^T\boldsymbol{\beta}),, \tag{4.51}$$

kde $f(\mathbf{x}_i^T \boldsymbol{\beta})$ je funkcia hustoty štandardného normálneho rozdelenia.

Pre **umelé nula-jednotkové** vysvetľujúce premenné je výpočet marginálnych efektov analogický ako v prípade logitového modelu, daných výrazom (4.42).

Odhadnuté modely logitového a probitového modelu dávajú veľmi podobné výsledky, avšak pri ich porovnávaní treba brať do úvahy rozdiely v pravdepodobnostných funkciach oboch modelov. Rozptyl náhodnej zložky u_i logistického rozdelenia logitového modelu je rovný $p^2/3$ avšak u probitového modelu s normálnym rozdelením je rozptyl náhodnej zložky u_i rovný hodnote 1. Znamená to, že kvôli porovnateľnosti oboch modelov, sú vynásobené odhadnuté regresné koeficienty $\hat{b}_i$ logitového modelu hodnotou $\sqrt{3}/p = 0{,}5513$.

Amemiya (1981) však doporučuje násobiť odhadnuté regresné koeficienty $\hat{b}_i$ logitového modelu hodnotou 1/1,6 = 0,625 namiesto hodnoty $\sqrt{3}/p = 0{,}5513$. Uvádza, že táto transformácia umožňuje bližšiu aproximáciu medzi logistickým rozdelením a distribučnou funkciou štandardného normálneho rozdelenia v okolí $p_i = 0{,}5$.

Okrem toho tiež navrhuje porovnanie lineárneho pravdepodobnostného a logitového modelu (LM) pre odhad konštanty nasledovne

$$\hat{b}_{\mathrm{LPM}} \cong 0{,}25\hat{b}_{\mathrm{LM}} + 0{,}5 \tag{4.52}$$

a pre odhad regresných koeficientov

$$\hat{b}_{\mathrm{LPM}} \cong 0{,}25\hat{b}_{\mathrm{LM}}\,. \tag{4.53}$$

Ďalej uvádza aj možnosť porovnania lineárneho pravdepodobnostného a probitového modelu (PM) v prípade konštanty

$$\hat{b}_{\mathrm{LPM}} \cong 2{,}5\hat{b}_{\mathrm{PM}} - 1{,}25 \tag{4.54}$$

a pre regresné koeficienty

$$\hat{b}_{\mathrm{LPM}} \cong 2{,}5\hat{b}_{\mathrm{PM}}. \tag{4.55}$$

Medzi ďalšie alternatívne spôsoby porovnania modelov diskrétnej voľby patria:

- výpočet sumy štvorcových odchýlok z predikovaných pravdepodobností,
- porovnanie percent správne predikovaných hodnôt,
- porovnanie parciálnych derivácií pravdepodobností vzhľadom ku konkrétnej vysvetľujúcej premennej.

4.5 Testovanie hypotéz u modelov diskrétnej voľby

Modely diskrétnej voľby je pomerne zložité testovať bežnými postupmi používanými v ekonometrickej analýze. Pri testovaní hypotéz o štatistickej významnosti jednotlivých odhadnutých parametrov premenných použijeme **Waldov test**. Táto testovacia štatistika má normálne rozdelenie a je podielom odhadnutého koeficientu jeho štandardnou chybou

$$W = \frac{\hat{b}}{s_{\hat{b}}} \approx N(0{,}1). \tag{4.56}$$

Testovanie modelu ako celku predstavuje test hypotézy, že sa všetky odhadnuté parametre daného moelu rovnajú nule.

U modelov diskrétnej voľby pri testovaní hypotézy o význame modelu ako celku využívame test pomerom vierohodností alebo test Langrangeovho multiplikátora.

Test pomerom vierohodností

Testovacia štatistika *LR* má tvar

$$LR = -2 \ln L_0 \,(-2 \ln L_{MAX}), \quad 0 \le LR \le \infty, \tag{4.56}$$

kde L_0 je maximum funkcie vierohodnosti L v prípade, keď všetky parametre modelu s výnimkou úrovňovej konštanty položíme rovné nule a L_{MAX} predstavuje maximum funkcie vierohodnosti L vypočítanej vzhľadom ku všetkým parametrom tejto funkcie.

Test Lagrangeovho multiplikátora

Alternatívou k testu pomerom vierohodností je test Lagrangeovho multiplikátora (LM), jeho testovaciu štatistiku vypočítame ako

$$LM = \left(\sum_{i=1}^{n} g_i \, \mathbf{x}_i \right)^T \left[\sum_{i=1}^{n} E[-h_i] \mathbf{x}_i \mathbf{x}_i^T \right]^{-1} \left(\sum_{i=1}^{n} g_i \, \mathbf{x}_i \right), \tag{4.58}$$

kde $E[-h_i]$ je definovaná v (4.38) u logitového modelu a v (4.46) u probitového modelu. Oba uvedené testy prepokladajú χ^2 rozdelenie s počtom stupňov voľnosti rovnému počtu odhadnutých parametrov bez konštanty.

Testy špecifikácie modelov binárnej voľby

Doteraz bolo predpokladané, že model binárnej voľby je správne špecifikovaný a odhad parametrov ako aj testovanie hypotéz neprinášalo žiadne problémy. V skutočnosti však existujú situácie, keď pri formulácii modelu vychádzame z jedného alebo viacerých nesprávnych predpokladov. Medzi najdôležitejšie problémy špecifikácie modelov binárnej voľby patrí vynechanie premenných a problém heteroskedasticity (Greene, 2003). **Vynechanie relevantných premenných** v klasickom lineárnom regresnom modeli spôsobí, že odhadová funkcia **b** je síce nevychýlená a konzistentná, ale neefektívna, takže nie je vhodné použitie klasických metód odhadu. U logitového a probitového modelu vynechanie premennej hoci aj neskorelovanej s ostatnými vysvetľujúcimi premennými modelu vedie k nekonzistentnému odhadu. Ďalším problémom je **výskyt heteroskedasticity**, ktorý rovnako ako u vynechania premenných vedie k nekonzistentnej odhadovej funkcii, najmä u probitového modelu. Na testovanie oboch problémov sú využívané vyššie uvedené testy pomerom vierohodností a Lagrangeovho multiplikátora.

4.6 Testovanie zhody modelu s údajmi

Keďže v modeloch binárnej diskrétnej voľby sú skutočné hodnoty vysvetľovanej premennej Y nula-jednotkové a jej predikované hodnoty Y sú podmienené pravdepodobnosti, pri použití a interpretácii štandardnej štatistiky koeficientu viacnásobnej determinácie R^2 ako miery celkovej významnosti modelu nastávajú určité problémy, uvedené v druhej kapitole. Z uvedených dôvodov boli pre modely diskrétnej voľby navrhnuté modifikované charakteristiky, ktoré sa používajú na overenie ich zhody s dátami podobne ako testovacia štatistika R^2. Tieto charakteristiky sú rovnaké v prípade lineárneho regresného modelu a platia aj pre modely mutlinomickej voľby.

Testy zhody založené na sume štvorcov reziduí

Štandardný koeficient determinácie R^2 v prípade klasického lineárneho regresného modelu, odhadnutého MNŠ alebo MZNŠ, je rovný

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \right] \tag{4.59}$$

Pre vysvetľovanú dichotomickú náhodnú premennú y_i platí

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = n_1 - n\left(\frac{n_1}{n}\right)^2 = \frac{n_1 n - n_1^2}{n} = \frac{n_1(n - n_1)}{n} = \frac{n_1 n_2}{n},$$

kde n_1 je početnosť alternatívy 1 a $n_2 = (n - n_1)$ je početnosť opačnej alternatívy.

Efron (1978) navrhol nasledovnú štatistiku vhodnú na posudzovanie zhody modelu diskrétnej voľby s dátami

$$R_E^2 = 1 - \frac{n}{n_1 n_2} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.60}$$

Amemyia (1981) navrhuje nahradiť sumu štvorcov reziduí vo vzorci (4.60) výrazom

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i(1-\hat{y}_i)}$$

Testy zhody založené na vierohodnostnom pomere

Pre klasický lineárny regresný model v tvare

$$y_i = b_0 + \sum b_i x_i + u_i \ , \qquad u_i \sim \mathrm{IN}(0,\sigma^2), \tag{4.61}$$

odhadnutý MMV, predpokladajme L_{MAX} maximum funkcie vierohodnosti L vypočítanej vzhľadom ku všetkým parametrom tejto funkcie a L_0 je maximum funkcie vierohodnosti L v prípade, keď všetky parametre model s výnimkou úrovňovej konštanty položíme rovné nule. Potom mieru celkovej zhody s dátami definujeme ako

$$R^2 = 1 - \left(\frac{L_0}{L_{MAX}} \right)^{2/n}. \tag{4.62}$$

V prípade logitového a probitového modelu môžeme formulovať obdobnú mieru celkovej zhody s dátami, pričom platí, že vierohodnostná funkcia nadobúda maximálnu hodnotu jedna, teda

$$L_0 \leq L_{\mathrm{MAX}} \leq 1.$$

Uvedený vzťah môžeme alternatívne vyjadriť ako

$$L_0 \leq \frac{L_0}{L_{MAX}} \leq 1,$$

alebo

$$L_0^{2/n} \leq 1 - R^2 \leq 1 \tag{4.63}$$

alebo

$$0 \le R^2 \le 1 - L_0^{2/n}.$$

Ako ďalšiu mieru zhody modelu diskrétnej voľby s dátami navrhli Cragg a Uhler (1970) na základe (4.61) tzv. **pseudoštatistiku *R²*** v tvare

$$R_{CU}^2 = \frac{L_{MAX}^{2/n} - L_0^{2/n}}{(1 - L_0^{2/n}) / L_{MAX}^{2/n}}, \tag{4.64}$$

ktorá nadobúda hodnoty v intervale $\langle 0,1 \rangle$.

McFadden (1974) navrhol inú modifikáciu R^2, tzv. **index podielu vierohodnosti**, ktorý tiež nadobúda hodnoty v intervale $\langle 0,1 \rangle$, v tvare

$$R_M^2 = 1 - \frac{ln L_{MAX}}{ln L_0}, \tag{4.65}$$

pričom jeho hodnota v rozmedzí intervalu 0,2 až 0,4 je považovaná za vysoko uspokojujúcu.

Cox a Snell (1989) definovali modifikovaný R^2 ako

$$R_{CS}^2 = 1 - \left(\frac{L_0}{L_{MAX}} \right)^{2/n}, \tag{4.66}$$

ktorý však aj pri perfektnej predpovedi zhody dát s modelom dosiahne maximálnej hodnoty 0,75.

Tento problém vyriešil o pár rokov neskôr Nagelkerke (1991), keď upravil R_{CS}^2 na tvar

$$\overline{R}_N^2 = \frac{R_{CS}^2}{R_{MAX}^2}, \tag{4.67}$$

kde $R_{MAX}^2 = 1 - L_0^{2/n}$ a je vlastne rovný výrazu (4.63).

Testy zhody založené na podiele správnych predpovedí

Po odhade binárneho modelu, kde jeho vysvetľujúca premenná nadobúda hodnoty nula alebo jedna, odhadnuté hodnoty $\hat{y}_i$ roztriedime tak, že ***i***-te pozorovanie bude patriť do skupiny 1 v prípade, že $\hat{y}_i \leq 0{,}5$ a do skupiny 2, ak $\hat{y}_i > 0{,}5$. Potom môžeme určiť počet správnych predpovedí. Definujme predikovanú hodnotu $\hat{y}_i^*$ ktorá je rovná

$$\hat{y}_i^* = \begin{cases} 1 & \text{ak } \hat{y}_i > 0{,}5 \\ 0 & \text{ak } \hat{y}_i \leq 0{,}5 \end{cases}.$$

Takže pseudoštatistiku R^2 definujeme ako

$$R^2_{PSP} = 1 - \frac{\text{počet správnych predpovedí}}{\text{celkový počet pozorovaní}} \tag{4.68}$$

Uvedená štatistika je dobre použiteľná vo všetkých problémoch, nemusí však mať dostatočnú diskriminačnú schopnosť. V mnohých aplikáciách modelov binárnej voľby majú napozorované údaje nerovnomerné rozdelenie medzi dvoma možnými výstupmi. Ak napríklad 80% pozorovaní má rovnaký výstup (čiže je napríklad rovné hodnote jedna), potom nie je problém predikovať 80% pozorovaní správne, čo môže byť zložitejšie v prípade rovnomerného rozdelenia pozorovaní. Ďalej sa táto štatistika nezdá byť nápomocná pri definovaní rozdielov medzi jednotlivými typmi modelov diskrétnej voľby.

Medzi ďalšie pseudoštatistiky patria R^2 autorov Aldricha a Nelsona (1984) a jeho modifikácia Hagleho a Mitchella (1992), ktoré sa nachádzajú napríklad v programe SPSS, alebo Veall a Zimmermann (1992), prípadne Zavoina a McKelvey (1975).

Ďalšími testami vypovedacej schopnosti modelu binárnej voľby je **ROC krivka** (napr. Hebák, 2005). Predstavuje prehľadnú grafickú metódu porovnania zhody modelu s údajmi, jej hodnotovým vyjadrením je ***c*-štatistika**. Ak sa jej hodnota blíži k jednej, znamená to vysokú diskriminačnú schopnosť modelu.

Ak je hodnotený model diskrétnej voľby s aspoň jednou spojitou vysvetľujúcou premennou, je možné na hodnotenie kvality modelu použiť **Hosmer-Lemeshow test** dobrej zhody (Hosmer, Lemeshow, 2000), použitie ktorého však vyžaduje dostatočne veľký súbor pozorovaní. Dátový súbor je rozdelený do niekoľkých skupín (u netriedených dát to je desať skupín, u triedených ich môže byť menej, v každej skupine by malo byť minimálne päť pozorovaní). V každej z týchto skupín zisťujeme, či daný stav nastal alebo nie. Testovacia štatistika má χ^2 rozdelenie s počtom stupňov voľnosti rovnom počtu vytvorených skupín mínus dva.

4.7 Výber vhodného modelu diskrétnej voľby

Zahrnutie ďalších vysvetľujúcich premenných do modelu so sebou prináša komplikácie spojené hlavne s prílišnou zložitosťou modelu a s následnou interpretáciou odhadnutých parametrov. Cieľom je nájsť model, ktorý pri čo najmenšom počte nezávislých premenných dosahuje čo najlepšej zhody modelu s dátami. Najznámejšími štatistikami sú Akaikeho, Schwarzovo a Hannan-Quinnovo informačné kritérium.

Akaikeho informačné kritérium

$$AIC = \frac{-2L_{MAX} + 2p}{N}, \tag{4.69}$$

kde

L_{MAX} – maximum funkcie vierohodnosti L vypočítanej vzhľadom ku všetkým parametrom tejto funkcie

p – počet parametrov odhadovaného modelu,

N – počet pozorovaní

Schwarzovo informačné kritérium

$$SIC = \frac{-2L_{MAX} + p\ln(N)}{N} \tag{4.70}$$

Hannan-Quinnovo informačné kritérium

$$HQIC = \frac{-2L_{MAX} + 2p\ln(\ln(N))}{N} \tag{4.71}$$

Za najlepší je považovaný model s najnižšími hodnotami informačných kritérií. Okrem uvedených kritérií sú často používanými štatistikami **Pearsonova štatistika** a **štatistika G** (napr. Hebák, 2005).

5 MODELY MULTINOMICKEJ DISKRÉTNEJ VOĽBY

Mnohé diskrétne závislé premenné sú dichotomickej povahy, stretávame sa však aj s prípadmi, kedy táto premenná nadobúda tri a viac alternatív, čiže individuálny ekonomický subjekt sa rozhoduje pre jednu z viac ako dvoch už dopredu známych variant. Príkladom môže byť rozhodovanie študenta o voľbe študijného odboru či rozhodnutie domácnosti o nákupe z niekoľkých dopredu vytipovaných predmetov dlhodobej spotreby. Tieto modely sú nazývané modelmi viacnásobnej (existuje viac alternatív) alebo **multinomickej** voľby.

V závislosti na charaktere odozvy (ohlasu) diskrétnej závislej premennej sú rozlišované dva typy modelov multinomickej voľby:

- **modely s usporiadanými odozvami**, (tzv. „ordered responses“) – príkladom môže byť vyjadrenie spokojnosti spotrebiteľa s kvalitou určitého výrobku alebo služby. Kvalitu môže považovať za výbornú, veľmi dobrú, dobrú, vyhovujúcu alebo nevyhovujúcu, čiže existuje päť možných odpovedí, ktoré sa evidentne dajú usporiadať prirodzeným spôsobom do určitej stupnice alebo škály. Ďalším príkladom je rating klientov banky.
- **modely s neusporiadanými odozvami**, (tzv. „unordered responses“) – ako možné neusporiadané odozvy slúži príklad dopravy individuálneho subjektu do práce, kde medzi nimi môžu byť: chôdza, bicykel, autobus či auto (vlastné, prípadne jazda so spolupracovníkom). Z uvedeného príkladu vyplýva, že je problematické uvedené možnosti vopred usporiadať logickým spôsobom.

Ako už bolo naznačené, s členením modelov multinomickej voľby úzko súvisí problém špecifikácie zodpovedajúcej **metriky**. Ak si napríklad subjekt volí z troch vytipovaných možností nákupu predmetu dlhodobej spotreby, definujeme diskrétnu vysvetľovanú premennú $Y = 0$ pri nákupe predmetu typu A, pri zvolení predmetu typu B to je $Y = 1$ a pri predmete typu C je diskrétna vysvetľovaná premenná $Y = 2$. Pritom musí platiť jedna z vlastností preferencií a to, že predmet typu C musí vyjadrovať dvojnásobnú preferenciu podľa dopredu zvoleného kritéria v porovnaní s predmetom dlhodobej spotreby typu B. U modelov binárnej voľby možno použiť umelé nula-jednotkové premenné, ktoré kvantifikujú **kvalitatívne rozdiely** medzi premennými, u modelov multinomickej voľby je potrebné definovať tieto umelé premenné tak, aby indikovali kvantitatívne rozdiely každej z uvedených variant. Ak nie je možné špecifikovať metriku tak, aby spĺňala uvedené požiadavky, nie je možné pri odhade parametrov modelov multinomickej voľby použiť MNŠ.

V oblasti výskumu verejnej mienky sa pomerne často stretávame so znakmi, ktoré nie sú vždy objektívne pozorovateľné, t.j existujú iba vo vedomí dotazovaných osôb (názory, postoje, hodnoty). Tieto znaky sú ťažko merateľné a ich zjednodušenie spočíva vo vyjadrení skúmaného javu, ktorý obsahuje množstvo znakov rôzneho druhu a obsahu, **jednou premennou** a jeho premietnutí na určitú slovne, číselne alebo graficky vyjadrenú stupnicu – **škálu** (napr. Pecáková a kol., 2004).

Najjednoduchšou škálou je škála **nominálna**, ktorá pri porovnávaní umožňuje zaznamenať iba zhodu alebo rozdiel medzi pozorovanými jednotkami v sledovanom znaku. U **ordinálnej** škály sú jednotlivé kategórie usporiadané jednoznačne a je teda umožnené ich porovnávanie. Kardinálnu škálu členíme na **intervalovú** a **pomerovú**. Uvedeným škálam zodpovedá definovanie typu premenných, podrobne uvedených tretej kapitole.

