Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Metody analýzy vícerozměrných kontingenčních tabulek

Vypracoval: Bc. Zdeněk Šulc

Vedoucí práce: doc. Ing. Iva Pecáková, CSc.

Rok vypracování: 2012

Čestné prohlášení:

Prohlašuji, že diplomovou práci *Metody analýzy vícerozměrných kontingenčních tabulek* jsem vypracoval samostatně. Veškeré použité podklady, ze kterých jsem čerpal informace, jsou uvedeny v seznamu použité literatury a citovány v textu podle normy ČSN ISO 690.

V Praze dne 8.5.2012 Zdeněk Šulc

Poděkování:

Touto cestou děkuji doc. Ing. Ivě Pecákové, CSc. za cenné rady a připomínky ke zpracování této práce. Dále děkuji společnosti GfK Czech za poskytnutá data a zkušenosti. V neposlední řadě děkuji své rodině a přátelům, kteří mě během studia podporovali.

Abstrakt

Tato práce se zabývá vztahem dvou významných metod analýzy vícerozměrných kontingenčních tabulek, a sice korespondenční analýzou a loglineárními modely. Práce je rozdělena na tři celky. První je věnován základním pojmům kategoriální analýzy dat, především kontingenčním tabulkám a jejich rozdělením. Důraz je kladen zejména na jejich vícerozměrnou formu. Druhý celek představuje nástroje a techniky obou metod v rozsahu, jaký je nutný k jejich praktickému použití a interpretaci jejich výsledků. Praktická aplikace obou metod je obsažena v třetím celku, která je prezentována na datech z marketingového průzkumu. Tento celek popisuje nastavení obou analýz ve statistickém softwaru SPSS i následnou interpretaci jejich výstupů. V závěru práce jsou analyzované metody porovnávány z hlediska jejich použití.

Abstract

This thesis occupies with a relationship of two significant methods of analyzing multivariate contingency tables, namely correspondence analysis and loglinear models. The thesis is divided into three parts. The first one is dedicated to basic terms of categorical data analysis, mainly to contingency tables and their distributions. Primarily, the emphasis is placed on their multidimensional form. The second part presents tools and techniques of both methods in a scope needed for their practical use and interpretation of their results. A practical application of both methods is included in the third part which is presented on the data from a market research. This part describes settings for both analyses in a statistical software SPSS and the subsequent interpretation of their outputs. A comparison of analyzed methods in terms of their use can be found in the conclusion.

Obsah

Úvod

Kontingenční tabulky představují přirozený způsob zobrazování kategoriálních dat. Je možné se s nimi setkat téměř všude, kde je potřeba sdělit nějakou informaci. Ve vědeckých pracích, podnikových reportech, v lékařství, psychologii, ale lze na ně narazit také při běžném čtení v novinách nebo na internetu. V kontingenčních tabulkách se vyskytují pouze kategoriální respektive kategorizovaná data. Analýza takových dat není zdaleka tak prozkoumanou oblastí jako v případě spojitých dat. Je to dáno skutečností, že kategoriální data kladou mnohem vyšší nároky na výpočetní náročnost. Některé statistiky běžné u spojitých dat, jako např. průměr nebo šikmost, je nutné pracně, a někdy nedostatečně, nahrazovat ekvivalenty vhodnými pro kategoriální data. Z těchto důvodů jsem se rozhodl věnovat analýze kontingenčních tabulek podrobněji. V této práci se budu věnovat metodám analýzy vícerozměrných kontingenčních tabulek, konkrétně korespondenční analýze a loglineárním modelům.

Korespondenční analýza je metoda určená k analyzování dat uspořádaných do dvourozměrné (*jednoduchá korespondenční analýza*) nebo vícerozměrné (*vícenásobná korespondenční analýza*) kontingenční tabulky. Klade důraz především na grafickou interpretaci analyzovaných dat. Původně byla vyvinuta ve Francii (*Benzérci*) v šedesátých letech dvacátého století, ale do celosvětového povědomí se dostala v podobě, jak je známá dnes, až v letech osmdesátých (*Greenacre*). Mezitím byla nezávisle pod různými názvy „vynalezena“ v několika dalších zemích. Byla známá jako optimální škálování (*optimal scaling*), optimální skórování (*optimal scoring*) nebo jako analýza homogenity (*homogeneity analysis*). V současné době se hojně využívá především v marketingových průzkumech a psychologii. Je však možné se s ní setkat také v biologii a ekologii, kde se používá např. k přiřazování živočišného druhu k určité oblasti výskytu.

Loglineární modely jsou metodou rovněž sloužící k analyzování kontingenčních tabulek, především vícerozměrných. Samotná podstata analýzy se však od korespondenční analýzy velmi odlišuje. Zaměřuje se hlavně na modelování četností buněk v kontingenční tabulce a odhalování závislostí mezi kategoriálními proměnnými. Teorie loglineárních modelů vznikla na počátku šedesátých let

dvacátého století. V souvislosti s rozvojem výpočetní techniky se jejich metodika zlepšovala až do poloviny sedmdesátých let, kdy došlo například k objevení důležitých vztahů mezi loglineárními a logitovými modely, viz např. (AGRESTI, 2002). V osmdesátých letech nastal rozvoj grafického řešení. Vývoj metodik loglineárních modelů neustal, pokračuje i v současnosti.

Ve své práci jsem se rozhodl prostudovat obě metody, zjistit, nakolik shodné či naopak rozdílné jsou jejich nástroje a postupy, a porovnat jejich výsledky. Dále bych chtěl určit, v jakých případech je vhodné použít korespondenční analýzu a v jakých případech loglineární modely. Jelikož budu aplikovat obě metody na stejná data, mám v plánu zjistit, zda je možné, popř. co je nutné pro to udělat, aby se tyto metody vzájemně doplňovaly. Samotná data v praktické části byla pořízena v rámci marketingového průzkumu, proto je mým dalším záměrem pokus o interpretaci vybraných výsledků a zamyšlení nad tím, jaký je přínos těchto metod v oblasti průzkumu trhu.

Práci jsem rozčlenil na tři celky. První celek je tvořen čtyřmi kapitolami, ve kterých se věnuji popisu kontingenčních tabulek a vztahů, které se v nich vyskytují. V této části se dále zabývám nástroji vhodnými pro diagnostiku kvality modelu, jako je například analýza reziduí nebo testy dobré shody.

V druhém celku, který zahrnuje kapitoly 5 a 6, představím metody korespondenční analýzu a loglineární modely, které budou představeny v rozsahu nutném pro pochopení principů a výstupů použitých v praktické části.

Praktická část je obsažena v třetím celku. Jedná se o data z marketingového průzkumu, na jehož základě se budu zabývat otázkami týkajícími se spokojenosti s podnikatelským prostředím v České republice. Zajímá mě, jak se odpovědi na tyto otázky liší v závislosti na pohlaví, ekonomické aktivitě nebo politické orientaci respondenta.

Účelem této práce je také přispět ke zkoumání použitelnosti korespondenční analýzy nebo loglineárních modelů, především řešení praktických úloh. Většina témat popsaných v této práci proto bude koncipována tak, aby podpořila právě praktickou část.

Kapitola 1
Typy kategoriálních proměnných

Na rozdíl od spojitých proměnných, kategoriální proměnné přirozeně vytváří určitý počet kategorií. Existuje několik typů členění, ta nejdůležitější budou představena v této kapitole.

1.1 Nominální, ordinální a kardinální proměnné

V kategoriální analýze dat je možné se setkat s nominálními, ordinálními a kardinálními proměnnými. Ty představují různé typy stupnic (škál). Nominální proměnné nelze uspořádat podle žádného objektivního kritéria. Může se jednat například o náboženskou příslušnost nebo o pohlaví. Číslice, kterými jsou značeny, slouží pouze jako kódy, tudíž nezáleží na jejich hodnotě. Statistické metody s nominálními proměnnými podávají stejné výsledky nezávisle na pořadí proměnných, a opírají se tak pouze o jejich četnosti. Tato skutečnost způsobuje, že možnosti statistických metod založených na nominálních proměnných jsou nejomezenější ze všech druhů kategoriálních proměnných.