Jednorozmerné škálovanie

Využíva škálovacie postupy založené na

a) vzájomnom porovnávaní jednotiek vzhľadom k sledovanému znaku. Medzi tieto postupy môžeme zaradiť metódu párových porovnávaní, zlomkové škály, škály konštantného súčtu, apod.,
b) ich jednotlivom hodnotení, medzi ktoré patrí grafická škála a bodovacie (známkovacie) škály. Medzi najznámejšie bodovacie metódy patrí (Powers, 2000):
   - **celočíselné bodovanie** – jedna z najjednoduchších metód, pomocou ktorej odpovede na položenú otázku bodujeme celočíselne, predpokladáme však rovnakú vzdialenosť medzi jednotlivými odpoveďami. Obyčajne začíname číslom jedna a o túto hodnotu zvyšujeme bodovanie v prípade prírastku v odpovedi. V prípade požitia Likertovej metódy u odpovedí ich kódujeme v tvare (1, 2, 3, 4, 5). Jej výhodou je jednoduchá interpretácia.
   - **bodovanie stredovým bodom** („midpoint scoring“) – táto metóda sa používa najmä u intervalovej premennej, ako je napríklad mesačný príjem domácnosti (menej ako 10 000 Kč, 10 000 až 20 000 Kč, 20 000 Kč a viac) a každému intervalu príradíme hodnotu, ktorá sa nachádza práve uprostred. Nevýhodou tejto metódy je, že často podhodnocuje dáta a často je zhora neohraničená (20 000 Kč a viac).

- **normálna bodová tranformácia** – uskutočníme výpočet relatívnej početnosti u každej kategórie, ďalej určíme kumulatívnu početnosť a 50% percentil kumulatívnej početnosti pre každú kategóriu. Nakoniec túto hodnotu transformujeme na *z*-skóre štandardizovaného normálneho rozdelenia, kde $z = \Phi^{-1}(p)$. Jej nevýhodou je jej aplikácia iba na konkrétny dátový súbor a bez znalosti postupu pri transformácii aj jej zložitá interpretácia.
- **bodovanie s dodatočnými informáciami –** je založené na využití dodatočných informácií z iných premenných súboru alebo z iného zdroja dát. Príkladom je socioekonomický index skonštruovaný Duncanom v roku 1961, ktorý kombinuje štyri priamo merateľné premenné (vzdelanie otca, vzdelanie matky, zamestnanecký status otca a príjmy oboch rodičov).
- **Osgoodova metóda sémantického diferenciálu**. Spočíva vo viackriteriálnom hodnotení určitej situácie a hlavnou podmienkou je, aby hodnotenie každého zo stimulov (u výrobku napríklad cena, vzhľad, chuť, apod.) bolo respondetovi ponúknuté v rovnako orientovanej škále (päť, sedem a desaťbodové škály, príp. škály špeciálne skonštruované).
- **Likertova metóda**, ktorou štandardizujeme kategórie odpovedí v dotazníkových prieskumoch. Vyznačujú sa pevnou alternatívnou otázkou, v odpovedi na ktorú respondent určuje, do akej miery súhlasí alebo nesúhlasí s daným výrokom (možnosti: silne nesúhlasím, nesúhlasím, mám neutrálny postoj k otázke, súhlasím, silne súhlasím).

Viacrozmerné škálovanie

Je použiteľné v prípade, kde je nemožné použiť zjednodušenie prevedenia sledovaného znaku iba do jednej dimenzie. Je to náročná metóda a jej použitie je nemysliteľné bez využitia výpočtovej techniky. Modul viacrozmerného škálovania je obsiahnutý takmer vo všetkých štatistických programoch ako *SPSS* (pod názvom *ALSCAL*), *SAS*, *S-plus* alebo v programe *Statistica - Vícerozměrné průzkumné techniky*, kde umožňuje analyzovať matice podobností, rozdielností alebo korelácií medzi premennými pri špecifikácii až 9 dimenzií.

5.1 Model neusporiadanej multinomickej voľby

Formuláciu modelu neusporiadanej viacnásobnej voľby a postup pri jeho odhade uvádza vo svojej práci napr. Hušek (1995) na príklade, kedy volič má odovzdať svoj hlas jednému z K kandidátov.

Pravdepodobnosť zvolenia k-teho kandidáta označíme ako P_k, pričom platí

$$\sum_{k=1}^{K} P_k = 1.$$

Ďalej nech výsledný vektor $y = (Y_1, Y_2, ..., Y_k)$, kde $Y_k = 1$, ak je vybraný práve k-ty kandidát, takže zostávajúce zložky sú nulové a

$$\sum_{k=1}^{K} Y_k = 1.$$

Ak skúmame volebné preferencie napríklad u M vybraných voličov, potom vierohodnosť, že

$$y_1 = (Y_{11}, Y_{12}, ..., Y_{1K}),..., y_M = (Y_{M1}, Y_{M2}, ..., Y_{MK}),$$

je daná funkciou vierohodnosti v tvare

$$e^L = \prod_{i=1}^{M} P_{i1}^{Y_{i1}} P_{i2}^{Y_{i2}} ... P_{iK}^{Y_{iK}}, \qquad (5.1)$$

pričom P_{iK} je pravdepodobnosť, že i-ty volič odovzdá svoj hlas K-temu kandidátovi.

Všeobecne možno povedať, že cieľom modelovania vyššie popísaného rozhodovacieho procesu je stanovenie pravdepodobností P_{iK}, ktoré maximalizujú hodnotu (5.1), ako funkcia charakteristík možných variant viacnásobnej voľby a atribútov individuálnych subjektov, ktorí sa rozhodujú. Je to možné dosiahnuť dvoma prístupmi, založenými na analýze tzv. podmienených logitov, príp. probitov, ktoré sa odlišujú stochastickou špecifikáciou pravdepodobností P_{iK}. Multinomický logitový model predpokladá Weibullovo rozdelenie a multinomický probitový model je založený na viacrozmernom normálnom rozdelení.

5.1.1 Multinomický model podmienených logitov

Východiskom pri konštrukcii modelov všeobecnej diskrétnej voľby je opäť teória užitočnosti, ktorú rozpracoval McFadden (1973), kde predpokladá existenciu priemerného rozhodujúceho sa individuálneho subjektu, t.j. takého, ktorý má priemerné preferencie týkajúce sa všetkých možných atribútov. Predpokladom je modelovanie rozhodovacieho procesu i-teho spotrebiteľa, ktorý si vyberá z K možných variant, ako $\mathbf{x}_{ik}$ je označený vektor K×1 pozorovaní premenných, ktoré sú funkciami charakteristík ľubovoľnej k-tej varianty u i-teho spotrebiteľa. Ďalej je definovaný vektor $\mathbf{s}_i$ ako vektor jednotlivých atribútov, ako napríklad vek, dosiahnuté vzdelanie alebo pohlavie i-teho spotrebiteľa. Všetky uvažované charakteristiky je možné zhrnúť do jedného vektora, napríklad $\mathbf{z}_{ik}$

= ($\mathbf{x}_{ik}$, $\mathbf{s}_i$), takže funkciu užitočnosti priemerného *i*-teho spotrebiteľa, ktorý sa rozhodol pre *k*-tu alternatívu za predpokladu linearity, môžeme zapisať v tvare

$$\overline{U}_{ik} = \mathbf{z}_{ik}^T \boldsymbol{\beta}, \tag{5.2}$$

kde $\boldsymbol{\beta}$ je K×1 vektor neznámych parametrov a je rovnaký pre celý základný súbor, v našom prípade pre všetkých spotrebiteľov. Ak nie je splnená táto požiadavka, je nutné základný súbor rozdeliť na skupiny a funkciu užitočnosti odhadnúť zvlášť pre každú z nich.

Kvôli rozlíšeniu konkrétneho spotrebiteľa od priemerného je formulovaná stochastickú funkciu užitočnosti

$$U_{ik} = \mathbf{z}_{ik}^T \boldsymbol{\beta} + \mathbf{u}_{ik}, \tag{5.3}$$

kde náhodná zložka $\mathbf{u}_{ik}$ reprezentuje vplyvy nepozorovateľných faktorov v každom výbere, pôsobenie náhodných rozdielov v správaní sa spotrebiteľov a dôsledky existencie chýb merania.

Za predpokladu, že individuálny subjekt, v našom prípade každý jednotlivý spotrebiteľ, sa rozhodne pre takú variantu, ktorá maximalizuje stochastickú funkciu užitočnosti (5.3), platí pre pravdepodobnosť zvolenia práve *k*-tej varianty (k = 1,…, K) *i*-tym spotrebiteľom vzťah

$$\begin{aligned} P_{ik} &= pr(U_{ik} > U_{ij}) \\ &= pr(\overline{U}_{ik} + u_{ik} > \overline{U}_{ij} + u_{ij}) \\ &= pr(u_{ij} - u_{ik} < \overline{U}_k - \overline{U}_j), \qquad \forall\, j \neq k, \end{aligned} \tag{5.4}$$

pričom pre zjednodušenie platí $\overline{U}_k = \overline{U}_{ik}$, alebo $\overline{U}_j = \overline{U}_{ij}$.

Podľa McFaddena (1973) nutnou a postačujúcou podmienkou popisu náhodnej užitočnosti je Weibullovo rozdelenie vzájomne nezávislých náhodných zložiek $\mathbf{u}_{ik}$, pretože rozdiel dvoch ľubovoľných náhodných zložiek s Weibullovým rozdelením má logistickú KDF a z nej vychádza aj konštrukcia multinomického logitového modelu.

Multinomický logitový model, ktorý vyjadruje pravdepodobnosť zvolenia k-tej varianty (alternatívy) z K možností i-tym individuálnym subjektom, môžeme zapísať v nasledovnom všeobecnom tvare

$$P_{ik} = \frac{e^{\mathbf{z}_{ik}^T \boldsymbol{\beta}}}{\sum_{k=1}^{K} e^{\mathbf{z}_{ik}^T \boldsymbol{\beta}}} . \tag{5.5}$$

Už na prvý pohľad je zrejmá analógia s binárnym logitovým modelom (4.32), ak je dosadený do výrazu (5.5) pre K = 2, takže napríklad pre pravdepodobnosť výberu prvej alternatívy z dvoch možných je rovná

$$P_{i1} = \frac{e^{\mathbf{z}_{i1}^T \beta}}{e^{\mathbf{z}_{i1}^T \beta} + e^{\mathbf{z}_{i2}^T \beta}} = \frac{1}{1 + e^{(\bar{U}_{i2} - \bar{U}_{i1})}} , \tag{5.6}$$

kde bol použitý vzťah (5.2).

Pre relatívnu šancu (vyjadrenú podielom pravdepodobností) multinomického logitového modelu, že i-ty subjekt dá prednosť k-tej variante pred j-tou platí

$$\frac{P_{ik}}{P_{ij}} = \frac{e^{\bar{U}_{ik}}}{e^{\bar{U}_{ij}}}$$

alebo

$$ln\left[\frac{P_{ik}}{P_{ij}}\right] = \bar{U}_{ik} - \bar{U}_{ij} = (\mathbf{z}_{ik} - \mathbf{z}_{ij})^T \boldsymbol{\beta}, \tag{5.7}$$

kde logaritmus relatívnej šance zvolenia jednej alternatívy oproti výberu inej z K možných, vyjadrený ako lineárna funkcia všetkých charakteristík jednotlivých rozhodnutí a atribútov individuálnych subjektov, sa nazýva **podmienený logit**. Z (5.7) vyplýva, že relatívne šance konkrétnej voľby nie sú ovplyvnené ostatnými variantami, pretože i multinomický logitový model je založený na porovnávaní iba dvojíc možných variant, takže si zachováva binárnu podstatu. Uvedená vlastnosť (5.7) je známa ako **nezávislosť nepodstatných variant** (NNV).

Nezávislosť nepodstatných variant

Vlastnosť nezávislosti nepodstatných (irelevantných) variant (independence of irrelevant alternatives) po prvýkrát použil Arrow v roku 1951 ako jednu z podmienok vo svojom **teoréme nemožnosti**. Uvádza, že sociálna preferencia medzi akýmikoľvek dvoma alternatívami závisí jedine na individuálnom hodnotení jedinca a na ničom inom. Uvedená vlastnosť predstavuje určité obmedzenie v použití multinomického modelu podmienených logitov v prípade, ak dve alebo viac z K možných variant sú značne substitučného charakteru, takže sa v podstate od seba nelíšia.

Vlastnosť NNV je ilustrovaná na príklade volieb, ktorý je obmenou príkladu červený/modrý autobus (McFadden, 1974). Predpokladom je, že individuálny subjekt sa pri voľbách rozhoduje medzi zvolením kandidáta A, kandidáta B alebo nebude voliť vôbec (N). Je zrejmé, že subjekt sa rozhodne nevoliť v prípade, ak

$$U(N) > \max \{U(A),U(B)\}.$$

Uvažujme, že do volieb vstúpi ďalší kandidát C s úplne identickými vlastnosťami ako má kandidát B. Vlastnosť NNV predpokladá, že relatívna pravdepodobnosť alternatív N a B je nezávislá na zahrnutí ďalšej identickej varianty do množiny alternatív. V skutočnosti sa však tieto pravdepodobnosti zmenia (pôvodní voliči kandidáta B sa budú teraz rozhodovať medzi alternatívami B a C, čím dôjde ku zmene relatívnej pravdepodobnosti alternatív N a B. Uvedený príklad je prípadom tzv. úplnej substituovateľnosti (nahraditeľnosti), ktorá sa v praxi príliš často nevyskytuje. Uvedenú vlastnosť NNV testujeme viacerými testami.

Hausman-McFaddenov test založený na podmnožine alternatív

Najviac rozšíreným testom vlastnosti NNV u podmieného logitového modelu viacnásobnej voľby je tzv. **Hausman-McFaddenov test** (McFadden, Tye a Train, 1978 alebo Hausman a McFadden, 1984). Základnou ideou tohto testu je testovanie opačného dôsledku vlastnosti nezávislosti nepodstatných variant a vyžaduje odhad parametrov dvoch modelov. Znamená to, že ak v modeli platí vlastnosť NNV, tak rozdiely v odhadnutých modeloch by nemali byť štatisticky významné. Obvyklou implikáciou tohto testu je existencia dvoch alternatív, napríklad auta a autobusu v prípade voľby dopravného prostriedku, kde pridanie ďalšej alternatívy by nemalo zmeniť pomer pravdepodobností dvoch pôvodných alternatív. Uvedený test je založený na eliminovaní jednej alebo viacerých alternatív z celej množiny alternatív tak, že základné výberové správanie z obmedzeného výberu je vedené vlastnosťou NNV. Odhadnuté parametre modelov obmedzeného a aj neobmedzeného výberu označíme ako $\hat{b}_O$ a $\hat{b}_N$. Je testovaná nulová hypotéza o rovnosti parametrov modelu $\hat{b}_O$ a $\hat{b}_N$. Testovacia štatistika q, ktorá má asymptotické χ^2 rozdelenie, ak je v modeli potvrdená vlastnosť NNV a s počtom stupňov voľnosti rovnom rozdielu v počte parametrov oboch modelov, je definovaná ako

$$q = (\hat{b}_N - \hat{b}_O)(V_O - V_N)^{-1}(\hat{b}_N - \hat{b}_O), \tag{5.8}$$

kde V_O a V_N sú variačné matice odhadnutých parametrov obmedzeného a neobmedzeného multinomického modelu podmienených logitov. Vysoká hodnota testovacej štatitiky q vedie k zamietnutiu nulovej hypotézy o rovnosti parametrov.

McFaddenov test vynechaných premenných

Tento test sa používa na testovanie vlastnosti NNV u **hniezdového logitového modelu** multinomickej voľby, ktorý je viac rozšírený než multinomický logitový model. Testuje sa tu nulová hypotéza o nulových interakčných efektoch, čiže platí existencia multinomického logitového modelu a zároveň aj vlastnosť NNV.

Je uskutočnený odhad parametrov $\boldsymbol{\beta}$ základného multinomického logitového modelu, ktorý zahŕňa všetky pozorovania. Je predpokladaná špecifická množina alternatív označenú ako A. Vytvoríme nové premenné v jednej z nasledujúcich troch foriem:

- Ak x_{in} sú premenné základného logitového modelu multinomickej voľby, tak definujeme novú premennú v tvare

$$z_{in} = \begin{cases} x_{in} - \left(\sum_{j \in A} P_{jn} x_{jn} \right) / \left(\sum_{j \in A} P_{jn} \right) & \text{ak } i \in A, \\ 0 & \text{ak } i \notin A, \end{cases} \tag{5.9}$$

kde pravdepodobnosť P_{jn} je vypočítaná zo základného odhadnutého modelu.

- Ak $V_{in} = x_{in}\boldsymbol{\beta}$ je reprezentatívna užitočnosť základného modelu, tak definujeme novú premennú ako

$$z_{in} = \begin{cases} V_{in} - \left(\sum_{j\in A} P_{jn} V_{jn}\right) / \left(\sum_{j\in A} P_{jn}\right) & \text{ak } i \in A, \\ 0 & \text{ak } i \notin A. \end{cases} \tag{5.10}$$

- Nová premenná s pravdepodobnosťami P_{in} vypočítaných z odhadov základného modelu

$$z_{in} = \begin{cases} -\log\left(P_{in} / \sum_{j\in A} P_{jn}\right) & \text{ak } i \in A, \\ 0 & \text{ak } i \notin A. \end{cases} \tag{5.11}$$

Testovacia štatistika má asymptotické χ^2 rozdelenie s počtom stupňov voľnosti rovným počtu parametrických obmedzení daných nulovou hypotézou. V prípade, že v modeli platí vlastnosť NNV, tak počet stupňov voľnosti testovacej štatistiky je rovný jednej (McFadden, 1976).

Zo špecifikácie logitového modelu (5.5) vyplýva, že žiadna z K premenných vektoru z_{ik} nie je konštantná pre všetky varianty, pretože príslušný parameter by nebol identifikovateľný. Vzhľadom k (5.7) je jasné, že ak sa rovnajú prvky z_{ij} a z_{ik}, tak príslušná premenná neovplyvňuje šance prednostnej voľby jednej z oboch uvažovaných možností a nemožno odhadnúť jej parameter. Ak je však vektor $\boldsymbol{\beta}$ konštantný ako pre všetky rozhodujúce sa subjekty, tak aj pre každú možnú variantu ich voľby, je potrebné zahrnúť do analýzy podmienených logitov iba tie faktory, ktoré sa menia v závislosti na možných variantách a tým prispievajú k vysvetleniu príčin, prečo konkrétna varianta má väčšiu šancu alebo pravdepodobnosť ako iná. Znamená to, že premenné ako príjem, vek alebo pohlavie, ktorých parametre sú spravidla takmer konštantné pre všetky prípustné varianty všeobecnej voľby, neposkytujú žiadnu informáciu, podstatnú pre modelovaný rozhodovací proces. Na druhej strane premenné ako náklady individuálneho subjektu spojené s konkrétnou variantou prepravy, výnosy vyplývajúce z príslušného rozhodnutia, kvalita či iná výhoda zvolenej varianty, sa obvykle vyznačujú značnou premenlivosťou v závislosti na výsledku voľby a sú prínosom skúmania rozhodujúcich sa subjektov (Boskin, 1974).