Často se vyskytují jevy, které mají přirozeně uspořádané kategorie. Tak je možné určit, která kategorie má vyšší popř. nižší stupeň sledovaného znaku. Takové proměnné se nazývají ordinální. Vyjadřovat mohou vzdělání (*základní, střední, vysokoškolské*), úroveň kvality výrobku (*nízká, střední, vysoká*) nebo třeba frekvenci konzumace mléčných výrobků (*nikdy, zřídka, středně, často, denně*). U ordinálních proměnných je pořadí sledovaných znaků známé, ale není možné určit, do jaké míry jsou jednotlivé znaky vzdálené. Statistické analýzy založené na ordinálních proměnných podávají výsledky beroucí v úvahu pořadí sledovaných znaků, jsou však poměrně výpočetně náročné.

Charakter ordinálních proměnných získávají také kardinální veličiny, pokud jsou transformovány do intervalů. Kardinální proměnné představují počty měrných jednotek k vyjádření úrovně určitého znaku, např. počtu let u proměnné *Věk*. Tyto proměnné je potřeba seskupit do několika intervalů, aby je bylo možné použít např.

v kontingenční tabulce. Někdy však není možné sledovanou proměnnou získat ve své původní kardinální škále. Především v oblasti dotazování bývají některé proměnné kardinálního typu zjišťovány přímo ve formě intervalů. Důvodem je taktika zjišťování některých údajů. Typickou proměnnou je *Příjem domácnosti*, na kterou respondenti velmi neradi odpovídají. Na transformovanou otázku do formy několika intervalů odpovědí mnohem častěji. Menší míra chybějících pozorování je tak vykoupena menším množstvím informace v datech a jejich náročnější analýzou.

Použitá škála ovlivňuje typ analýzy dat, kterou je možné použít. Pro danou škálu lze použít metody pro ni určené, ale také metody pro škály postavené v hierarchii níže. Např. u ordinálních proměnných mohou být použity metody pro ordinální a nominální proměnné, ale nikoliv metody určené pro kardinální proměnné. Jsou-li použity metody nižší škály, než je možné, dochází ke ztrátě informace. Proto je doporučeno používat takovou škálu, která odpovídá studovaným datům.

1.2 Vysvětlující a vysvětlované proměnné

Proměnné, které se vyskytují v kontingenčních tabulkách, mohou být vysvětlující (*explanatory*) nebo vysvětlované (*response*). Vysvětlovanou proměnnou je např. *Souhlas s trestem smrti*, vysvětlující proměnnou *Věk*. Většinou je předmětem zájmu situace, jak se vysvětlovaná proměnná mění při různých úrovních vysvětlující proměnné. V některých případech může být také zkoumán vztah mezi dvěma vysvětlovanými proměnnými, např., souvislost S*ouhlasu s trestem smrti* se souhlasem s výrokem, že S*oudy provádějí svoji práci kvalitně*.

Kapitola 2
Pravděpodobnostní rozdělení kategoriálních dat

Aby bylo možné provádět statistické úsudky, je nutné znát empirické pravděpodobnostní rozdělení analyzovaných proměnných. U analýzy spojitých proměnných hraje dominantní roli normální rozdělení. Naproti tomu, při analýze kategoriálních dat mají výjimečné postavení binomické, multinomické a Poissonovo rozdělení.

2.1 Binomické rozdělení

V analýze kategoriálních dat je poměrně častý případ, že n nezávislých náhodných pozorování $y_1, y_2, \ldots, y_n$ proměnné Y může nabývat pouze dvou stavů, které vyjadřují buď „úspěch“, značený hodnotou jedna, nebo „neúspěch“, značený nulou. Úspěch nastává s pravděpodobností $P(Y_i = 1) = \pi$ a neúspěch s pravděpodobností $P(Y_i = 0) = 1 - \pi$. Jednotlivé veličiny Y_i mají alternativní rozdělení. Veličina $Y = \sum_i^n Y_i$, která představuje celkový počet „úspěchů“, má binomické rozdělení. Pravděpodobnost, že u vysvětlované veličiny Y s binomickým rozdělením s parametry n a π, bude právě y úspěšných pokusů, je vyjádřena vztahem:

$$P(y) = \binom{n}{y} \pi^y (1-\pi)^{(n-y)} \qquad y = 1, 2, \ldots, n \tag{2.1}$$

Typickou proměnnou, která má binomické rozdělení, je proměnná *Pohlaví*. Pro každé pozorování může nabývat dvou hodnot. Hodnotou jedna se obvykle značí muž, hodnotou nula žena. Pravděpodobnost „výskytu“ obou pohlaví je téměř shodná, tedy $\pi = (1 - \pi) = 0{,}5$. Při výběru o rozsahu $n = 10$ bude pravděpodobnost vybrání pravě jedné ženy podle rovnice (2.1) rovna $P(1) = 0{,}0098$, tedy necelé jedno procento.

Střední hodnotu binomického rozdělení je možné vyjádřit jako součin počtu pozorování n a parametru π. Pro výše uvedený příklad je střední hodnota rovna pěti.

$$E(Y) = n\pi \tag{2.2}$$

Rozptyl binomického rozdělení je vyjádřen vztahem (2.3). Pro uvedený příklad je roven hodnotě 2,5.

$$D(Y) = n\pi(1 - \pi) \tag{2.3}$$

Binomické rozdělení je nesymetrické kromě případu, kdy $\pi = 0{,}5$. Čím více se π blíží hodnotě nula nebo jedna, tím je zešikmenější. Při velkém n může být aproximováno normálním rozdělením. Potřebná velikost n záleží na parametru π. Pokud se parametr blíží $0{,}5$, stačí, aby n bylo větší než 10. Při velmi šikmých rozděleních ($\pi = 0{,}9$) musí být n větší než 50.

2.2 Multinomické rozdělení

Vysvětlovaná proměnná Y má multinomické rozdělení, pokud může nabývat v n nezávislých pokusech k stavů (tj. kategorií). I-tý vícerozměrný pokus s k možnými kategoriemi může být zapsán jako $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \dots, y_{ik})$, přičemž $\sum_j^k y_{ij} = 1$. Pro každou kategorii je určena pravděpodobnost úspěchu $\pi_1, \pi_2, \dots, \pi_k$, přičemž jejich součet $\sum_j^k \pi_j$ se musí rovnat jedné. Pro n nezávislých náhodných pozorování je pravděpodobnost, že právě n_1 pozorování nabude kategorie 1, n_2 kategorie 2, … n_k kategorie k, kde $\sum_j^k n_j = n$, rovna:

$$P(n_1, n_2, \dots, n_k) = \left(\frac{n!}{n_1!, n_2!, \dots, n_k!}\right) \pi_1^{n_1} \pi_2^{n_2} \dots \pi_k^{n_k} \tag{2.4}$$

Binomické rozdělení je zvláštním případem multinomického za situace, že $k = 2$. Multinomické rozdělení je vícerozměrné, jeho marginální rozdělení n_j jsou binomická. V (AGRESTI, 2007) je uvedeno, že většina metod určených pro kategoriální data předpokládá binomické rozdělení u jedné kategorie a multinomické rozdělení pro sadu několika kategorií.

Multinomické rozdělení má např. proměnná *Věková kategorie*, která je tvořena pěti možnými věkovými intervaly. Každý z n respondentů tak představuje náhodný vícerozměrný pokus. Je-li např. $\boldsymbol{y_1} = (0,0,0,1,0)$, pak první respondent patří do čtvrté věkové kategorie.

2.3 Poissonovo rozdělení

Často není přesný počet pokusů n známý. Většinou proto, že je příliš vysoký nebo pozorovaný jev stále ještě probíhá. Nejčastěji se v takových případech používá Poissonovo rozdělení. Pravděpodobností funkce

$$p(y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = 0,1,2,\ldots \tag{2.5}$$

má pouze jeden parametr λ, který je současně střední hodnotou i rozptylem. Jeho celočíselná část je zároveň i módem. Jedná se o zešikmené rozdělení, které se s rostoucím n blíží k normálnímu. V případě, že jde o velmi vysokou celkovou četnost ($n > 1000$) a pravděpodobnost nastoupení jevu π je velmi nízká ($\pi < 0{,}001$), dá se binomické rozdělení aproximovat rozdělením Poissonovým pomocí vztahu: $\lambda = n\pi$.