Model (5.5) možno modifikovať pre prípad, kedy vysvetľujúce premenné majú rôzny vplyv na relatívne šance voľby jednotlivých variant, čiže vektor parametrov **β** je špecifikovaný variantne a pre konkrétnu k-tu variantu ho značíme ako $\boldsymbol{\beta}_k$. Tak výraz (5.6) je rovný

$$P_{ik} = \frac{e^{z_{ik}^T \boldsymbol{\beta}_k}}{\sum_{k=1}^{K} e^{z_{ik}^T \boldsymbol{\beta}_k}}. \tag{5.12}$$

Relatívna šanca k-tej varianty vzhľadom k j-tej je podľa (5.12) daná

$$\frac{P_{ik}}{P_{ij}} = e^{(z_{ik}^T \boldsymbol{\beta}_k - z_{ij}^T \boldsymbol{\beta}_j)},$$

alebo

$$ln\left[\frac{P_{ik}}{P_{ij}}\right] = z_{ik}^T \boldsymbol{\beta}_k - z_{ij}^T \boldsymbol{\beta}_j, \qquad k \neq j. \tag{5.13}$$

Ak vektory z_{ij} a z_{ik} obsahujú iba premenné, ktoré sú pre všetky varianty multinomickej voľby konštantné, potom $z_{ij} = z_{ik} = z_i$, a pre (5.13) platí

$$\frac{P_{ik}}{P_{ij}} = e^{z_i^T (\boldsymbol{\beta}_k - \boldsymbol{\beta}_j)}, \qquad k \neq j. \tag{5.14}$$

Pri použití normovacieho pravidla sa najčastejšie predpokladá, že parameter $\beta_j = 0$, pričom táto podmienka spolu s K-1 rovnicami (5.14) umožňuje jednoznačné určenie všetkých pravdepodobností voľby i-teho subjektu a zároveň zaručuje, že ich suma je pre každý subjekt rovná jednej. Pre takto špecifikovaný model majú pravdepodobnosti tvar

$$P_{ik} = \frac{e^{z_i^T \boldsymbol{\beta}_k}}{1 + \sum_{k=2}^{K} e^{z_i^T \boldsymbol{\beta}_k}}, \qquad k = 2, 3, ..., K. \tag{5.15}$$

Metódy odhadu multinomického modelu podmienených logitov

Podobne ako v prípade modelu binárnej voľby možno pri dosť veľkom počte pozorovaní v jednotlivých výberových experimentoch použiť na odhad parametrov multinomického logitového modelu MZNŠ (Zellner a Lee, 1965, alebo Theil, 1967). Častejší je však prípad, kedy potrebný počet pozorovaní nemáme k dispozícii, takže logitový model odhadneme pomocou **MMV**, ktorá zaručuje konzistentné odhady jeho parametrov.

Je predpokladané zjednodušenie, že každý individuálny subjekt sa rozhoduje v procese viacnásobnej voľby na základe rovnakého počtu variant a jeho pravdepodobnosti sú určené na základe (5.15), po dosadení (5.15) do funkcie vierohodnosti platí

$$e^L = \prod_{i=1}^{M} \left[\frac{1}{1+\sum_{k=2}^{K} e^{z_i^T \beta_k}} \right] \prod_{i=1}^{M} \prod_{k=2}^{K} exp(\mathbf{z}_i^T \boldsymbol{\beta}_k)^{Y_{ik}} . \tag{5.16}$$

Po zlogaritmovaní vyššie uvedeného výrazu (5.16) dostaneme

$$L = \sum_{i=1}^{M} \left[\sum_{k=2}^{K} Y_{ik} \mathbf{z}_i^T \boldsymbol{\beta}_k - ln(1+\sum_{k=2}^{K} exp(\mathbf{z}_i^T \boldsymbol{\beta}_k)) \right], \tag{5.17}$$

z čoho vyplýva, že funkcia vierohodnosti, prípadne jej logaritmus, sú v parametroch vektoru $\boldsymbol{\beta}_k$ nelineárne, takže na odhad logitového modelu je použitý niektorý z nelineárnych optimalizačných postupov, ako napríklad Newton-Rhapsonova iteračná metóda.

MMV zaisťuje konzistentné a asymptoticky normálne rozdelené odhady, takže je možné stanoviť s dostatočnou presnosťou intervaly spoľahlivosti bodových odhadov a aplikovať testovacie štatistiky, vhodné pre veľký výber pozorovaní.

Alternatívny spôsob testovania významnosti modelu podmienených logitov, ktorý vychádza z teórie informácie, navrhol Theil (1967).

Theil (1967) definuje **informáciu**, ktorá je obsiahnutá v správe ako

$$log\frac{P_1}{P_0}, \tag{5.18}$$

kde P_1 je pravdepodobnosť výskytu udalosti po obdržaní správy a P_0 je pravdepodobnosť výskytu udalosti pred obdržaním správy. Znamená to, čím vyšší je podiel P_1/P_0, tým je dosiahnutá vyššie zmenšenie neurčitosti prípadného výskytu udalosti.

Očakávaná (stredná hodnota) informácie je rozšírením konceptu informácie na situáciu, v ktorej správa obsiahnutá v informácii mení pravdepodobnosť rozdelenia J vzájomne vylučujúcich sa udalostí $\mathbf{y} = (y_1,...,y_J)$. Nech pravdepodobnosť rozdelenia pred a po prijatí správy je $P_0 = (P_{01},...,P_{0J})$ a $P_1 = (P_{11},...,P_{1J})$. Potom očakávaná informácia je

$$\sum_{j=1}^{J} P_{1j}\log(P_{1J}/P_{0j}) \tag{5.19}$$

V modeli multinomickej voľby, správa, že pozorovaná jednotka i má charakteristiku x_i mení pravdepodobnosť pozorovania jednotky, ktorá volí alternatívu j na začiatku $P(y_j)$ ku konci $P(y_j|x_i)$ to je pravdepodobnosť výskytu hodnoty y_j náhodnej veličiny Y za podmienky výskytu hodnoty x_i náhodnej veličiny X. Takže očakávaná (stredná hodnota) informácia modelu má tvar

$$E = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} P(y_j\,|\,x_i)\log\,(P(y_j\,|\,x_i)/\,P(\mathrm{y_j})) \tag{5.20}$$

Multinomický model predikuje pravdepodobnosť $\hat{P}_{ij}$ o $P(y_j|x_i)$. Empirická informácia má podobu

|

$$I(y\,|\,X) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} d_{\mathrm{ij}}\log(\hat{P}_{ij}/P(y_j)), \tag{5.21}$$

kde δ_{ij} je rovné 1, ak *i* zvolí *j* a je rovné 0 v opačnom prípade. Podľa Hausera ak je model správny, potom $\sqrt{n}(I(y\,|\,X) - E(y\,|\,X))$ je asymptoticky normálne rozdelený so strednou hodnotou $E(y|X)$ a rozptylom rovným

$$\mathrm{var}(y\,|\,X) = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{J} P(y_j\,|\,x_i)[\log(P(y_j\,|\,x_i)/\mathrm{P}(y_j))^2\right\}.. \tag{5.22}$$

5.1.2 Štandardný multinomický logitový model

MLM má viacero podôb v závislosti na charakteristike rozdelenia náhodnej zložky modelu. U **štandardného MLM** je pravdepodobnosť zvolenia *k*-tej varianty (alternatívy) z *K* možností ***i***-tym individuálnym subjektom nasledovná

$$P_{ik} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_k}}{\sum_{k=1}^{K} e^{\mathbf{x}_i^T \boldsymbol{\beta}_k}}. \tag{5.23}$$

Znamená to, že v štandardnom MLM modeli sú vysvetľujúce premenné nemenné na výsledných kategóriách a ich parametre sa menia v závislosti na výsledku, v podmienenom multinomickom logitovom modeli sa vysvetľujúce premenné menia v závislosti na možných variantach a ich parametre sú konštantné.

Existuje mnoho metód, ktoré pracujú s týmto predpokladom, uvedieme aspoň niektoré z nich (McFadden, 1973):

- hniezdový multinomický logitový model a špeciálne prípady modelov extrémnych hodnôt,
- multinomický probitový model so špeciálnou faktorovo-analytickou štruktúrou, ktorá umožňuje realizovateľný výpočet numerickej integrácie,
- multinomický probitový model so simulačnými odhadovými funkciami, ktoré zvládnu vysokú dimenziu úlohy,

- zmiešaný multinomický logitový model náhodných koeficientov so simulačnou procedúrou pre náhodné koeficienty.

5.1.3 Multinomický model podmienených probitov

Pri porovnaní s logitovým modelom viacnásobnej voľby, probitový multinomický model nie založený na tak reštriktívnych predpokladoch. Vyžaduje iba splnenie požiadavky, že pre každý rozhodujúci sa individuálny subjekt sú náhodné zložky vo funkcii užitočnosti pre rôzne varianty **viacrozmerne normálne rozdelené**.

Analýza podmienených probitov vychádza z úvahy, že každý indviduálny subjekt volí z radu variant, ktorých hodnotu alebo užitočnosť možno z hľadiska jednotlivých subjektov rozdeliť na fixnú alebo priemernú časť a na náhodnú zložku. Náhodná zložka sa mení v závislosti na nepozorovaných charakteristikách možných variant a na odchýlke individuálnych preferencií od preferencií priemerného subjektu. Tým sa oslabuje predpoklad nezávislosti náhodných zložiek u_{ik} a nahrádza sa reálnejším predpokladom, že pre každý subjekt môže závisieť intenzita korelácie dvoch náhodných zložiek, zodpovedajúcich rôznym možným variantám, na tom, do akej miery ich ten, kto sa rozhoduje, pokladá za podobné. Ak je tento predpoklad zahrnutý do analýzy podmienených probitov, tak pre dva sledované atribúty X_1 a X_2 je možné stochastickú funkciu užitočnosti ***i***-teho subjektu vyjadriť ako

$$\begin{aligned} U(x_i) &= \overline{U}(x_i, s_i) + u_i \\ &= (\overline{b}_1 + b_1)X_{i1} + (\overline{b}_2 + b_2)X_{i2} + u_i^* \\ &= \overline{b}_1 X_{i1} + \overline{b}_2 X_{i2} + b_1 X_{i1} + b_2 X_{i2} + u_i^*, \end{aligned} \tag{5.24}$$

kde

$\overline{U}_i = \overline{b}_1 X_{i1} + \overline{b}_2 X_{i2}$,

$u_i = b_1 X_{i1} + b_2 X_{i2} + u_i^*$,

b_1, b_2, u_i^* sú nezávislé náhodné premenné.

Vzťah (5.23) popisuje model s náhodnými parametrami, v ktorom koeficienty β_1 a β_2 prezentujú individuálnu odchýlku od priemerných preferencií a u_i^* je náhodná chyba.

Rozptyl náhodnej zložky, ktorá zodpovedá voľbe *j*-tej varianty *i*-tym subjektom je rovný

$$\operatorname{var}(u_{ij}) = s_{ij}^2 = s_{b1}^2 X_{1,ij}^2 + s_{b2}^2 X_{2,ij}^2 + s_{wij}^2 , \tag{5.25}$$

kde $s_{b_1}^2$ predstavuje rozptyl preferencií jednotlivých subjektov vzhľadom k premennej X_1.

Podmienenú pravdepodobnosť zvolenia varianty 1 *i*-tym individuálnym subjektom definujeme ako

$$P(y_i = 1 | X_i) = \int_{-\infty}^{\infty} \mathbf{K} \int_{-\infty}^{\infty} 1(u_{ij}^1 < X_{ij}^1 \boldsymbol{\beta}) dF(u_{ij}^1), \qquad \forall j \neq 1. \tag{5.26}$$

Multinomický probitový model nie je tak rozšírený ako logitový model hlavne z dôvodu, ak je počet variant $K > 2$, tak je odhad parametrov pomerne komplikovaný, ako vidíme z (5.26) a vyžduje použitie simulačných techník. Medzi najpoužívanejšie metódy odhadu probitového modelu viacnásobnej voľby patrí Clarkova metóda (Clark, 1961) simulačná metóda momentov (McFadden, 1989), metóda simulovaných skóre (Hajivassiliou a McFadden, 1990) či bayesovská analýza (McCulloch a Rossi, 1994) so zaujímavou aplikáciou spotrebiteľského správania pri nákupe potravín (výber zo šiestich variant) v závislosti na cene.

5.2 Model usporiadanej multinomickej voľby

V prípade, že diskrétne pozorovania majú usporiadaný charakter, hovoríme o **usporiadaných modeloch diskrétnej voľby**. Pri predpoklade prístupu latentnej náhodnej premennej, premenná *Y* nadobúda nasledovné hodnoty

$$\begin{aligned} y_i &= 0 \quad \text{ak } y_i^* \leq 0, \\ &= 1 \quad \text{ak } 0 < y_i^* \leq \mu_1, \\ &= 2 \quad \text{ak } \mu_2 < y_i^* \leq \mu_2, \\ &\vdots \\ &= J \quad \text{ak } \mu_{J-1} \leq y_i^*. \end{aligned}$$

Hodnoty μ predstavujú tzv. **prahové logity** a podmienené pravdepodobnosti sú vypočítané v **kumulatívnom logitovom** modeli multinomickej voľby ako

$$
\begin{aligned}
&P(y_i = 0|\mathbf{x}_i) = \frac{1}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_1)} \\
&P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_2)} - \frac{1}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_1)} \\
&P(y_i = 2|\mathbf{x}_i) = \frac{1}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_3)} - \frac{1}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_2)} \\
&\vdots \\
&P(y_i = J|\mathbf{x}_i) = 1 - \frac{1}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_{J-1})},
\end{aligned}
\tag{5.27}
$$

kde musí platiť

$$0 < \mu_1 < \mu_2 < \ldots < \mu_{J\text{-}1}.$$

Marginálne efekty vysvetľujúcich premenných u zvolenia *k*-tej varianty počítame ako

$$\frac{\partial P(y_i = k|\mathbf{x}_i)}{\partial \mathbf{x}_i} = -\beta_j \left(\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_k)}{(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_k))^2} - \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_{k-1})}{(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta} - \mu_{k-1}))^2} \right). \tag{5.28}$$

Podobne počítame pravdepodobnosti a marginálne efekty u kumulatívneho probitového modelu, kde u náhodnej zložky uvažujeme, že je normálne rozdelená (Greene, 2003).

5.3 Viacrozmerné modely diskrétnej voľby

Viacrozmerné modely obsahujú simultánne závislé endogénne premenné, ktoré sú kvalitatívnej povahy. Nerlove a Press (1973) sa zaoberali viacrozmerným logaritmicko-lineárnym a logistickým modelom. Probitový model tohto typu skúmali v epidemiologickej štúdií Ashford a Sowden (1970), uvažovali dvojrovnicový probitový model. Na túto štúdiu nadviazal Amemiya (1978), kde na rovnaký datový súbor použil odlišnú metódu odhadu. Heckman (1978) sa vo svojej práci zaoberal systémom simultánne závislých rovníc s niektorými alebo všetkými kvalitatívnymi endogénnymi premennými.

Dvojrovnicový probitový model (Greene, 2003) môžeme definovať v tvare latentných náhodných premenných

$$
\begin{aligned}
Y_{1i}^{*} &= \mathbf{x}_{1i}^{T}\boldsymbol{\beta}_1 + u_{1i}, \quad & Y_{1i} &= 1 \quad & &\text{ak } Y_{1i}^{*} > 0, \\
& & Y_{1i} &= 0 & &\text{ak } Y_{1i}^{*} \le 0, \\
Y_{2i}^{*} &= \mathbf{x}_{2i}^{T}\boldsymbol{\beta}_2 + u_{2i}, & Y_{2i} &= 1 & &\text{ak } Y_{2i}^{*} > 0, \\
& & Y_{2i} &= 0 & &\text{ak } Y_{2i}^{*} \le 0, \qquad i = 1,2,\dots,n.
\end{aligned}
\tag{5.29}
$$

V modeli (5.29), ktorý odhadujeme metódou MMV, existujú tri typy pozorovaní s nasledovnými nepodmienenými pravdepodobnosťami

$$
\begin{aligned}
&Y_{2i} = 0 & &P(Y_{2i} = 0 & &\mid \mathbf{x}_{1i}, \mathbf{x}_{2i}) = & &1 - \Phi(\mathbf{x}_{2i}^{T}\boldsymbol{\beta}_2), \\
&Y_{2i} = 0,\ Y_{2i} = 1 & &P(Y_{1i} = 0,\ Y_{2i} = 1 & &\mid \mathbf{x}_{1i}, \mathbf{x}_{2i}) = & &\Phi_2[-\mathbf{x}_{1i}^{T}\boldsymbol{\beta}_1, \mathbf{x}_{2i}^{T}\boldsymbol{\beta}_2, -r), \\
&Y_{2i} = 1,\ Y_{2i} = 1 & &P(Y_{1i} = 1,\ Y_{2i} = 1 & &\mid \mathbf{x}_{1i}, \mathbf{x}_{2i}) = & &\Phi_2[\mathbf{x}_{1i}^{T}\boldsymbol{\beta}_1, \mathbf{x}_{2i}^{T}\boldsymbol{\beta}_2, r),
\end{aligned}
$$

kde

$$\mathrm{cov}(u_{i1}, u_{i2} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \rho.$$

Zovšeobecnením tvaru modelu (5.37) je model

$$Y_{it}^{*} = \mathbf{x}_{it}^{T}\boldsymbol{\beta}_{i} + u_{it}, \qquad Y_{it} = 1 \quad \text{ak } Y_{it}^{*} > 0,$$
$$Y_{it} = 0 \quad \text{ak } Y_{it}^{*} \leq 0, \quad i = 1,2,...,n, \quad t = 1,2,...,T,$$

ktorý vlastne predstavuje **probitový model panelových dát**.

6 VYBRANÉ APLIKÁCIE MODELOV DISKRÉTNEJ VOĽBY

Pri aplikácii modelov diskrétnej voľby je predpokladom uvedená teória. V prvej aplikácii je analyzovaná a predikovaná vybavenosť domácností vybranými predmetmi dlhodobej spotreby, druhá pochádza z finančnej oblasti a je venovaná analýze kreditného skórovania. Obe aplikácie vychádzajú z **reálnych výberových šetrení veľkého rozsahu**.

6.1 Vybavenosť domácností predmetmi dlhodobej spotreby

Problémom aplikácie nelineárnych pravdepodobnostných modelov binárnej voľby pri anticipácii vývoja vybavenosti domácností v Českej republike z **triedených údajov** sa zaoberali vo svojej práci Hušek a Moravová (2002). V uvedenej aplikácii bol odhadnutý logitový a probitový model binárnej voľby závislosti percentuálnej vybavenosti domácnosti mobilným telefónom na hrubom ročnom príjme jedného člena domácnosti v Kč, resp. na príjmovej skupine. Domácnosti boli roztriedené do desiatich príjmových skupín s rovnakou početnosťou. Medzi ďalšie vysvetľujúce premenné bola postupne zahrnutá príslušnosť domácnosti k jednej zo štyroch sociálnych skupín (zamestnanec, poľnohospodáriaci, samostatne pracujúci a dôchodca). Merateľnosť jednotlivých vysvetľujúcich bola dosiahnutá použitím techniky umelých nula-jednotkových premenných.

6.1.1 Dátový súbor a premenné modelu binárnej voľby

Český statistický úřad v súlade so zákonom č. 89/1995 Sb., o státní statistické službě uskutočnil v Českej republike na prelome mája a júna roku 2001 výberové šetrenie *Sociální situace domácností v roce 2001* (ďalej SSD 2001). Jeho cieľom bolo získať najnovšie reprezentatívne údaje o sociálno-demografických charakteristikách osôb a domácností v ČR, úrovni a stupni diferenciácie ich príjmov, informácie o vybavenosti domácností a charakteristikách bývania. Okrem uvedených objektívnych charakteristík boli do zisťovania zahrnuté otázky týkajúce sa životných podmienok respondentov.

Šetrenie bolo vykonané podľa vzoru a metodiky podobne zameraných šetrení organizované Eurostatom v krajinách Európskej únie od roku 1994. Výhodou tohto šetrenia je jeho komplexnosť – poskytuje informácie z rôznych oblastí života skúmaných domácností (napr. zamestnanie, príjmy, vzdelanie, zdravie, bývanie, životné prostredie), ďalej umožňuje analyzovať súvislostí medzi individuálnymi charakteristikami, definovať problémy a hľadať cestu k ich riešeniu. Dlhodobý charakter šetrenia vo forme panelových údajov tiež zaisťuje získavanie informácie o vývoji sociálno-ekonomických charakteristík domácností. Výberové šetrenie je harmonizované, čím je možná porovnateľnosť výsledkov medzi jednotlivými krajinami EU. Česká republika ako nedávno prijatý člen EU sa pripojí k tomuto šetreniu od roku 2006. Z tohoto dôvodu sa Český statistický úřad rozhodol uskutočniť vyššie uvedené šetrenie o sociálnej situácií domácností len ako jednorazové a so zredukovaným obsahom, pričom však uplatnil jednotnú medzinárodnú metodiku tak, aby bola zaistená využiteľnosť týchto výsledkov i v európskom meradle.