Vzhledem k faktu, že střední hodnota Poissonova rozdělení je rovna jeho rozptylu, je v případě většího průměrného počtu nastoupených jevů i jejich rozptyl větší. To souvisí s jevem zvaným *overdispersion*, při kterém napozorované četnosti vykazují vyšší variabilitu, než která je predikována modelem. *Overdispersion* je způsobena mylným předpokladem stejných pravděpodobností všech výběrových jednotek. Ve skutečnosti se však pravděpodobnosti liší kvůli zanedbatelným vlivům, které nejsou sledovány. Tato nadbytečná variabilita způsobuje větší variabilitu četností, než je předpovězeno Poissonovým modelem.

Overdispersion se neomezuje pouze na Poissonovo rozdělení. Vyskytuje se také u binomického a multinomického rozdělení v případech, ve kterých je skutečné rozdělení tvořeno mixem různých binomických nebo multinomických rozdělení, která mají různé parametry v důsledku nezahrnutí zanedbatelných vlivů.

Kapitola 3
Kontingenční tabulky

Kontingenční tabulky představují východisko pro analyzování vztahů mezi kategoriálními proměnnými. Jsou výsledkem dvourozměrného nebo vícerozměrného třídění dat souboru.

3.1 Dvourozměrné kontingenční tabulky

V dvourozměrném třídění jsou kontingenční tabulky definovány jako převážně obdélníkové tabulky s r řádky kategorie X a s sloupci kategorie Y, v jejichž průsečících se nacházejí četnosti všech ij kombinací. Symbolika používaná v kontingenčních tabulkách je představena v tabulce T1.

T1 – Kontingenční tabulka; zdroj: (PECÁKOVÁ, 2011)

	y_1	y_2	…	y_s	n_{i+}
x_1	n_{11}	n_{12}	…	n_{1s}	n_{1+}
x_1	n_{21}	n_{22}	…	n_{2s}	n_{2+}
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rs}	n_{r+}
n_{+j}	n_{+1}	n_{+2}	…	n_{+s}	n

Hodnoty x_i a y_j znázorňují kategorie veličin X a Y. Prvky n_{ij} představují sdružené absolutní četnosti kategorií veličin X a Y. V posledním řádku a sloupci se nacházejí absolutní marginální četnosti n_{i+} a n_{+j}. Jedná se vlastně o řádkové a sloupcové součty. Marginální četnosti podávají pohled na požadovanou proměnnou bez ohledu na hodnoty druhé proměnné.

3.2 Vícerozměrné kontingenční tabulky

Kontingenční tabulka s vícerozměrným tříděním dat je definována jako tabulka s r řádky kategorie X, s sloupci kategorie Y, v vrstvami kategorie Z, atd. Může být pojata dvěma způsoby.

První způsob je vlastní hlavně statistickým softwarům. Jde o vrstvení jednotlivých rozměrů kontingenční tabulky „nad sebe". Přestože se jedná o zobrazení, které nejvíce odpovídá struktuře dat, v praxi se nepoužívá, protože neumožňuje zobrazit vícerozměrnou kontingenční tabulku v dvourozměrném prostoru. Proto se používá alternativní zobrazení, které ke kontingenční tabulce přidává další sloupec resp. řádek obsahující kategorie dalších proměnných, podle kterých jsou data dále tříděna. Schéma trojrozměrné kontingenční tabulky je zobrazeno v tabulce T2. Při jeho srovnání s dvourozměrnou tabulkou T1 je dobře patrné, jak se vztahy s třetím rozměrem zkomplikovaly.

T2 – Vícerozměrná kontingenční tabulka; zdroj: (PECÁKOVÁ, 2011)