Definícia **hospodáriacej domácnosti** (ďalej HD) je založená na prehlásení spoločne bývajúcich osôb, že spoločne bývajú a hospodária, čiže hradia základné výdavky domácnosti (stravu, výdavky na bývanie a ostatné prevádzkové náklady).

U vybavenosti HD predmetmi dlhodobej spotreby bolo zisťované vlastníctvo týchto predmetov v kombinácii s vekom daného predmetu (do 5 rokov, nad 5 rokov). U predmetov, ktoré HD nevlastnila boli ďalej rozlišované prípady, kedy domácnosť predmet chcela, ale nemohla si ho dovoliť z finančných dôvodov, a prípady, kedy domácnosť predmety nechcela alebo nemala z iných dôvodov. Vybavenosť HD bola zisťovaná za predmety chladnička, mraznička a ich kombinácie, automatická práčka, sušička prádla, umývačka riadu, mikrovlná rúra, farebný televízor, video, pevný telefón, mobilný telefón, osobný počítač, prístup na internet, osobný automobil, chata a chalupa. Uvedená vybavenosť HD predmetmi dlhodobej spotreby tvorí v našej analýze závislú premennú, ktorá nadobúda dve alternatívy – vybavenosť HD predmetom dlhodobej spotreby bola kódovaná hodnotou 1 a HD, ktorá ho nevlastnila bola kódovaná hodnotou 0.

6.1.2 Analýza a predikcia modelov binárnej voľby

Analýza a predikcia modelov binárnej voľby je konkrétne aplikovaná na lineárny pravdepodobnostný, logitový a probitový model s ich vzájomným porovnaním a uvedením rozdielov. Rozsah výberového datového súboru tvorilo **10 599 domácností,** výberovú vzorku , po viacero uskutočnených simulačných pokusoch, tvorilo 3000 domácností.

<u>Lineárny pravdepodobnostný model</u>

Ako prvý je uvedený lineárny pravdepodobnostný model (LPM), ktorý predpokladá lineárnu závislosť medzi vysvetľovanou a vysvetľujúcimi premennými modelu. V modeli LPM je vysvetľovanou premennou výberová relatívna početnosť vybavenosti domácností predmetmi dlhodobej spotreby (mobilný telefón, pripojenie na internet, osobný automobil), prezentovaná pravdepodobnosťou p_i, že dichotomická premenná $y_i = 1$. Predvýber vysvetľovaných premenných bol uskutočnený pomocou χ^2 **testu o vzájomnej nezávislosti premenných** (Hebák, 2004) u kategoriálnych premenných a prostredníctvom **Kolmogorov-Smirnovovho** testu u premenných spojitých. Spojité premenné boli následne diskretizované (všetky majú charakter umelých nula-jednotkových premenných), ako ich uvádza tabuľka 6.1.

Tabuľka 6.1 Vysvetľujúce premenné LPM

Premenná	**Typ premennej**
PRIJEM	priemerná výška ročného hrubého príjmu na jedného člena domácnosti v v *i*-tej príjmovej skupine, uvažujeme 30 príjmových skupín,
POHLAVIE	pohlavie osoby v čele HD, 1 – muž, 0 – žena,
VEK	vek osoby v čele HD rozčlenený do siedmich vekových skupín,
VZDELANIE	dosiahnuté vzdelanie osoby v čele HD agregované do troch skupín[3],
STAV	rodinný stav osoby v čele HD, 1 – slobodný(á), 2 – ženatý, vydatá, 3 – rozvedený(á), 4 – ovdovelý(á)

V LPM bola testovaná nulová hypotéza o homoskedasticite prostredníctvom Whiteovho testu na päťpercentnej hladine významnosti a táto nulová hypotéza nebola zamietnutá. Z tohoto

[3] HD sú triedené podľa najvyššieho dosiahnutého vzdelania osoby v čele HD (v úplných rodinách potom v kombinácii so vzdelaním manželky). Stupne vzdelania boli pre tento účel agregované do troch skupín:

1) základné - základné vzdelanie, vyučení a nižšie stredné vzdelanie bez maturity,
2) stredné - úplné stredné s maturitou, nadstavbové štúdium, vyššie odborné vzdelanie,
3) vysokoškolské - bakalárske, vysokoškolské a doktorandské štúdium, vedecká príprava.

dôvodu nebolo potrebné použiť odhadovú metódu MZNŠ. Parametre LPM jednotlivých závislých premenných odhadneme MNŠ, ich hodnoty spolu s *t*-štatistikou a s pravdepodobnosťou P > *t* ich sú prezentované v tabuľke 6.2.

Tabuľka 6.2 Odhadnuté parametre LPM *Pripojenie na internet*

Premenná	Hodnota parametra	*t* štatistika	P > *t*
Konštanta	0,2995	8,82	<,0001
$PRIJEM_1$	-0,2032	-5,15	<,0001
$PRIJEM_2$	-0,1784	-4,61	<,0001
$PRIJEM_3$	-0,1563	-3,95	<,0001
$PRIJEM_4$	-0,1702	-4,49	<,0001
$PRIJEM_5$	-0,2011	-5,22	<,0001
$PRIJEM_6$	-0,1652	-4,30	<,0001
$PRIJEM_7$	-0,1667	-4,31	<,0001
$PRIJEM_8$	-0,1175	-2,99	0,0029
$PRIJEM_9$	-0,1497	-3,84	0,0001
$PRIJEM_{10}$	-0,1404	-3,54	0,0004
$PRIJEM_{11}$	-0,1530	-3,83	0,0001
$PRIJEM_{12}$	-0,1619	-4,24	<,0001
$PRIJEM_{13}$	-0,1533	-3,96	<,0001
$PRIJEM_{14}$	-0,1330	-3,59	0,0004
$PRIJEM_{15}$	-0,1950	-4,91	<,0001
$PRIJEM_{16}$	-0,1504	-3,86	0,0001
$PRIJEM_{17}$	-0,1608	-4,20	<,0001
$PRIJEM_{18}$	-0,1554	-4,00	<,0001
$PRIJEM_{19}$	-0,1686	-4,27	<,0001
$PRIJEM_{20}$	-0,1715	-4,49	<,0001
$PRIJEM_{21}$	-0,1668	-4,45	<,0001
$PRIJEM_{22}$	-0,1326	-3,53	0,0004
$PRIJEM_{23}$	-0,1250	-3,17	0,0015
$PRIJEM_{24}$	-0,0951	-2,59	0,0096
$PRIJEM_{25}$	-0,1810	-4,73	<,0001
$PRIJEM_{26}$	-0,1235	-3,19	0,0014
$PRIJEM_{27}$	-0,0965	-2,56	0,0104
$PRIJEM_{28}$	-0,1276	-3,41	0,0007
$PRIJEM_{29}$	-0,1072	-2,88	0,0040
POHLAVIE	0,0050	-0,29	0,7694
VEK_1	0,0598	1,47	0,1428
VEK_2	0,0814	3,53	0,0004
VEK_3	0,1485	6,74	<,0001
VEK_4	0,0977	4,68	<,0001
VEK_5	0,0023	0,11	0,9153
VEK_6	-0,0046	-0,22	0,8221
$VZDELANIE_1$	-0,1681	-9,80	<,0001
$VZDELANIE_2$	-0,0743	-4,18	<,0001
$STAV_1$	-0,0484	-1,95	0,0518
$STAV_2$	0,0161	0,76	0,4486
$STAV_3$	-0,0462	-2,16	0,0309

Test vypovedacej schopnosti modelu

U odhadnutého LPM je hodnota R^2, podľa očakávania, veľmi nízka a to rovná 0,1328 ($\overline{R}^2$ = 0,1210). Hodnota ***F***-štatistiky potvrdila významnosť modelu ako celku na jednopercentnej hladine významnosti.

Kladné hodnoty odhadnutých regresných koefientov zvyšujú pravdepodobnosť vybavenia domácnosti pripojením na internet, u záporných to je naopak. Odhadnutý regresný koeficient umelej nula-jednotkovej premennej $PRÍJEM_6$ je štatisticky významný a prezentuje konštantnú marginálnu zmenu p_i o 0,1652% (znamienko mínus predstavuje zníženie) pri zvýšení príjmu o jednu príjmovú skupinu. U všetkých umelých nula-jednotkých premenných $PRÍJEM_1$ až $PRÍJEM_{29}$ sú znamienka odhadnutých parametrov záporné, t.j, čím má osoba v čele domácnosti vyšší príjem, tým sa znižuje pravdepodobnosť, že domácnosť je vybavená pripojením na internet. Podobná interpretácia platí pre odhadnuté parametre u ostatných vysvetľujúcich premenných.

Predikované pravdepodobnosti

Odhadnutú predikovanú pravdepodobnosť p_i vybavenia domácnosti pripojením na internet je uved pre tri prípady:

1. Prvým prípadom je žena v čele HD, z druhej vekovej skupiny (32 rokov), zo 6. príjmová skupina, slobodná a vysokoškolsky vzdelaná. Predikovaná pravdepodobnosť p_i vybavenia domácnosti pripojením na internet tejto osoby v čele HD je 0,2312, teda viac ako 23%.

2. Muža v čele HD, zo štvrtej vekovej skupiny, 14. príjmová skupina, ženatý so základným vzdelaním. Predikovaná pravdepodobnosť p_i vybavenia domácnosti pripojením na internet u tohto muža je 0,1672.

3. Posledným príkladom je žena v čele HD, z tretej vekovej skupiny, z 19. príjmovej skupiny, rozvedená a stredoškolsky vzdelaná. Predikovaná pravdepodobnosť vybavenia domácnosti pripojením na internet je –0,1531, teda nulová.

Na poslednom príklade je naznačený jeden z problémov vyskytujúcich sa pri odhade LPM a to záporná hodnota predikovanej pravdepodobnosti.

Marginálne efekty

Marginálne efekty u LPM predstavujú parciálne derivácie závislej premennej podľa príslušnej nezávislej premennej a u lineárnych regresných modelov sú to priamo jednotlivé odhadnuté regresné koeficienty nezávislých premenných, ktoré sú konštantné pre každé pozorovanie.

Napriek uvedeným nedostatkom LPM je vhodným nástrojom analýzy a prognózy. Ďalej predpokladáme nelineárny vzťah závislosti vybavenia domácnosti predmetmi dlhodobej spotreby na vybraných charakteristikách, a to logitový a probitový model.

Nelineárny logitový a probitový model

Odhad nelineárneho logitového a probitového modelu vybavenia domácnosti pripojením na internet bol uskutočnený metódou MMV opäť z výberu o rozsahu 3000 pozorovaní ako u LPM. Tabuľka 6.3 obsahuje odhadnuté parametre modelov spolu s χ^2 štatistikou a P > χ^2. Hodnoty odhadnutých parametrov logitového modelu boli, z dôvodu porovnateľnosti (kapitola 4.4.2), upravené prenásobením 0,551.

Tabuľka 6.3 Odhadnuté parametre logitového a probitového modelu

	Logitový model			**Probitový model**		
Premenná	**Hodnota parametra pôvodná (upravená)**	**χ^2 štatistika**	**P > χ^2**	**Hodnota parametra**	**χ^2 štatistika**	**P > χ^2**
Konštanta	3,0790 (1,6965)	0,00	0,9820	-3,4757	0,01	0,9204
$PRIJEM_1$	0,9508 (0,5239)	10,48	0,0012	0,5263	12,07	0,0005
$PRIJEM_2$	0,6393 (0,3523)	7,68	0,0056	0,3493	8,03	0,0046
$PRIJEM_3$	0,5314 (0,2928)	5,51	0,0189	0,2923	5,74	0,0166
$PRIJEM_4$	0,5768 (0,3178)	6,54	0,0105	0,3308	7,36	0,0067
$PRIJEM_5$	0,8600 (0,4739)	11,60	0,0007	0,4904	13,22	0,0003
$PRIJEM_6$	0,6572 (0,3621)	7,24	0,0071	0,3591	7,48	0,0062
$PRIJEM_7$	0,6708 (0,3696)	6,88	0,0087	0,3881	7,82	0,0052
$PRIJEM_8$	0,2919 (0,1608)	1,89	0,1689	0,1607	1,88	0,1706
$PRIJEM_9$	0,5503 (0,3032)	4,59	0,0321	0,3207	5,33	0,0210
$PRIJEM_{10}$	0,4934 (0,2719)	3,45	0,0633	0,3038	4,16	0,0415
$PRIJEM_{11}$	0,6412 (0,3533)	5,33	0,0209	0,3607	5,69	0,0171
$PRIJEM_{12}$	0,6354 (0,3501)	5,37	0,0204	0,3159	5,15	0,0232
$PRIJEM_{13}$	0,5336 (0,2940)	3,11	0,0779	0,3123	3,62	0,0571
$PRIJEM_{14}$	0,4354 (0,2399)	3,31	0,0689	0,2320	3,17	0,0751
$PRIJEM_{15}$	1,4797 (0,8153)	7,88	0,0050	0,8059	9,93	0,0016
$PRIJEM_{16}$	0,6153 (0,3390)	4,24	0,0395	0,3452	4,82	0,0282
$PRIJEM_{17}$	0,5977 (0,3293)	5,99	0,0144	0,3385	6,48	0,0109
$PRIJEM_{18}$	0,6571 (0,3621)	4,89	0,0270	0,3682	5,56	0,0184
$PRIJEM_{19}$	0,8512 (0,4690)	6,51	0,0107	0,4679	7,40	0,0065
$PRIJEM_{20}$	0,6112 (0,3368)	5,74	0,0166	0,3638	6,81	0,0091
$PRIJEM_{21}$	0,6349 (0,3498)	6,15	0,0131	0,3532	6,68	0,0098
$PRIJEM_{22}$	0,3697 (0,2037)	3,22	0,0725	0,2191	3,61	0,0573
$PRIJEM_{23}$	0,3773 (0,2079)	2,91	0,0882	0,2184	3,06	0,0804
$PRIJEM_{24}$	0,1837 (0,1012)	0,94	0,3332	0,1199	1,22	0,2690
$PRIJEM_{25}$	0,6960 (0,3835)	9,25	0,0024	0,4015	10,24	0,0014
$PRIJEM_{26}$	0,3010 (0,1659)	2,04	0,1536	0,1690	2,10	0,1469
$PRIJEM_{27}$	0,3014 (0,1661)	2,66	0,1027	0,1484	2,03	0,1543
$PRIJEM_{28}$	0,3673 (0,2024)	3,74	0,0532	0,2067	3,72	0,0538
$PRIJEM_{29}$	0,2654 (0,1462)	2,08	0,1488	0,1430	1,89	0,1694
POHLAVIE	-0,0966 (-0,0532)	0,39	0,5329	-0,0499	0,37	0,5425
VEK_1	-5,5077 (-3,0347)	0,03	0,8719	-1,8014	0,04	0,8356
VEK_2	-5,6924 (-3,1365)	0,03	0,8677	-1,8713	0,05	0,8293
VEK_3	-6,0000 (-3,3060)	0,03	0,8606	-2,0542	0,06	0,8129
VEK_4	-5,7875 (-3,1889)	0,03	0,8655	-1,9281	0,05	0,8242
VEK_5	-5,0734 (-2,7954)	0,02	0,8819	-1,5567	0,03	0,8576
VEK_6	-4,7394 (-2,6114)	0,02	0,8897	-1,4101	0,03	0,8709
$VZDELANIE_1$	0,9668 (0,5327)	86,52	<0,0001	0,5200	88,78	<,0001
$VZDELANIE_2$	0,2168 (0,1195)	5,93	0,0149	0,1267	6,19	0,8356
$STAV_1$	-0,0651 (-0,0359)	0,04	0,8338	-0,0008	0,00	0,9956
$STAV_2$	-0,4845 (-0,2670)	2,64	0.1044	-0,2391	2,71	0,0994
$STAV_3$	-0,1031 (-0,0568)	0,12	0,7284	-0,0230	0,03	0,8727

Testy vypovedacej schopnosti modelu

Vhodnosť zvoleného modelu bola potvrdená prostredníctvom Akaikeho a Schwarzovho informačného kritéria (bol vybraný model, ktorý minimalizuje tieto informačné kritéria). Štatistickú významnosť odhadnutých regresných koeficientov testujeme pomocou χ^2 štatistiky a to testom pomeru pravdepodobností (4.55) a Waldovým testom (4.56). Na rozdiel od LPM, u logitového a probitového modelu sú napr. umelé nula-jednotkové premenné VEK_3 a VEK_4 štatisticky nevýznamné a odhadnuté parametre u premenných $PRIJEM_1$ až $PRIJEM_{29}$ majú opačné znamienka pri porovnaní s LPM. Tabuľka 6.4 uvádza štatistiky, ktoré hodnotia vypovedaciu schopnosť modelu.

Tabuľka 6.4 Testy vypovedacej schopnosti logitového a probitového modelu

	Hodnota štatistiky	
Štatistika	**Logitový model**	**Probitový model**
R^2_{PSP}	0,8641	0,8642
R^2_M	0,2345	0,2387

Test vypovedacej schopnosti založený na podiele správnych predpovedí (4.67) v tvare R^2_{PSP} dosiahol pomerne vysokú hodnotu, rovnako McFaddenov index podielu vierohodnosti R^2_M daný výrazom (4.64). Vysoké hodnoty boli dosiahnuté aj u pseudoštatistík R^2 Aldricha a Nelsona (1984) a Hagleho a Mitchella (1992) v programe *SPSS*.

Predikované pravdepodobnosti logitového a probitového modelu

Hodnoty predikovaných pravdepodobností vybavenia domácnosti pripojením na internet u logitového a probitového modelu sú uvedené pre tri prípady, podobne ako u lineárneho pravdepodobnostného modelu:

1. Prvým predpokladaným prípadom je žena v čele HD, z 2. vekovej skupiny (napríklad 32 rokov), zo 6. príjmovej skupiny, slobodná a vysokoškolsky vzdelaná. Predikovaná pravdepodobnosť vybavenia tejto domácnosti pripojením na internet u logitového modelu je 0,5575, u probitového modelu to je 0,5914.

2. Ďalej je uvažovaný muž v čele HD, zo 4. vekovej skupiny, zo 14. príjmovej skupiny, ženatý so základným vzdelaním. Predikovaná pravdepodobnosť vybavenia domácnosti tohoto muža pripojením na internet je u logitového modelu 0,5417, u probitového modelu to je 0,5664.

3. Ako posledný príklad je uvažovaná žena v čele HD, z 3. vekovej skupiny, z 19. príjmovej skupiny, rozvedená a stredoškolsky vzdelaná. Predikovaná pravdepodobnosť vybavenia danej domácnosti pripojením na internet u logitového modelu je 0,4618, u probitového modelu to je 0,4392.

Vo všetkých troch prípadoch dosiahli predikované podmienené pravdepodobnosti p_i u logitového a probitového modelu veľmi podobné výsledky. V poslednom prípade je táto pravdepodobnosť najnižšia u oboch modelov, podobne ako u LPM. Aplikovaním nelineárneho vzťahu medzi zaávislú a nezávislé premenné sme zároveň dosiahli to, že všetky predikované pravdepodobnosti sa nachádzajú v nula-jednotkovom intervale.

Marginálne efekty u logitového a probitového modelu

Odhadnuté regresné koeficienty u logitového a probitového modelu nepredstavujú marginálne efekty zmeny vysvetľujúcej premennej na vysvetľovanú ako to je u LPM. Nie sú konštantné, čiže je potrebné ich počítať pre každú vysvetľujúcu premennú zvlášť. Napríklad marginálny efekt pomienenej pravdepodobnosti p_i predstavujúcej domácnosť vybavenú pripojením na internet pre umelú nula-jednotkovú premennú $VZDELANIE_1$ je podľa (4.42) rovný u logitového modelu hodnote –0,1202, a u probitového modelu je tento marginálny efekt rovný hodnote –0,1132, kde za ostatné hodnoty vysvetľujúcich premenných (tiež v tvare umelých nula-jednotkovách premenných) boli dosadené ich priemerné hodnoty.

U logitového modelu (na rozdiel od probitového modelu) je možná ešte ďalšia podrobnejšia interpretácia dosiahnutých výsledkov.

Šanca zvolenej varianty (odds), čiže koľkokrát je pravdepodobnejšie, že domácnosť bude vybavená pripojením na internet je daná podielom pravdepodobností $p_i/1 - p_i$, kde $p_i = 1$. Napríklad u ženy v čele HD, z 2. vekovej skupiny (napríklad 32 rokov), zo 6. príjmovej skupiny, slobodnej a vysokoškolsky vzdelanej bola dosiahnutá šanca zvolenej varianty, a teda šanca, že jej domácnosť bude vybavená pripojením na internet je 0,5575/1–0,5575 = 1,2599-krát vyššia než pravdepodobnosť opačná.

Pomer šancí (odds ratio), tj. koľkokrát je vyššia šanca, že hodnota podmienenej pravdepodobnosti $p_i = 1$, než že $p_i = 0$ pre každú z vysvetľujúcich premenných je získaná odlogaritmovaním príslušného odhadnutého parametra vysvetľujúcej premennej. Napríklad pre premennú $VZDELANIE_1$ je pomer šancí rovný $e^{0,9668} = 2,63$, čiže pravdepodobnosť, alebo šanca, že domácnosť s osobou v čele so vzdelaním z prvej skupiny, bude vybavená pripojením na internet je 2,63-krát vyššia než domácnosť s osobou v čele so vzdelaním z tretej (referenčnej) skupiny.

Hodnoty pomeru šancí uvádzajú automaticky spolu s odhadnutými parametrami logitového modelu mnohé štatistické a ekonometrické programy (napr. program *SAS*, viz príloha dizertačnej práce). Obe charakteristiky, šanca i pomer šancí je často využívaná najmä v epidemiologických štúdiách (Zvárová a Malý, 2003).

6.2 Skórovanie kreditných rizík

Pojem **kreditné skórovanie** používajú finančné inštitúcie (banky, leasingové spoločnosti, apod.) pri stanovení bonity klienta, tj, jeho schopnosti dodržať finančné záväzky voči danej inštitúcii. Úlohou je určiť pravdepodobnosť, že klient sa bude správať určitým spôsobom a snaha o vybudovanie systému hodnotenia u každého klienta. Tento systém je založený na **historických dátach** (informáciách) o jednotlivých klientoch, pomocou ktorého sú klienti rozdelení na skupinu tzv. „dobrých“ a „zlých“ klientov, prípadne sú zaradení do skupiny tzv. „neurčitých“ klientov.

Každá finančná inštitúcia má vlastný systém hodnotenia a sledovania údajov, ktorého cieľom je správna identifikácia týchto dvoch, prípadne troch skupín. Problém nastáva pri určení hraníc medzi týmito skupinami, ktoré sa obvykle určí pomocou **defaultu klienta**. Default klienta predstavuje nedodržanie alebo nesplnenie záväzku klienta voči finančnej inštitúcii, ktoré sú určené v jeho úverovej zmluve. Na rozlíšenie druhu klienta sa obvykle predpokladá 12-mesačné obdobie s definíciou 90 alebo 120 dní po dátume splatnosti. Medzi obe hranice rozlišujúce dobrého a zlého klienta sa vkladá interval, pre ktorý klienta definujeme ako neurčitého.

Úlohou skórovnia kreditného rizika je odvodenie matematického pravidla na výpočet tzv. skóre alebo skórovej karty. **Skórová karta** je formulár obsahujúci charakteristiky, ktoré boli určitým spôsobom definované ako významné (štatisticky významné) pri rozlišovaní jednotlivých skupín klientov. Každá charakteristika alebo premenná obsahuje jednotlivé atribúty (alternatívy), napr. premenná vek môže obsahovať tri atribúty (vek do 35 rokov, 36-60 rokov, 61 rokov a viac), ktorým je priradené osobitné **skóre**. Toto skóre však musí uvažovať s prediktívnou silou premennej, koreláciou medzi charakteristikami ako aj s ďalšími operatívnymi faktormi. Celkové skóre potenciálneho klienta potom predstavuje suma bodov jednotlivých atribútov, ktoré tvoria súčasť skórovej karty.

Skórové karty sa od seba odlišujú definíciou správania sa klienta, premennými a procesom jej implementácie:

- **aplikačná** skórová karta sa týka hodnotenia nových klientov v danej finančnej inštitúcii,
- **behaviorálna** skórová karta definuje správanie už existujúcich klientov so známou históriou,
- **kolektívna** (collection) skórová karta sa uplatňuje v procese vymáhania.

Na modelovanie a predikovanie kreditného skórovania sa používajú rozličné prístupy, napr. diskriminačná analýza, rozhodovacie stromy, matematické programovanie, modely diskrétnej voľby či neurónové siete. Analýza je zameraná na aplikáciu modelov diskrétnej voľby v kombinácii s technikou dolovania dát.

6.2.1 Dátový súbor a premenné modelu multinomickej voľby

Na analýzu a predikciu kreditného skórovania boli poskytnuté údaje leasingovej spoločnosti. Zloženie výberového datového súboru podľa typu vyhodnotenia zmlúv sledovaného obdobia je uvedené v nasledujúcej tabuľke:

Charakteristika	**Počet klientov**	**Percentuálny podiel**
Dobrí klienti	15 026	96,04%
Zlí klienti	180	1,15%
Neurčití klienti	440	2,81%
Celkom	**15 646**	**100%**

O každom klientovi boli k dispozícii nasledovné údaje (tab. 6.5), ktoré zároveň predstavujú vysvetľujúce premenné multinomického logitového modelu (MLM).

Tabuľka 6.5 Vysvetľujúce premenné MLM

Premenná	**Typ premennej**
Typ úverového produktu	Kategoriálna premenná (0 – spotrebný úver, 1 – finančný leasing)
Dĺžka leasingu	Spojitá premenná (v mesiacoch)
Cena leasingového produktu	Spojitá premenná (v Kč)
Výška depozitu	Spojitá premenná (v %)
Doba v trvalom bydlisku	Spojitá premenná (v rokoch)
Vek	Spojitá premenná (v rokoch)
Rodinný stav	Kategoriálna premenná (1 – slobodný(á), 2 – ženatý, vydatá, 3 – rozvedený(á), 4 – ovdovelý(á)
Pohlavie	Kategoriálna premenná (0 – žena, 1 – muž)
Vzdelanie	Kategoriálna premenná (0 – nižšie ako vysokoškolské, 1 – vyššie ako vysokoškolské)
Typ bývania	Kategoriálna premenná (1- vlastný byt, 2 – vlastný dom, 3 – u rodičov, 4 – iné, 5 – podnájom, 6 – prenajatý byt, 7 – prenajatý dom)

Jednorozmerná analýza vysvetľujúcich premenných

Pri vytváraní skórovej funkcie je dôležitý výber a konštrukcia vysvetľujúcich premenných. U **kategoriálnych** premenných sme použili χ^2 **test o vzájomnej nezávislosti premenných** (Hebák, 2004) s nasledovnými dosiahnutými výsledkami:

- ***Typ bývania vs. Typ klienta –*** zamietnutá hypotéza o nezávislosti na hladine významnosti α = 0,001, kde hodnota testovacej štatistiky χ^2 je rovná 22,086. V ďalšej analýze budeme uvažovať s premennou *Typ bývania*.
- ***Rodinný stav vs. Typ klienta*** – zamietnutá hypotéza o nezávislosti na hladine významnosti α = 0,0001, kde hodnota testovacej štatistiky χ^2 je rovná 89,01. V ďalšej analýze budeme uvažovať s premennou *Rodinný stav*.
- ***Vzdelanie vs. Typ klienta –*** zamietnutá hypotéza o nezávislosti na hladine významnosti α = 0,0001, kde hodnota testovacej štatistiky χ^2 je rovná 20,86. V ďalšej analýze budeme uvažovať s premennou *Vzdelanie*.
- ***Pohlavie vs. Typ klienta*** – zamietnutá hypotéza o nezávislosti na hladine významnosti α = 0,491, kde hodnota testovacej štatistiky χ^2 je rovná 0,618. V ďalšej analýze nebudeme uvažovať s premennou *Pohlavie*.
- ***Typ úverového produktu vs. Typ klienta*** – zamietnutá hypotéza o nezávislosti na hladine významnosti α =0,811 kde hodnota testovacej štatistiky χ^2 je rovná 0,057. V ďalšej analýze nebudeme uvažovať s premennou *Typ úverového produktu*.

U spojitých premenných bol na ich predvýber použitý **Kolmogorov-Smirnov test** pre dva závislé výbery a **Mann-Whitneyov test** pre dva nezávislé výbery (Hebák, 2004). V ďalšej analýze boli na základe uvedených testov ponechané všetky spojité nezávislé premenné.

Všetky spojité vysvetľujúce premenné boli z dôvodu možnej existencie nízkej citlivosti skórovej funkcie (malé zmeny v premenných by spôsobili veľkú zmenu u závislej premennej) a nebezpečenstva falošnej monotónnosti v predikcii správania sa klienta v závislosti na vysvetľujúcich premenných, prevedené na premenné kategoriálne.

6.2.2 Analýza a predikcia modelu multinomickej voľby

Voľba tzv. **„cut off"** skóre alebo hraničného skóre pri definovaní typu klienta je dosiahnutá použitím **modelu multinomickej voľby** pre prípad **ordinálnej vysvetľovanej premennej**. Uvažovali sme prípad logitového modelu multinomickej voľby, výsledky odhadnutého probitového modelu boli veľmi podobné.

Pri analýze uvedených dát bola použitá techniku **dolovania dát,** ktorú je definovaná ako určitý proces výberu, prehľadávania, modelovania a predikcie veľkého objemu údajov (napr. Berka, 2003). Je veľmi často využívaná v oblasti financií, telekomunikácií, marketingu, plánovaní či v zdravotníctve. Medzi najznámejšie komerčné systémy dolovania dát patria produkty SAS Enterprise Miner, IBM Intelligent Miner, SPSS Clementine, Oracle Darwin alebo SGI MineSet, z nekomerčných to je Weka alebo Orange. Pri aplikácii kreditného skórovania autorka použila produkt **SAS Enterprise Miner** z dôvodu jeho najširšieho využitia v praxi. Tento produkt vychádza z vlastnej **metodológie SEMMA**, čo je skratka pre nasledovné po sebe idúce kroky:

- Sample – vzorkovanie, použitie z dôvodu existencie veľkého objemu údajov,
- Explore – vizuálna explorácia a redukcia údajov,
- Manipulate – vytvorenie najvhodnejšej skupiny a zhlukov dát pre ďalšiu analýzu,
- Model – analýza dát, neurónové siete, rozhodovacie stromy a regresná analýza,
- Access – porovnanie modelov a ich interpretácia.

Samotné dolovanie dát je programované pomocou procesných diagramov vizuálne prístupných najmä koncovému užívateľovi, jeho ukážka je uvedená v prílohe dizertačnej práce.

Odhad logitového modelu multinomickej voľby

Výberová vzorka bola zložená z 720 „dobrých" klientov, 180 „zlých" klientov a 220 klientov „neurčitých". Odhady parametrov modelu multinomickej voľby boli uskutočnené u rôzne veľkých výberových vzoriek so záverom, že výsledky sa významne od seba nelíšili. Záverečný výber vysvetľujúcich premenných v modeli bol vykonaný pomocou postupnej (stepwise) regresie. Ako vysvetľujúce premennými v multinomickom logitovom modeli (MLM) boli zvolené relevantné premenné:

- *DĹŽKA LEASINGU* – po diskretizácii vznikli tri kategórie: do 28 mesiacov (kódovaná ako 1), 29-52 mesiacov (2), 53 a viac mesiacov (3),
- *VÝŠKA DEPOZITU* – rozdelená do šiestich kategórií: do 13 % (1), 27,01-41 % (2), 41,01-70 % (3), nad 70,01 % (4),
- *VEK* – rozdelený do troch kategórií: do 38 rokov (1), 39-60 rokov (2), nad 61 rokov (3).

V modeli MLM je z dôvodu možnej prítomnosti perfektnej multikolinearity každá kategoriálna premenná zastúpená *k*-1 umelými nula-jednotkovými premennými, kde *k* predstavuje počet kategórií. Väčšina programových systémov volí referenčnú kategóriu samostatne (napr. systém SAS vyberá vždy tú najvyššiu). Výsledkom po odhade skórovej funkcie bude skórová karta, ktorá bude priraďovať klientovi body podľa tvaru odhadnutých regresných koeficientov. Z praktického hľadiska boli požadované body kladné, dosiahnuté tým, že ako referenčná kategória bola zvolená premenná s najnižšou pravdepodobnosťou byť ohodnotený ako „dobrý“ klient.

Vysvetľovanou premennou v modeli MLM je *Typ klienta*, kódovaný 0 – „zlý“ klient, 1 – „neurčitý“ klient a 2 – „dobrý“ klient. Keďže hodnoty závislej premennej bolo možné usporiadať, je ordinálna, bol uvažovaný prípad MLM s ordinálnou závislou premennou. Parametre MLM boli odhadnuté metodou MMV a spolu hodnotou χ^2 štatistiky sú uvedené v tabuľke 6.6.

Tabuľka 6.6 Odhadnuté parametre skórovej funkcie MLM

Premenná	**Hodnota parametra**	**χ^2 štatistika**	**P > χ^2**	**Pomer šancí**
Konštanta 2	-1,7406	89,6395	<0,0001	-
Konštanta 1	-0,4515	6,6112	0,0101	-
DĹŽKALEASINGU2	0,2709	2,1143	0,1459	1,311
DĹŽKALEASINGU3	1,2719	59,1051	<0,0001	3,567
VEK2	0,4225	7,6266	0,0058	1,526
VEK3	0,8730	27,8038	<0,0001	2,394
VÝŠKADEPOZITU2	1,3541	61,8912	<0,0001	3,873
VÝŠKADEPOZITU3	1,6980	60,5515	<0,0001	5,463
VÝŠKADEPOZITU4	1,9258	103,7550	<0,0001	6,861

Z tabuľky 6.6 je zrejmé, že odhadnutý MLM s ordinálnou závislou premennou obsahuje dve konštanty (počet kategórií závislej premennej mínus jedna), predstavujú tzv. hraničné alebo prahové skóre a platí predpoklad, že konštanta dva je nižšia ako konštanta jedna.

Okrem premennej *DĹŽKALEASINGU2* sú odhadnuté regresné koeficienty ostatných premenných na základe hodnoty χ^2 štatistiky štatisticky významné na jednopercentnej hladine významnosti. Jednotlivé odhadnuté parametre predstavujú vplyv vysvetľujúcej premennej na zmenu pravdepodobnosti, že vysvetľovaná premenná (napr. *VEK2*) sa bude nachádzať práve v tejto druhej vekovej skupine a nie v referenčnej skupine (*VEK1*). Kladné znamienka odhadnutých parametrov vysvetľujúcich premenných znamenajú, že s ich nárastom (vysvetľujúcich premenných) klesá pravdepodobnosť, že vysvetľovaná premenná nadobudne vyššiu a nie nižšiu kategóriu a naopak (Hebák, 2005). Napríklad u uchádzača z druhej vekovej skupiny v porovnaní s uchádzačom z prvej vekovej skupiny je pravdepodobnosť, že bude ohodnotený ako „neurčitý“ alebo „dobrý“ vyššia a je vyššia aj vo všetkých ostatných vekových skupinách vyššia: najviac u poslednej vekovej skupiny (*VEK3*), je viac ako dvakrát vyššia ($e^{0,8730}$= 2,394). Tieto hodnoty **pomeru šancí** pre každú z vysvetľujúcich premenných sú uvedené v poslednom stĺpci tabuľky 6.6.

Testy vypovedacej schopnosti modelu

Na základe minimálnych hodnôt Akaikeho a Schwarzovo informačného kritéria bola potvrdená vhodnosť zvoleného modelu. Štatistická významnosť všetkých odhadnutých regresných koeficientov zároveň bola testovaná pomocou χ^2 štatistiky (test pomeru vierohodností, skórový test a Waldov test). Tabuľka 6.7 uvádza štatistiky, ktoré hodnotia vypovedaciu schopnosť modelu.

Tabuľka 6.7 Testy vypovedacej schopnosti MLM

Štatistika	Hodnota štatistiky	Štatistika	Hodnota štatistiky
R^2_{PSP}	0,698	Giniho koeficient	0,436
R^2_M	0,113	Gamma	0,455
c-štatistika	0,718	Tau-a	0,229

Celková zhoda každého odhadnutého regresného koeficienta s údajmi meraná pomocou štatistiky podielu správnych odpovedí R^2_{PSP} dosahuje takmer 70%, hodnota McFaddenovho indexu podielu vierohodnosti R^2_M je viac ako 11%. Odhad c-štatistiky je rovný takmer 72% a predstavuje pravdepodobnosť, že dobrý klient bude mať vyššiu hodnotu (skóre) než klient neurčitý a zlý a je upravenou funkciou asymetrického Somersovho koeficienta (napr. Hebák, 2005). Hodnota Giniho koeficientu je 0,436, program *SAS* ju uvádza pod názvom Somers’ D (výberová miera asociácie).

Skórová karta

Výsledná skórová funkcia po odhade MLM je prezentovaná vo forme skórovej karty. Je to formulár, kde sú hodnotenému klientovi priraďované body v závislosti od úrovne, ktorú dosiahol v jednotlivých charakteristikách. Skórová karta je odvodená z odhadnutých regresných koeficientov skórovej funkcie, kde boli hodnoty jednotlivých koeficientov kvôli prehľadnosti a jednoduchšej interpretácii vynásobné hodnotou 100 a zaokrúhlené. Skórová karta má tvar uvedený v tabuľke 6.8.

Tabuľka 6.8 Skórová karta

Parameter	Hodnota	Body
Dĺžka leasingu	do 28 mesiacov	0
	29-52 mesiacov	27
	53 mesiacov a viac	127
Vek	do 38 rokov	0
	39-60 rokov	42
	61 rokov a viac	87
Výška depozitu	do 13%	0
	13,01-41%	135
	41,01-55%	170
	55,01 a viac	193

Predikované pravdepodobnosti u MLM

Vyrovnané hodnoty pravdepodobnosti p_i uchádzača ohodnoteného ako dobrého, neurčitého alebo zlého klienta sú definované na príkladoch troch reprezentatívnych zástupcov z každej skupiny podľa (5.27).

1. **Skupina dobrých klientov –** klient vo veku 43 rokov, ktorý žiada o leasing na 52 mesiacov a poskytol depozit vo výške 37% dosiahne podľa navrhnutej skórovej karty 197 bodov. Pravdepodobnosť, že uchádzač bude finančnou spoločnosťou ohodnotený ako dobrý klient je 0,5568, teda takmer 56%, neurčitým klientom s pravdepodobnosťou 0,2633 a zlým klientom s pravdepodobnosťou 0,1798.

2. **Skupina neurčitých klientov** – 40-ročný klient, ktorý žiada o leasing na 50 mesiacov a poskytol depozit vo výške 40% dosiahne podľa navrhnutej skórovej karty 204 bodov. Pravdepodobnosť, že uchádzač bude finančnou spoločnosťou ohodnotený ako dobrý klient je 0,5761, neurčitým klientom s pravdepodobnosťou 0,2553 a zlým klientom s pravdepodobnosťou 0,1685.

3. **Skupina zlých klientov** – klient vo veku 36 rokov, ktorý žiada o leasing na 54 mesiacov a poskytol depozit vo výške 23% dosiahne podľa navrhnutej skórovej karty 262 bodov. Pravdepodobnosť, že uchádzač bude finančnou spoločnosťou ohodnotený ako dobrý klient je 0,7079, teda takmer 71%, neurčitým klientom s pravdepodobnosťou 0,19 a zlým klientom s pravdepodobnosťou 0,1021.

Marginálne efekty u MLM

Marginálne efekty zmeny v priemeroch vysvetľujúcich premenných *Dĺžka leasingu, Vek* a *Výška depozitu* na závislú premennú *Typ klienta* (dobrý, neurčitý alebo zlý) uvádza tabuľka 6.9.

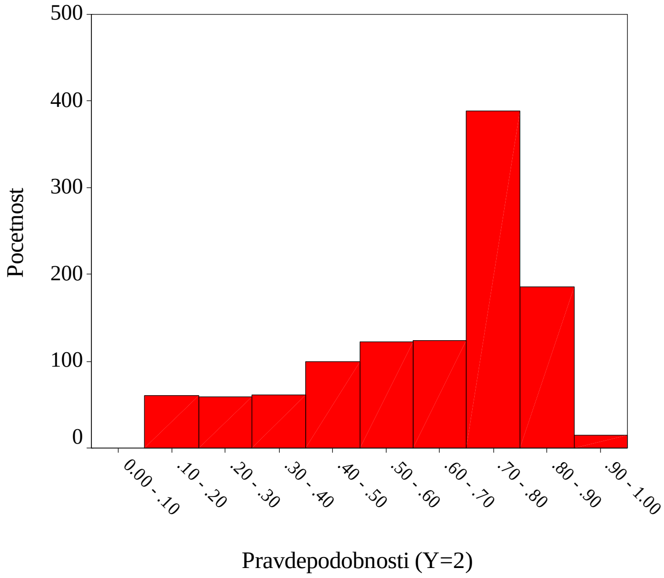
Tabuľka 6.9 Marginálne efekty MLM

Premenná		***Typ klienta – dobrý***	***Typ klienta – neurčitý***	***Typ klienta – zlý***
DĹŽKALEASINGU2	= 0	0,6410	0,2695	0,0895
DĹŽKALEASINGU2	= 1	0,7007	0,0697	0,2296
ZMENA		*-0,0597*	*0,1998*	*-0,1401*
DĹŽKALEASINGU3	= 0	0,4965	0,3525	0,1510
DĹŽKALEASINGU3	= 1	0,7787	0,0475	0,1738
ZMENA		-0,2822	0,3050	-0,0228
VEK2	= 0	0,6196	0,2832	0,0972
VEK2	= 1	0,7131	0,0659	0,2210
ZMENA		*-0,0935*	*0,2172*	*-0,1238*
VEK3	= 0	0,5833	0,3053	0,1114
VEK3	= 1	0,7702	0,0497	0,1801
ZMENA		*-0,1869*	*0,2556*	*-0,0687*
VÝŠKADEPOZITU2	= 0	0,5487	0,3252	0,1261
VÝŠKADEPOZITU2	= 1	0,8248	0,0359	0,1392
ZMENA		*-0,2761*	*0,2893*	*-0,0132*
VÝŠKADEPOZITU3	= 0	0,5893	0,3018	0,1089
VÝŠKADEPOZITU3	= 1	0,8869	0,0219	0,0913
ZMENA		*-0,2976*	*0,2799*	*0,0177*
VÝŠKADEPOZITU4	= 0	0,5156	0,3429	0,1415
VÝŠKADEPOZITU4	= 1	0,8796	0,0235	0,0970
ZMENA		*-0,3639*	*0,3194*	*0,0445*

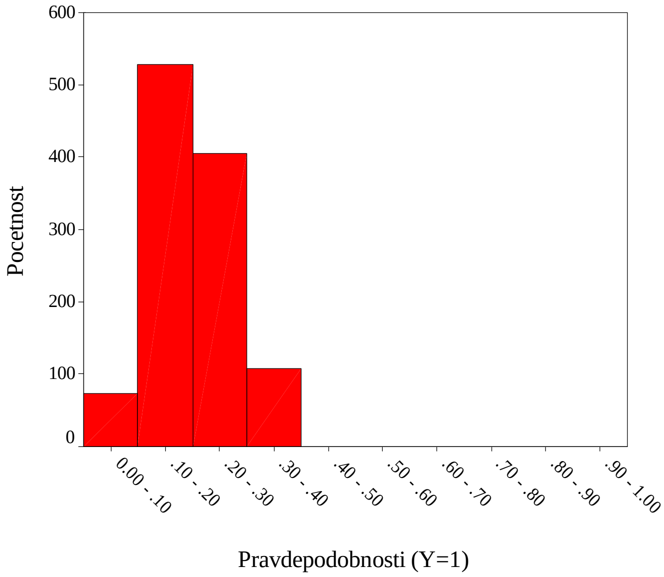
Marginálne efekty sú vypočítané vždy pre jednotlivú umelú nula-jednotkovú premennú za podmienky, že za hodnoty ostatných umelých nula-jednotkových premenných dosadíme ich priemery podľa (5.28). Suma marginálnych efektov každej premennej jej vždy rovná jednej a zmena je jednoducho rozdielom medzi zahrnutím a nezahrnutím danej premennej do MLM.

Grafy 6.1 až 6.3 znázorňujú pravdepodobnosti toho, že uchádzač bude ohodnotený ako dobrý, neurčitý alebo zlý klient, kde na vertikálnem osi sú znázornené ich početnosti.

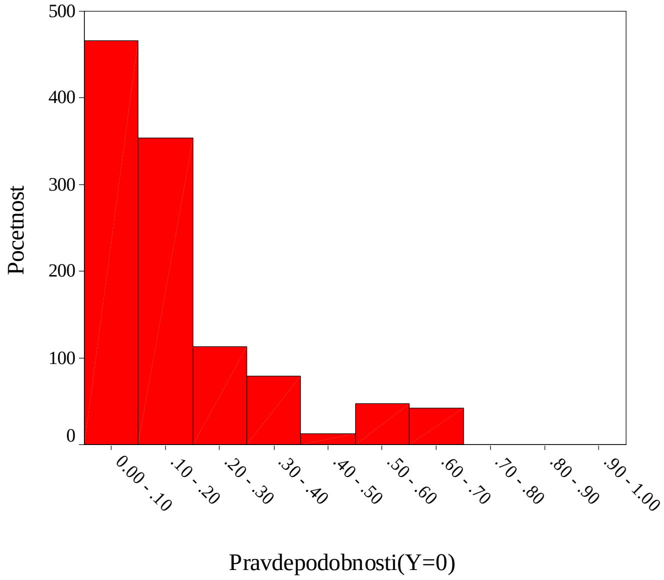
Graf 6.1 Graf pravdepodobnosti ohodnotenia uchádzača ako dobrého klienta



Graf 6.2 Graf pravdepodobnosti ohodnotenia uchádzača ako neurčitého klienta



Graf 6.3 Graf pravdepodobnosti ohodnotenia uchádzača ako zlého klienta



Multinomický probitový model

Závislosť *Typu klienta* na vybraných charakteristikách sme analyzovali aj pomocou **multinomického probitového modelu** pomocou systému **DCM** (Discrete Choice Modeling), ktorý pracuje na platforme ekonometrického programu *GiveWin*. Nie je bežnou súčasťou tohto programu, vyžaduje komplikovanú dodatočnú inštaláciu a pokročilejšie programovacie znalosti. Samotná procedúra odhadu je časovo veľmi náročná (odhad nášho modelu trval približne jeden a pol hodiny) v porovnaní s odhadom multinomického logitového modelu (niekoľko sekúnd napr. v *SAS*, *EViews* alebo v *Stata*). Rozdiel medzi oboma modelmi nebol štatisticky významný, preto ho v dizertačnej práci neuvádzame.

7 ZÁVER

Dizertačná práca prináša historické a teoretické súvislosti vývoja modelov diskrétnej voľby, ich odvodenie a zhodnotenie praktického využitia pri analýze dvoch rozsiahlych originálnych dátových súborov z praxe. Mikroekonomické východiská teórie spotrebiteľského správania neoklasickej ekonomickej školy a história vývoja modelov diskrétnej voľby sú popísané a objasnené v druhej a tretej kapitole. Vo štvrtej kapitole dizertačnej práce sú odvodené modely s **binárnou** závislou premennou, a to lineárny pravdepodobnostný, logitový a probitový model. Uvedená je ich matematická formulácia, metódy odhadu, testovanie hypotéz a problémy spojené s interpretáciou týchto modelov. Modelom usporiadanej a neusporiadanej **multinomickej** voľby a možnostiam ich využitia je venovaná piata kapitola.

Pri hodnotení aplikovateľnosti modelov diskrétnej voľby je jedným z hlavných problémov nedostupnosť vhodných dát, prípadne nedostatočný rozsah dátového súboru určeného na analýzu modelov diskrétnej voľby. V dizertačnej práci sú analyzované dva reálne dátové súbory veľkého rozsahu. Prvý dátový súbor tvorilo 10 599 domácností z výberového šetrenia *Sociální situace domácností 2001*, ktorý uskutočnil Český statistický úřad. Dátový súbor obsahuje podrobné informácie o socio-ekonomických charakteristikách domácností, ich názoroch na bývanie, apod. V tejto aplikácii bola analyzovaná a predikovaná vybavenosť domácností predmetmi dlhodobej spotreby v závislosti na vybraných socio-ekonomických charakteristikách prostredníctvom modelov binárnej voľby. Vysvetľovanou binárnou premennou bolo pripojenie domácnosti na internet, vysvetľujúcimi premennými boli pomocou metódy pomocnej regresie vybrané premenné čistý príjem, pohlavie, vek, vzdelanie a rodinný stav osoby, ktorá je v čele hospodáriacej domácnosti. Spojité vysvetľujúce premenné museli byť diskretizované resp. kategorizované z dôvodu nízkej robustnosti odhadovej funkcie vzhľadom k malým zmenám týchto premenných. Boli skúmané vlastnosti odhadnutých parametrov modelov na rôzne veľkých výberových vzorkách. So zmenami rozsahu výberu nedochádzalo ku zmenám hodnôt odhadnutých parametrov a ich štatistickej významnosti, takže konečnú výberovú vzorku tvorilo 3000 domácností. Lineárny pravdepodobnostný model bol, vzhľadom k nemeniacemu sa rozptylu náhodných zložiek, odhadnutý MNŠ. U tohto modelu boli zistené, v súlade s predpokladom, určité nedostatky spojené s jeho odhadom a interpretáciou. Patrili medzi ne nízka hodnota, predikované pravdepodobnosti ležali mimo nula-jednotkového intervalu, či samotný predpoklad linearity medzi závislou a nezávislými premennými modelu. Uvedené problémy boli odstránené použitím nelineárnych modelov binárnej voľby, a to logitovým a probitovým modelom, ktoré boli odhadnuté MMV. Štatistická významnosť všetkých parametrov u oboch modelov bola testovaná pomocou χ^2 štatistikya odhadnuté parametre sú štatisticky významné na jednopercentnej hladine významnosti. Pomerne vysoké hodnoty McFaddenovho indexu podielu vierohodností R_M^2 potvrdili vhodnosť použitých modelov. Logitový model binárnej voľby je citlivý na oblasť interpretácie dosiahnutých výsledkov v podobe šance, pomeru šancí skúmanej alternatívy, či pri skúmaní marginálnych efektov.

Druhá aplikácia pochádza z oblasti hodnotenia bonity klientov z hľadiska úverového rizika. Reálne údaje (15 646 pozorovaní) leasingovej spoločnosti boli analyzované pomocou kumulatívneho multinomického logitového modelu. Aplikácia uvedeného modelu patrí, podľa dostupných informácii, medzi prvé aplikácie multinomického logitového modelu na reálne dáta v Českej republike aj na Slovensku. Vysvetľovanou premennou v modeli bolo hodnotenie žiadateľa o leasing ako tzv. dobrého, neurčitého alebo zlého klienta. Každý žiadateľ poskytol leasingovej spoločnosti 21 socio-ekonomických charakteristík, ktoré tvorili východiskovú množinu vysvetľujúcich premenných modelu. Pri spracovávaní tohto dátového súboru bola uskutočnená podrobná prvotná kontrola údajov z hľadiska ich vecnej správnosti (napr. preklepy, zdvojené hodnoty, apod.), ďalej nasledovalo prekódovanie premenných s pozorovaniami uvedenými slovne do číselnej podoby. Predvýber vysvetľujúcich premenných bol uskutočnený pomocou jednorozmernej analýzy a to χ^2 testom dobrej zhody u kategoriálnych premenných a Kolmogorov-Smirnovovým testom u spojitých premenných. Finálnym výberom vysvetľujúcich premenných modelu na základe postupnej regresie boli vybrané len tri charakteristiky z 21 (dĺžka poskytnutého leasingu, výška depozitu a vek žiadateľa vo forme umelých nula-jednotkových premenných), ktoré vplývajú na rozhodnutie poskytovateľa úveru pri hodnotení žiadateľov o tento úver alebo leasing. Vlastnosti odhadnutých parametrov boli skúmané simulovaním na rôzne veľkých výberových vzorkách, konečnú výberovú vzorku tvorilo 720 uchádzačov ohodnotených ako dobrí klienti, 180 zlých klientov a 220 klientov neurčitých. Kumulatívny multinomický logitový model bol odhadnutý MMV v programe *SAS* pomocou modulu *SAS Enterprise Miner*. Tento modul umožňuje prepojenie techniky dolovania dát s ekonometrickou analýzou a pomocou procesných diagramov vytvorenie celkovej schémy procesu analýzy a predikovania modelov. Na základe vlastných skúseností je *SAS Enterprise Miner* veľmi často využívaný bankami alebo leasingovými spoločnosťami pri analýze úverového rizika klienta. Pomocou skórovej karty, vytvorenej z odhadnutých parametrov modelu, bolo priradené každému atribútu bodové skóre na základe ekonometrickej analýzy, ktoré zohľadňuje prediktívnu silu charakteristiky, korelácie medzi charakteristikami a ďalšie faktory. Celkové skóre uchádzača je výsledkom súčtu bodov jednotlivých atribútov, ktoré sú súčasťou skórovej karty. Predikcia vyrovnaných hodnôt pravdepodobností ohodnotenia uchádzača bola prezentovaná na troch typických predstaviteľoch z každej kategórie, uvedené sú aj marginálne efekty zmien v priemeroch vysvetľujúcich umelých nula-jednotkových premenných.

Pri oboch aplikáciách modelov dikrétnej voľby boli použité štatistické a ekonometrické programy *EViews, SPSS, Stata, SAS, GiveWin* s modulmi *PcGive* a *DCM*. Z hľadiska predprípravy dátového súboru, ktorá zahŕňala kódovanie, triedenie a náhodný výber považujem za užívateľsky najvhodnejší program *SPSS*. Programy *EViews* a *Stata* boli využité najmä pri testovaní štatistických hypotéz modelov diskrétnej voľby. Modul *SAS Enterprise Miner*, ktorý bol využitý v druhej aplikácii, predstavuje, po zvládnutí techniky dolovania dát a SAS programového kódu, veľmi vhodný nástroj pri spracovaní veľkého množstva údajov. Ocenila som jeho možnosť vytvorenia komplexného procesného diagramu ekonometrickej analýzy, ktorá umožňuje bezproblémový návrat k niektorým procedúram a ich jednoduchú modifikáciu pri úprave premenných, voľbe odhadovej funkcie, apod. Ďalšou veľkou výhodou tohto programu je, že podporuje súčasne prácu viacerých užívateľov na jednom projekte.

Multinomický probitový model bol odhadnutý prostredníctvom programu *PcGive* a modulom *DCM* (*Discrete Choice Modeling*). Multinomický probitový model patrí medzi modely časovo veľmi náročné na výpočet, takže programový produkt potrebný na jeho odhad a následnú analýzu nie je bežnou súčasťou ekonometrických programov a vyžaduje znalosť programovacieho jazyka Ox. Bola uskutočnená celá rada simulačných experimentov na rôzne veľkých výberových vzorkách a s rôzne definovanými počiatočnými podmienkami (napr. voľba tvaru kovariančnej matice, požadovaná presnosť metódy Monte Carlo, apod.) Výpočty u niektorých modelov boli komplikovanejšie, na osobnom počítači s bežným parametrami trvali až hodinu a pol. Hodnoty odhadnutých parametrov probitového modelu multinomickej voľby sa, podľa očakávania, nelíšili od multinomického logitového modelu, z tohto dôvodu neboli zaradené medzi výstupy.

Záverom možno na základe výsledkov oboch uskutočnených aplikácií konštatovať, že sa preukázala vhodnosť modelov diskrétnej voľby pre analýzu problémov z ekonomickej praxe. Domnievam sa, že modely diskrétnej voľby predstavujú jeden z najpraktickejších nástrojov mikroekomickej analýzy.

LITERATÚRA

[1] Adam, D.: *Les Réactions du Consommateur Devant les Prix*, Number 15 in Observation Economique, Sedes, Paris, 1958.

[2] Aitchison, J., Brown, J.A.C.: *The Lognormal Distribution*, Number 5 in University of Cambridge, Department of Applied Economics Monoghraphs, Cambridge University Press, Cambridge, 1957.

[3] Aitchison, J., Silvey S.D.: The Generalization of Probit Analysis to the Case of Multiple Responses, *Biometrika 44*, str. 131-140, 1957.

[4] Amemiya, T.: Bivariate Probit Analysis: Minimum Chi-Square Methods, *Journal of the American Statistical Association 69*, str. 940-944, 1994.

[5] Amemiya, T.: Qualitative Response Model: A Survey, *Journal of Economic Literature 19*, str. 1483-1536, 1981.

[6] Ashford, J.R., Sowden, R.R.: Multi-Variate Probit Analysis, *Biometrics 26*, str. 535-546, 1970.

[7] Berka, P.: *Dobývání znalostí z databází*, 1. vyd., Academia, Praha, 2003.

[8] Berkson, J.: Application of the Logistic Function to Bio-Assay, *Journal of the American Statistical Association 39*, str. 357-365, 1944.

[9] Berkson, J.: Why I Prefer Logits of Probits, *Biometrics 7*, str. 327-339, 1951.

[10] Bliss, C.I.: The Methods of Probits, *Science 79*, str. 38-39, 1934.

[11] Boskin, M.J.: A Conditional Logit Model of Occupational Choice, *Journal of Political Economy 82*, str. 389-398, 1974.

[12] Clark, C.E.: The Greatest of a Finite Set of Random Variables, *Operations Research 9*, str. 145-162, 1961.

[13] Cramer, J.S.: *The Origins and Development of the Logit Model*, 2003 (dostupné na www stránkach: http://publishing.cambridge.org/resources/0521815886/1208_default.pdf).

[14] Cox, D.R.: The Regression Analysis of Binary Sequences, *Journal of the Royal Statistical Society, Series B 20*, str. 215-242, 1958.

[15] Cox, D.R.: *Some Procedures Connected with the Logistic Qualitative Response Curve*, (in David, F. (ed.): Research Papers in Statistics: Festschrift for J. Neyman, str. 55-71, Wiley, London, 1966.

[16] Cox, D.R.: *Analysis of Binary Data*, Chapman and Hall, London, 1969.

[17] Farrell, M.J.: The Demand for Motorcars in the United States, *Journal of the Royal Statistical Society, series A 117*, str. 171-200, 1954.

[18] Fechner, G.T.: *Elemente der Psychophysik*, Breitkopf und Härtel, Leipzig, 1860.

[19] Finney, D.: *Probit Analysis*, 3rd edition (1st edition – 1947), Cambridge University Press, Cambridge, 1971.

[20] Fíglová, Z.: Aplikácia modelov diskrétnej voľby v oblasti zdravotno-sociálnej starostlivosti: prvé výsledky, *Participácia doktorandov na vedecko-výskumnej činnosti,* str. 33-38, EU FHI, Bratislava, 2001.

[21] Fíglová, Z.: Analýza nákupných úmyslov na základe modelov kvalitatívnej voľby, *Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze,* str. 101-113, VŠE, Praha, 2002.

[22] Fíglová, Z.: Predicting the Level of Institutional Care for the Elderly by the EQ-5D. Ostrava 03.09.2002 – 05.09.2002. In: RAMÍK, Jaroslav (ed.). *Mathematical Methods in Economics 2002*, str. 83-87, TU Ostrava, Ostrava, 2002.

[23] Fíglová, Z., Dlouhý, M.: Využití modelů binární volby v případě institucionalizované péče o starší populaci. *Zdravotní politika a ekonomika 5*, str. 70-77, 2002.

[24] Gaddum, J.H.: *Reports on Biological Standard III. Methods of Biological Assay Depending on a Quantal Response*, Medical Research Council, Special Report Series of the Medical Research Council 183, London, 1933.

[25] Green, W.H.: *Econometric Analysis,* 5th edition, Prentice Hall, New Jersey, 2003.

[26] Gujarati, D. N.: *Basic Econometrics*, 3rd edition, McGraw-Hill, Inc., New York, 1995.

[27] Gurland, J., Lee, I., Dahm, P.A.: *Polychotomous quantal response in biological assay*, Biometrics 16, str. 382-398, 1960.

[28] Hajivassiliou, V., McFadden, D.: The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises, *Cowles Discussion Paper 967*, 1990.

[29] Hausman, J.A., McFadden, D.: Specification Tests for the Multinomial Logit Model, *Econometrica 5,* str. 1219-1240, 1984.

[30] Hausman, J.A., Wise, D.A.: A Conditional Probit Model for Qualitative Choice: Discrete Desicions Recognizing Interdependence and Heterogenous Preferences, *Econometrica 46*, str. 403-426, 1978.

[31] Hebák, P., Hustopecký, J., Jarošová, E., Pecáková, I.: *Vícerozměrné statistické metody (1),* Informatorium, Praha, 2004.

[32] Hebák, P., Hustopecký, J., Malá, I.:. *Vícerozměrné statistické metody (2),* Informatorium, Praha, 2005.

[33] Hebák, P. a kol.: *Vícerozměrné statistické metody (3),* Informatorium, Praha, 2005.

[34] Heckman, J.J.: Dummy Endogenous Variables in a Simultaneous Equation System, *Econometrica 46*, str. 931-959, 1978.

[35] Hlaváček a kol.: *Mikroekonomie sounáležitosti se společenstvím*, Karolinum, Praha, 1999.

[36] Hosmer, D.W., Lemeshow, S.: *Applied Logistic Regression*, John Wiley & Sons, New York, 2000.

[37] Hušek, R.: *Základy ekonometrické analýzy I*, VŠE, Praha, 1995.

[38] Hušek, R.: *Základy ekonometrické analýzy II, Speciální postupy a techniky*, VŠE, Praha, 1998.

[39] Hušek, R.: *Ekonometrická analýza*, Ekopress, Praha, 1999.

[40] Hušek, R., Moravová, J.: Analýza a prognóza na základě modelů binární diskrétní volby, *Politická ekonomie 4*, VŠE, Praha, str. 579-593, 2002.

[41] Hušek, R., Pelikán, J.: *Aplikovaná ekonometrie*, Professional Publishing, Praha, 2003.

[42] Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decision under Risk, *Econometrica 47*, str. 263-291, 1979.

[43] Kleinbaum, D.G.: *Logistic Regression*, Springer-Verlag, New York, 1994.

[44] Kuznets, S.: *Secular Movement in Production and Prices*, Houghton Mifflin, Riverside Press, Boston, 1930. Reprinted by A.M. Kelley, 1967.

[45] Maddala, G.S.: *Introduction to Econometrics*, 3rd edition, John Wiley & Sons, Chichester, 2001.

[46] Maddala, G.S., Phillips, P.C.B., Srinivasan, T.N.: *Advances in Econometrics and Quantitative Economics*, Hartnolls Limited, Bodmin, Cornwall, 1995.

[47] Mantel, N.: Models for Complex Coningency Tables and Polychotomous Dosage Response Curves, *Biometrics 22*, str. 83-95, 1966.

[48] McCulloch, R., Rossi, P.: An Exact Likelihood Analysis of the Multinomial Probit Model, *Journal of Econometrics 64*, str. 207-240, 1994.

[49] McFadden, D.: Conditional Logit Analysis of Qualitative Choice Behavior, in Zarembka, P. (ed.), *Frontiers in Econometrics*, Academic Press, New York, str. 105-142, 1974.

[50] McFadden, D.: The Choice Theory Approach to Market Research, *Marketing Science 5*, Special Issue on Consumer Choice Models, str. 275-297, 1986.

[51] McFadden, D.: Regression-Based Specification Tests for the Multinomial Logit Model, *Journal of Econometrics 1/2*, str. 63-82, 1987.

[52] McFadden, D.: A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica 57(5)*, str. 995-1026, 1989.

[53] McFadden, D.: Econometric Analysis of Qualitative Response Models, *Handbook of Econometrics 2*, Engle, R.F. (editor spolu s McFadden), North Holland: Amsterdam, str. 1395-1457, 1994.

[54] McFadden, D.: Rationality for Economists?, *Journal of Risk and Uncertainty 19*, str. 73-105, 1999.

[55] McFadden, D.: Economic Choices Nobel Lecture, *American Economic Review 3*, str. 351-378, 2001, (dostupné na http://emlab.berkeley.edu/users/mcfadden/nobel/final-nobel.pdf).

[56] McFadden, D., Manski, Ch.F. (ed.): *Structural Analysis of Discrete Data with Econometric Applications*, The MIT Press, Cambridge, London, 1990.

[57] McFadden, D., Train, K., Tye, W.B.: An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model, *Transportation Research Record 637*, str. 39-45, 1976.

[58] McKelvey, R.D., Zavoina, W.: A statistical model for the analysis of ordinal level dependent variables, *Journal of Mathematical Sociology 4*, str. 103-120, 1975.

[59] Mittelhammer, R.C., Judge, G.G., Miller, D.J.: *Econometric Foundations*, Cambridge University Press, Cambridge, 2000.

[60] Morrison, D.G.: Upper Bounds for Correlations between Binary Outcomes and Probabilistic Predictions, *Journal of the American Statistical Association 67*, str. 68-70, 1972.

[61] Motlová, L., Dragomirecká, E., Španiel, F., Goppoldová, E., Záleský, R., Šelepová, P., Fíglová, Z., Höschl, C.: Relapse prevention on schizophrenia:does group family psychoeducation matter? One-year prospective follow-up field study. *International Jornal of Psychiatry in Clinical Practice 10*, str. 38-44, 2006.

[62] Nagelkerke, N.J.D.: A Note on a General Definition of the Coefficient of Determination, *Biometrika 3*, str. 691-692, 1991.

[63] Pearl, R., Reed, L.J.: On the Rate of Growth of the Population of the United States since 1870 and its Mathematical Representation, *Proceedings of the National Academy of Sciences 6*, str. 275-288, 1920.

[64] Pecáková, I., Novák, I., Herzmann, J.: *Pořizování a vyhodnocování dat ve výzkumech veřejného mínění*, Praha, VŠE, 2004.

[65] Pindyck, R., Rubinfeld, D.: *Econometric Models and Economic Forecasts*, 4th edition, McGraw-Hill, New York, 1998.

[66] Powers, D.A., Xie, Y.: *Statistical Methods for Categorical Data Analysis*, Academic Press, San Diego, 2000.

[67] Quandt, R., Baumol, W.: The Demand for Abstract Transport Models: Theory and Measurement, *Journal of Regional Science 6*, 1966.

[68] Ray, P.: Independence of Irrelevant Alternatives, *Econometrica 5*, str. 987-991, 1973.

[69] Reed, L.J., Berkson, J.: The Application of the Logistic Function to Experimental Data, *Journal of Physical Chemistry 33*, str. 760-779, 1929.

[70] Rodini, M., Ward, M.R., Woroch, G.A.: Going mobile: substitutability between fixed and mobile access, *Telecommunications Policy 27(5-6)*, str. 457-476, 2003.

[71] Sirůček, P.: Ekonomický člověk jako základ moderní ekonomie. *Marathon 3*, 1997, str. 5-14, (dostupné na www stránkach: http://misc.eunet.cz/marathon).

[72] Sirůček, P.: Pojetí člověka a racionality v ekonomických teoriích. *Marathon 7*, 2002, str. 4-20, (dostupné na www stránkach: http://misc.eunet.cz/marathon).

[73] Small, K.A., Hsiao, C.: Multinomial Logit Specification Tests, *International Economic Review 3*, str. 619-627, 1985.

[74] Soukup, J.: *Mikroekonomická analýza (vybrané kapitoly)*, Melandrium, Slaný 2001.

[75] Stuchlý, J.: Econometrical Analysis of Exam Results, In: *Mathematical Methods in Economics 1999*, Faculty of Management, University of Economics Prague, Jindřichův Hradec, 1999.

[76] Theil, H.: A Multinomial Extension of the Linear Logit Model, *International Economic Review 10*, str. 251-259, 1969.

[77] Thurstone, L.L.: A Law of Comparative Judgement, *Psychological Review 34*, str. 273-286, 1927, (znovu publikované in: *Psychological Review 2*, 1994, str. 266-270).

[78] Tobin, J.: Estimation of Relationships for Limited Dependent Variables, *Econometrica 26*, str. 24-36, 1958.

[79] Warner, S.: *Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice*, Northwestern University Press, Evanston, 1962.

[80] Wilson, E.B.: The logistic or autocatalytic grid, *Proceedings of the National Academy of Sciences 11*, str. 451-456, 1925.

[81] Wilson, E.B., Worcester, J.: The determination of L.D. 50 and its sampling error in bio-assay, *Proceedings of the National Academy of Sciences 29*, First series of three articles, 1943.

[82] Winsor, C.P.: A comparison of certain symmetrical growth curves, Journal *of the Washington Academy of Sciences 22*, str. 73-84, 1932.

[83] Yule, G.U.: The Growth of Population and the Factors which Control it, *Journal of the Royal Statistical Society 138*, str. 1-59, 1925.

[84] Zellner, A., Lee, T.H.: Joint Estimation of Relationships Involving Discrete Random Variables, *Econometrica 33*, str. 382-394, 1965.

[85] Zvárová, J., Malý, M. (ed.): *Statistické metody v epidemiologii*, svazek 2, Karolinum, Praha, 2003.

PRÍLOHA K DIZERTAČNEJ PRÁCI

ZOZNAM VEDECKÝCH PUBLIKÁCIÍ (JSTOR)

Ekonómia

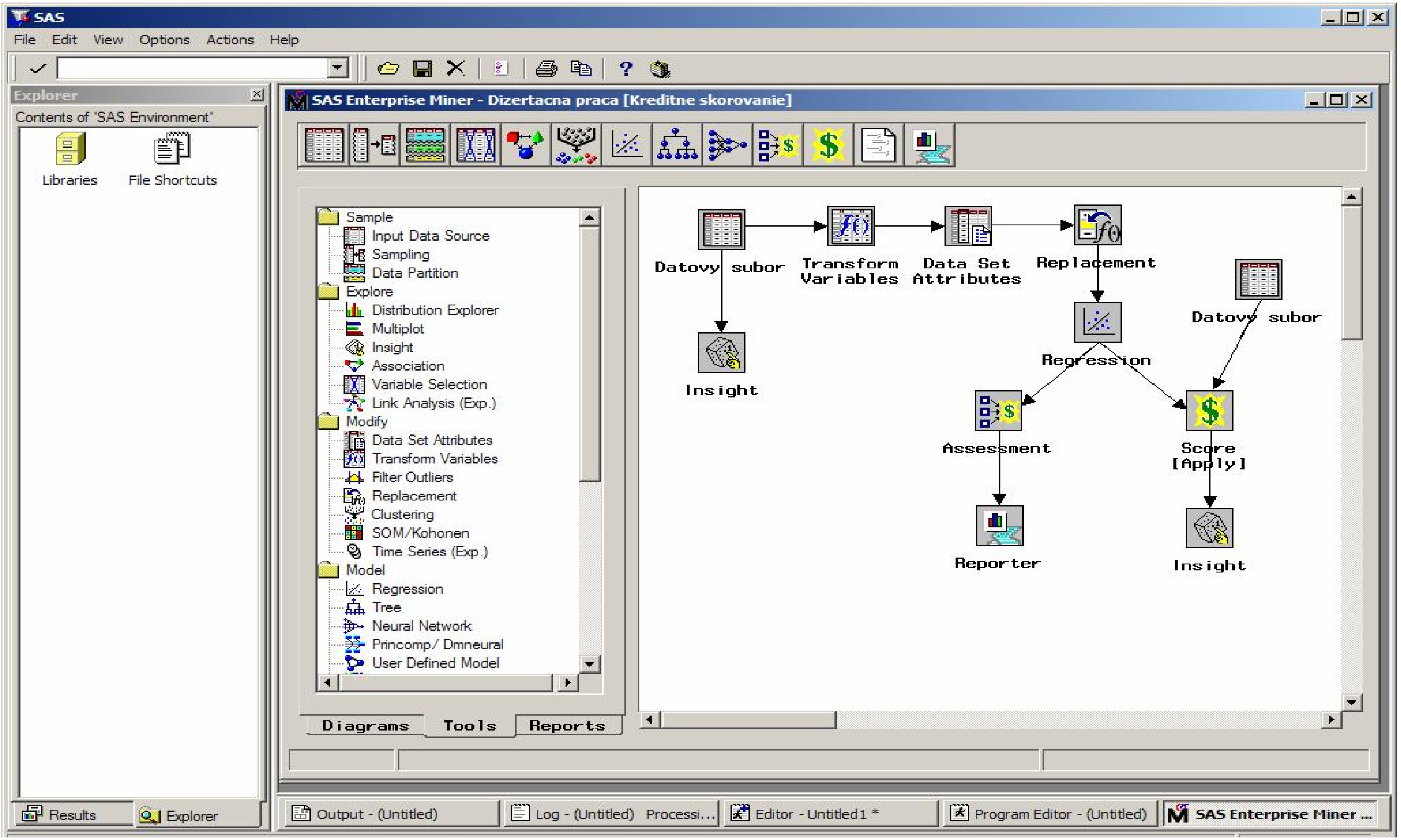
American Economic Review
Bell Journal of Economics and Management Science
Brookings Papers on Economic Activity
Canadian Journal of Economics
Econometrica
Economic Geography
Economic History Review
Economic Journal
Economica
Industrial and Labor Relations Review
International Economic Review
Journal of Economic History
Journal of Economic Literature
Journal of Human Resources
Journal of Industrial Economics
Journal of Money
Credit and Banking
Journal of Political Economy
Journal of Risk and Insurance
Oxford Economic Papers
Quarterly Journal of Economics
Review of Economic Studies,
Review of Economics and Statistics
Canadian Journal of Economics and Political Science
Bell Journal of Economics
Journal of Economic Abstracts
Contributions to Canadian Economics
Journal of Labor Economics
RAND Journal of Economics
Journal of Applied Econometrics
Journal of Economic Perspectives
Journal of Insurance
Publications of the American Economic Association
Brookings Papers on Economic Activity
Microeconomics
American Economic Association Quarterly

Journal of the American Association of University Teachers of Insurance
Proceedings of the Annual Meeting (American Association of University Teachers of Insurance)

Štatistika

American Statistician
Annals of Mathematical Statistics
Biometrics
Biometrika
Journal of the Royal Statistical Society. Series A (General)
Journal of the Royal Statistical Society. Series B (Methodological)
Applied Statistics
Statistician
Annals of Statistics
Annals of Probability
Biometrics Bulletin
Journal of the American Statistical Association
Statistical Science
Journal of the Royal Statistical Society
Journal of the Statistical Society of London
Journal of the Royal Statistical Society. Series A (Statistics in Society)
Annals of Applied Probability, Journal of the Royal Statistical Society. Series B (Statistical Methodology)
Supplement to the Journal of the Royal Statistical Society
Incorporated Statistician
Publications of the American Statistical Association
Quarterly Publications of the American Statistical Association

UKÁŽKA SAS ENTERPRISE MINER



ZOZNAM GRAFOV, OBRÁZKOV A TABULIEK

GRAFY

OBRÁZKY

TABUĽKY

VÝSTUPY ODHADNUTÝCH MODELOV

Lineárny pravdepodobnostný model binárnej voľby

The REG Procedure
Model: MODEL1
Dependent Variable: net net

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	41	32.18783	0.78507	11.05	<.0001
Error	2958	210.22684	0.07107		
Corrected Total	2999	242.41467			

Root MSE	0.26659	R-Square	0.1328
Dependent Mean	0.08867	Adj R-Sq	0.1208
Coeff Var	300.66620		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	0.29953	0.03395	8.82	<.0001	0
dprij1	dprij1	1	-0.20324	0.03948	-5.15	<.0001	-0.12520
dprij2	dprij2	1	-0.17840	0.03871	-4.61	<.0001	-0.11211
dprij3	dprij3	1	-0.15631	0.03959	-3.95	<.0001	-0.09481
dprij4	dprij4	1	-0.17022	0.03792	-4.49	<.0001	-0.11352
dprij5	dprij5	1	-0.20107	0.03855	-5.22	<.0001	-0.12697
dprij6	dprij6	1	-0.16518	0.03841	-4.30	<.0001	-0.10531
dprij7	dprij7	1	-0.16668	0.03863	-4.31	<.0001	-0.10526
dprij8	dprij8	1	-0.11753	0.03937	-2.99	0.0029	-0.07166
dprij9	dprij9	1	-0.14972	0.03899	-3.84	0.0001	-0.09316
dprij10	dprij10	1	-0.14040	0.03968	-3.54	0.0004	-0.08471
dprij11	dprij11	1	-0.15300	0.03995	-3.83	0.0001	-0.09032
dprij12	dprij12	1	-0.16187	0.03822	-4.24	<.0001	-0.10465
dprij13	dprij13	1	-0.15328	0.03875	-3.96	<.0001	-0.09818
dprij14	dprij14	1	-0.13295	0.03706	-3.59	0.0003	-0.09238
dprij15	dprij15	1	-0.19498	0.03973	-4.91	<.0001	-0.11638
dprij16	dprij16	1	-0.15043	0.03896	-3.86	0.0001	-0.09407
dprij17	dprij17	1	-0.16079	0.03824	-4.20	<.0001	-0.10202
dprij18	dprij18	1	-0.15538	0.03883	-4.00	<.0001	-0.09717
dprij19	dprij19	1	-0.16863	0.03945	-4.27	<.0001	-0.10228
dprij20	dprij20	1	-0.17150	0.03822	-4.49	<.0001	-0.10830
dprij21	dprij21	1	-0.16677	0.03749	-4.45	<.0001	-0.10978
dprij22	dprij22	1	-0.13262	0.03755	-3.53	0.0004	-0.08613
dprij23	dprij23	1	-0.12499	0.03946	-3.17	0.0016	-0.07419
dprij24	dprij24	1	-0.09510	0.03673	-2.59	0.0097	-0.06424
dprij25	dprij25	1	-0.18106	0.03824	-4.73	<.0001	-0.11378
dprij26	dprij26	1	-0.12347	0.03872	-3.19	0.0014	-0.07567
dprij27	dprij27	1	-0.09649	0.03767	-2.56	0.0105	-0.06152
dprij28	dprij28	1	-0.12779	0.03751	-3.41	0.0007	-0.08186
dprij29	dprij29	1	-0.10702	0.03719	-2.88	0.0040	-0.06951
kodpoh	kodpoh	1	-0.00503	0.01715	-0.29	0.7694	-0.00735
dvek1	dvek1	1	0.05983	0.04081	1.47	0.1428	0.02947
dvek2	dvek2	1	0.08135	0.02301	3.53	0.0004	0.10402
dvek3	dvek3	1	0.14852	0.02203	6.74	<.0001	0.19822
dvek4	dvek4	1	0.09772	0.02090	4.68	<.0001	0.14605
dvek5	dvek5	1	0.00226	0.02126	0.11	0.9153	0.00292
dvek6	dvek6	1	-0.00455	0.02024	-0.22	0.8221	-0.00566
dvzdel1	dvzdel1	1	-0.16805	0.01716	-9.80	<.0001	-0.28937
dvzdel2	dvzdel2	1	-0.07432	0.01780	-4.18	<.0001	-0.11835
dstav1	dstav1	1	-0.04839	0.02487	-1.95	0.0518	-0.05091
dstav2	dstav2	1	0.01611	0.02126	0.76	0.4486	0.02746
dstav3	dstav3	1	-0.04615	0.02138	-2.16	0.0309	-0.05454

Test of First and Second
Moment Specification

DF	Chi-Square	Pr > ChiSq

```
                         427         559.58         <.0001


                     Durbin-Watson D                1.993
                     Number of Observations          3000
                     1st Order Autocorrelation      0.003
```

Logitový model binárnej voľby

```
                                  The DMREG Procedure

Training Data Set:           EMDATA.DMDB6UVZ
DMDB Catalog:                EMPROJ.DMDB6UVZ
Target Variable:             net (net)
Target Measurement Level:    Ordinal
Number of Target Categories: 2
Error:                       MBernoulli
Link Function:               Logit
Number of Model Parameters:  42
Number of Observations:      3000


                                  Target Profile

                              Ordered          Total
                                Value   net  Frequency

                                    1     1        266
                                    2     0       2734

                          Dual Quasi-Newton Optimization

                  Dual Broyden - Fletcher - Goldfarb - Shanno Update (DBFGS)

                     Parameter Estimates              42

                              Optimization Start

Active Constraints                          0  Objective Function                898.32622356
```

Max Abs Gradient Element 210.264

Iter	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Step Size	Slope of Search Direction
1	0	5	0	755.15725	143.2	142.4	0.941	-398.3
2	0	7	0	707.57035	47.5869	25.4002	1.000	-107.7
3	0	9	0	698.31988	9.2505	22.8475	1.000	-16.054
4	0	11	0	692.64878	5.6711	8.8341	1.000	-9.458
5	0	13	0	690.79666	1.8521	6.9858	1.000	-2.933
6	0	15	0	690.23773	0.5589	12.2565	2.397	-1.931
7	0	19	0	689.04686	1.1909	4.0054	1.503	-1.859
8	0	22	0	688.71250	0.3344	1.8214	2.179	-0.318
9	0	24	0	688.51577	0.1967	4.2640	3.987	-0.210
10	0	26	0	688.25649	0.2593	2.4638	1.440	-0.341
11	0	28	0	687.98391	0.2726	2.8635	4.185	-0.155
12	0	31	0	687.87066	0.1132	0.8871	1.330	-0.170
13	0	33	0	687.77511	0.0955	2.1751	5.159	-0.0549
14	0	35	0	687.72503	0.0501	2.6054	2.380	-0.100
15	0	37	0	687.68692	0.0381	1.5716	1.681	-0.0745
16	0	40	0	687.67013	0.0168	0.3859	1.406	-0.0241
17	0	42	0	687.66094	0.00920	0.8069	6.599	-0.0064
18	0	44	0	687.65690	0.00404	0.8890	2.329	-0.0099
19	0	46	0	687.65214	0.00476	0.1441	1.271	-0.0079
20	0	49	0	687.64932	0.00282	0.2027	4.639	-0.0013
21	0	52	0	687.64794	0.00139	0.1299	1.700	-0.0017
22	0	55	0	687.64715	0.000785	0.0795	2.190	-0.0008
23	0	58	0	687.64677	0.000383	0.0653	2.150	-0.0004

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-square	Pr > Chi-square	exp(Est)
Intercept		1	3.0790	136.7	0.00	0.9820	21.736
dprij1	0	1	0.9508	0.2937	10.48	0.0012	2.588
dprij10	0	1	0.4934	0.2657	3.45	0.0633	1.638
dprij11	0	1	0.6412	0.2777	5.33	0.0209	1.899
dprij12	0	1	0.6354	0.2741	5.37	0.0204	1.888
dprij13	0	1	0.5336	0.3027	3.11	0.0779	1.705
dprij14	0	1	0.4354	0.2394	3.31	0.0689	1.546
dprij15	0	1	1.4797	0.5270	7.88	0.0050	4.392
dprij16	0	1	0.6153	0.2989	4.24	0.0395	1.850
dprij17	0	1	0.5977	0.2443	5.99	0.0144	1.818
dprij18	0	1	0.6571	0.2972	4.89	0.0270	1.929
dprij19	0	1	0.8512	0.3336	6.51	0.0107	2.342
dprij2	0	1	0.6393	0.2308	7.68	0.0056	1.895
dprij20	0	1	0.6112	0.2552	5.74	0.0166	1.843
dprij21	0	1	0.6349	0.2559	6.15	0.0131	1.887
dprij22	0	1	0.3697	0.2059	3.22	0.0725	1.447
dprij23	0	1	0.3773	0.2213	2.91	0.0882	1.458
dprij24	0	1	0.1837	0.1898	0.94	0.3332	1.202
dprij25	0	1	0.6960	0.2289	9.25	0.0024	2.006
dprij26	0	1	0.3010	0.2109	2.04	0.1536	1.351
dprij27	0	1	0.3014	0.1847	2.66	0.1027	1.352
dprij28	0	1	0.3673	0.1900	3.74	0.0532	1.444
dprij29	0	1	0.2654	0.1838	2.08	0.1488	1.304
dprij3	0	1	0.5314	0.2263	5.51	0.0189	1.701
dprij4	0	1	0.5768	0.2256	6.54	0.0105	1.780
dprij5	0	1	0.8600	0.2525	11.60	0.0007	2.363
dprij6	0	1	0.6572	0.2442	7.24	0.0071	1.929
dprij7	0	1	0.6708	0.2558	6.88	0.0087	1.956
dprij8	0	1	0.2919	0.2122	1.89	0.1689	1.339
dprij9	0	1	0.5503	0.2568	4.59	0.0321	1.734
dstav1	0	1	-0.0651	0.3101	0.04	0.8338	0.937
dstav2	0	1	-0.4845	0.2984	2.64	0.1044	0.616
dstav3	0	1	-0.1031	0.2969	0.12	0.7284	0.902
dvek1	0	1	-5.5077	34.1652	0.03	0.8719	0.004
dvek2	0	1	-5.6924	34.1639	0.03	0.8677	0.003
dvek3	0	1	-6.0000	34.1638	0.03	0.8606	0.002
dvek4	0	1	-5.7875	34.1638	0.03	0.8655	0.003
dvek5	0	1	-5.0734	34.1640	0.02	0.8819	0.006
dvek6	0	1	-4.7394	34.1644	0.02	0.8897	0.009
dvzdel1	0	1	0.9668	0.1039	86.52	<.0001	2.630
dvzdel2	0	1	0.2168	0.0890	5.93	0.0149	1.242
kodpoh	0	1	-0.0966	0.1550	0.39	0.5329	0.908

Odds Ratio Estimates

Input Odds Ratio

```
  dprij1  0 vs 1            6.697
  dprij10 0 vs 1            2.682
  dprij11 0 vs 1            3.606
  dprij12 0 vs 1            3.564
  dprij13 0 vs 1            2.907
  dprij14 0 vs 1            2.389
  dprij15 0 vs 1           19.286
  dprij16 0 vs 1            3.423
  dprij17 0 vs 1            3.305
  dprij18 0 vs 1            3.721
  dprij19 0 vs 1            5.487
  dprij2  0 vs 1            3.592
  dprij20 0 vs 1            3.396
  dprij21 0 vs 1            3.560
  dprij22 0 vs 1            2.095
  dprij23 0 vs 1            2.127
  dprij24 0 vs 1            1.444
  dprij25 0 vs 1            4.023
  dprij26 0 vs 1            1.826
  dprij27 0 vs 1            1.827
  dprij28 0 vs 1            2.085
  dprij29 0 vs 1            1.700
  dprij3  0 vs 1            2.894
  dprij4  0 vs 1            3.170
  dprij5  0 vs 1            5.585
  dprij6  0 vs 1            3.722
  dprij7  0 vs 1            3.825
  dprij8  0 vs 1            1.793
  dprij9  0 vs 1            3.006
  dstav1  0 vs 1            0.878
  dstav2  0 vs 1            0.379
  dstav3  0 vs 1            0.814
  dvek1   0 vs 1            0.000
  dvek2   0 vs 1            0.000
  dvek3   0 vs 1            0.000
  dvek4   0 vs 1            0.000
  dvek5   0 vs 1            0.000
  dvek6   0 vs 1            0.000
  dvzdel1 0 vs 1            6.914
  dvzdel2 0 vs 1            1.543
  kodpoh  0 vs 1            0.824

         The FREQ Procedure

      Table of F_net by I_net

F_net(From: net)     I_net(Into: net)

Frequency‚
Percent  ‚
Row Pct  ‚
Col Pct  ‚0       ‚1       ‚  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
0        ‚   2718 ‚     16 ‚   2734
         ‚  90.60 ‚   0.53 ‚  91.13
         ‚  99.41 ‚   0.59 ‚
         ‚  91.55 ‚  51.61 ‚
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
1        ‚    251 ‚     15 ‚    266
         ‚   8.37 ‚   0.50 ‚   8.87
         ‚  94.36 ‚   5.64 ‚
         ‚   8.45 ‚  48.39 ‚
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total        2969       31     3000
            98.97     1.03   100.00
```

Probitový model binárnej voľby

The DMREG Procedure

Training Data Set:	EMDATA.DMDB6UVZ
DMDB Catalog:	EMPROJ.DMDB6UVZ
Target Variable:	net (net)
Target Measurement Level:	Ordinal
Number of Target Categories:	2
Error:	MBernoulli
Link Function:	Probit
Number of Model Parameters:	42
Number of Observations:	3000

Target Profile

Ordered Value	net	Total Frequency
1	1	266
2	0	2734

Dual Quasi-Newton Optimization

Dual Broyden - Fletcher - Goldfarb - Shanno Update (DBFGS)

Parameter Estimates 42

Optimization Start

Active Constraints	0	Objective Function	898.32622356
Max Abs Gradient Element	417.89382904		

Iter	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Step Size	Slope of Search Direction
1	0	5	0	711.07506	187.3	107.4	1.000	-351.8
2	0	9	0	689.28306	21.7920	20.7839	2.046	-21.432
3	0	13	0	686.49174	2.7913	15.5324	2.000	-2.965
4	0	17	0	685.04840	1.4433	6.3451	2.401	-1.196
5	0	21	0	684.61127	0.4371	5.3162	2.225	-0.392
6	0	23	0	684.47141	0.1399	11.0209	6.114	-0.179
7	0	27	0	684.17289	0.2985	1.5288	1.423	-0.492
8	0	30	0	684.05481	0.1181	3.3595	3.205	-0.0811
9	0	32	0	683.97294	0.0819	2.4454	3.741	-0.0789
10	0	35	0	683.92066	0.0523	0.8398	1.423	-0.0734
11	0	38	0	683.89134	0.0293	0.9541	4.255	-0.0164
12	0	41	0	683.88507	0.00626	0.3708	2.230	-0.0057
13	0	44	0	683.88146	0.00362	0.3936	2.577	-0.0030
14	0	47	0	683.87943	0.00203	0.3081	2.455	-0.0018
15	0	50	0	683.87837	0.00106	0.1926	2.680	-0.0009
16	0	53	0	683.87803	0.000343	0.1139	2.314	-0.0003

Optimization Results

Iterations	16	Function Calls	55
Gradient Calls	24	Active Constraints	0
Objective Function	683.87802566	Max Abs Gradient Element	0.1138518389
Slope of Search Direction	-0.000332232		

GCONV convergence criterion satisfied.

NOTE: At least one element of the (projected) gradient is greater than 1e-3.

Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
-2 LOG L	1796.652	1367.756	428.896 with 41 DF (p<.0001)

The SAS System 13:59 Sunday, March 12, 2000 15

The DMREG Procedure

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > Chi-Square
dprij1	1	12.0743	0.0005
dprij10	1	4.1563	0.0415
dprij11	1	5.6866	0.0171
dprij12	1	5.1510	0.0232
dprij13	1	3.6185	0.0571
dprij14	1	3.1683	0.0751
dprij15	1	9.9324	0.0016
dprij16	1	4.8156	0.0282
dprij17	1	6.4756	0.0109
dprij18	1	5.5550	0.0184
dprij19	1	7.4020	0.0065
dprij2	1	8.0281	0.0046
dprij20	1	6.8122	0.0091
dprij21	1	6.6760	0.0098
dprij22	1	3.6149	0.0573
dprij23	1	3.0571	0.0804
dprij24	1	1.2221	0.2690
dprij25	1	10.2379	0.0014
dprij26	1	2.1039	0.1469
dprij27	1	2.0292	0.1543
dprij28	1	3.7196	0.0538
dprij29	1	1.8882	0.1694
dprij3	1	5.7388	0.0166
dprij4	1	7.3606	0.0067
dprij5	1	13.2152	0.0003
dprij6	1	7.4834	0.0062
dprij7	1	7.8189	0.0052
dprij8	1	1.8776	0.1706
dprij9	1	5.3296	0.0210
dstav1	1	0.0000	0.9956
dstav2	1	2.7149	0.0994
dstav3	1	0.0257	0.8727
dvek1	1	0.0431	0.8356
dvek2	1	0.0465	0.8293
dvek3	1	0.0560	0.8129
dvek4	1	0.0494	0.8242
dvek5	1	0.0322	0.8576
dvek6	1	0.0264	0.8709
dvzdel1	1	88.7760	<.0001
dvzdel2	1	6.1947	0.0128
kodpoh	1	0.3709	0.5425

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-square	Pr > Chi-square
Intercept		1	-3.4757	34.7736	0.01	0.9204
dprij1	0	1	0.5263	0.1515	12.07	0.0005
dprij10	0	1	0.3038	0.1490	4.16	0.0415
dprij11	0	1	0.3607	0.1513	5.69	0.0171
dprij12	0	1	0.3159	0.1392	5.15	0.0232
dprij13	0	1	0.3123	0.1642	3.62	0.0571
dprij14	0	1	0.2320	0.1303	3.17	0.0751
dprij15	0	1	0.8059	0.2557	9.93	0.0016
dprij16	0	1	0.3452	0.1573	4.82	0.0282
dprij17	0	1	0.3385	0.1330	6.48	0.0109
dprij18	0	1	0.3682	0.1562	5.56	0.0184
dprij19	0	1	0.4679	0.1720	7.40	0.0065
dprij2	0	1	0.3493	0.1233	8.03	0.0046
dprij20	0	1	0.3638	0.1394	6.81	0.0091
dprij21	0	1	0.3532	0.1367	6.68	0.0098
dprij22	0	1	0.2191	0.1153	3.61	0.0573
dprij23	0	1	0.2184	0.1249	3.06	0.0804
dprij24	0	1	0.1199	0.1085	1.22	0.2690
dprij25	0	1	0.4015	0.1255	10.24	0.0014
dprij26	0	1	0.1690	0.1165	2.10	0.1469
dprij27	0	1	0.1484	0.1042	2.03	0.1543
dprij28	0	1	0.2067	0.1072	3.72	0.0538
dprij29	0	1	0.1430	0.1041	1.89	0.1694
dprij3	0	1	0.2923	0.1220	5.74	0.0166
dprij4	0	1	0.3308	0.1219	7.36	0.0067
dprij5	0	1	0.4904	0.1349	13.22	0.0003
dprij6	0	1	0.3591	0.1313	7.48	0.0062
dprij7	0	1	0.3881	0.1388	7.82	0.0052
dprij8	0	1	0.1607	0.1173	1.88	0.1706
dprij9	0	1	0.3207	0.1389	5.33	0.0210
dstav1	0	1	-0.00084	0.1527	0.00	0.9956
dstav2	0	1	-0.2391	0.1451	2.71	0.0994
dstav3	0	1	-0.0230	0.1436	0.03	0.8727

dvek1	0	1	-1.8014	8.6790	0.04	0.8356
dvek2	0	1	-1.8713	8.6778	0.05	0.8293
dvek3	0	1	-2.0542	8.6777	0.06	0.8129
dvek4	0	1	-1.9281	8.6777	0.05	0.8242
dvek5	0	1	-1.5567	8.6778	0.03	0.8576
dvek6	0	1	-1.4101	8.6781	0.03	0.8709
dvzdel1	0	1	0.5200	0.0552	88.78	<.0001
dvzdel2	0	1	0.1267	0.0509	6.19	0.0128
kodpoh	0	1	-0.0499	0.0819	0.37	0.5425

```
                 The FREQ Procedure

            Table of F_net by I_net

      F_net(From: net)     I_net(Into: net)

      Frequency,
      Percent  ,
      Row Pct  ,
      Col Pct  ,0       ,1       ,  Total
      ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
      0        ,   2720 ,     14 ,   2734
               ,  90.67 ,   0.47 ,  91.13
               ,  99.49 ,   0.51 ,
               ,  91.52 ,  50.00 ,
      ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
      1        ,    252 ,     14 ,    266
               ,   8.40 ,   0.47 ,   8.87
               ,  94.74 ,   5.26 ,
               ,   8.48 ,  50.00 ,
      ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
      Total        2972       28     3000
                  99.07     0.93   100.00
```

Multinomický logitový model

```
                              The LOGISTIC Procedure

                                Model Information

          Data Set                      _PROJ_.KREDITNESKOROVANIE
          Response Variable             REC012    REC012
          Number of Response Levels     3
          Number of Observations        1120
          Model                         cumulative logit
          Optimization Technique        Fisher's scoring


                                 Response Profile

                       Ordered                       Total
                         Value       REC012      Frequency

                             1            2            720
                             2            1            220
                             3            0            180

         Probabilities modeled are cumulated over the lower Ordered Values.


                            Model Convergence Status

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


                    Score Test for the Proportional Odds Assumption

                          Chi-Square       DF     Pr > ChiSq

                             51.7821        7         <.0001


                             Model Fit Statistics

                                                   Intercept
                                      Intercept          and
                    Criterion              Only    Covariates

                    AIC                2008.236      1796.713
                    SC                 2018.265      1841.846
                    -2 Log L           2004.236      1778.713


                    Testing Global Null Hypothesis: BETA=0

            Test                    Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio          225.5227        7         <.0001
            Score                     211.1512        7         <.0001
            Wald                      201.6665        7         <.0001



                       Analysis of Maximum Likelihood Estimates

                                                Standard          Wald
       Parameter       DF     Estimate         Error    Chi-Square    Pr > ChiSq

       Intercept 2      1      -1.7406        0.1838       89.6395        <.0001
       Intercept 1      1      -0.4515        0.1756        6.6112        0.0101
       DZLUSLE2         1       0.2709        0.1863        2.1143        0.1459
       DZLUSLE3         1       1.2719        0.1654       59.1051        <.0001
       VEKTRI2          1       0.4225        0.1530        7.6266        0.0058
       VEKTRI3          1       0.8730        0.1656       27.8038        <.0001
       DDEPNOV2         1       1.3541        0.1721       61.8912        <.0001
       DDEPNOV3         1       1.6980        0.2182       60.5515        <.0001
       DDEPNOV4         1       1.9258        0.1891      103.7550        <.0001


                              Odds Ratio Estimates

                                 Point          95% Wald
                  Effect      Estimate      Confidence Limits

                  DZLUSLE2       1.311        0.910        1.889
```

DZLUSLE3	3.567	2.580	4.934
VEKTRI2	1.526	1.130	2.059
VEKTRI3	2.394	1.731	3.312
DDEPNOV2	3.873	2.764	5.427
DDEPNOV3	5.463	3.562	8.379
DDEPNOV4	6.861	4.736	9.938

Association of Predicted Probabilities and Observed Responses

Percent Concordant	69.8	Somers' D	0.436
Percent Discordant	26.2	Gamma	0.455
Percent Tied	4.1	Tau-a	0.229
Pairs	324800	c	0.718