		y_1	y_2	…	y_s	n_{i+k}
z_1	x_1	n_{111}	n_{121}	…	n_{1s1}	n_{1+1}
	x_2	n_{211}	n_{221}	…	n_{2s1}	n_{2+1}
	…	…	…	…	…	…
	x_r	n_{r11}	n_{r21}	…	n_{rs1}	n_{r+1}
	n_{+j1}	n_{+11}	n_{+21}	…	n_{+s1}	n_{++1}
z_2	x_1	n_{112}	n_{122}	…	n_{1s2}	n_{1+2}
	x_2	n_{212}	n_{222}	…	n_{2s2}	n_{2+2}
	…	…	…	…	…	…
	x_r	n_{r12}	n_{r22}	…	n_{rs2}	n_{r+2}
	n_{+j2}	n_{+12}	n_{+22}	…	n_{+s2}	n_{++2}
…	…	…	…	…	…	…
z_v	x_1	n_{11v}	n_{12v}	…	n_{1sv}	n_{1+v}
	x_2	n_{21v}	n_{22v}	…	n_{2sv}	n_{2+v}
	…	…	…	…	…	…
	x_r	n_{r1v}	n_{r2v}	…	n_{rsv}	n_{r+v}
	n_{+jk}	n_{+1v}	n_{+2v}	…	n_{+sv}	n_{++v}
n_{+j+}		n_{+1+}	n_{+2+}	…	n_{+s+}	n

Zobrazení kontingenční tabulky bez ohledu na třídění podle hodnot vrstvy, která není v daném okamžiku předmětem zájmu, se nazývá *marginální kontingenční tabulka*. Například četnosti marginální tabulky vytvořené z tabulky T2 bez ohledu na

proměnnou Z, jsou pak určeny jako $n_{11+} = \sum_k n_{11k}$, $n_{12+} = \sum_k n_{12k}$, atd. Vztahy v marginální tabulce se nazývají *marginální asociace* (*marginal asociations*).

Parciální tabulky, které třídí X a Y při různých úrovních proměnné Z, rozdělují vícerozměrnou kontingenční tabulku na několik tabulek o menším rozměru. Například trojrozměrná kontingenční tabulka T2 je roztříděna na tři dvourozměrné kontingenční tabulky podle úrovní proměnné Z. Asociační vztahy mezi proměnnými v parciální tabulce se nazývají *podmíněné asociace* (*conditional associations*). Ty se někdy mohou i výrazně lišit od marginálních asociací. Proto je při analýzách vhodné zkoumat jak marginální, tak parciální asociace.

3.3 Pravděpodobnostní struktura kontingenčních tabulek

U kontingenčních tabulek jsou rozlišovány tři typy pravděpodobnostní struktury:

- sdružená
- marginální
- podmíněná

3.3.1 Pravděpodobností struktura dvourozměrných kontingenčních tabulek

Sdružené pravděpodobnosti π_{ij} vyjadřují pravděpodobnost, že kombinace veličin X a Y nabude konkrétní hodnoty ij, tedy $P(X = i, Y = j) = \pi_{ij}$. Součet přes všechny π_{ij} je roven jedné.

Marginální pravděpodobnosti jsou řádkové $\pi_{i+} = \sum_j \pi_{ij}$ a sloupcové $\pi_{+j} = \sum_i \pi_{ij}$ součty sdružených pravděpodobností. Vyjadřují pravděpodobnostní strukturu jedné proměnné bez ohledu na úroveň druhé proměnné.

Kontingenční tabulky, které mají jednu proměnnou vysvětlovanou a druhou vysvětlující, mají často pro každou úroveň vysvětlující proměnné jiné pravděpodobnostní rozdělení. Rozdíly mezi rozděleními jednotlivých úrovní se určují pomocí pravděpodobností vztažených ke každé úrovni vysvětlující proměnné X. Ty se nazývají *podmíněné pravděpodobnosti* a jsou definovány jako podíl sdružených a příslušných marginálních pravděpodobností:

$$\pi_{j/i} = \frac{\pi_{ij}}{\pi_{i+}} \tag{3.1}$$

V praxi nejsou pravděpodobnosti π_{ij} známy předem, proto se používají jejich výběrové protějšky, které se získají z výběrových četností n_{ij} v kontingenční tabulce. Značí se malým písmenem p a jsou definovány podle následujících vztahů:

- sdružené relativní četnosti: $p_{ij} = \frac{n_{ij}}{n}$
- marginální relativní četnosti: $p_{i+} = \frac{n_{i+}}{n}$ resp. $p_{+j} = \frac{n_{+j}}{n}$
- podmíněné relativní četnosti: $p_{j/i} = \frac{p_{ij}}{p_{i+}}$

Sdružené a marginální relativní četnosti mohou být uspořádány do tabulky relativních četností představené v tabulce T3. Suma všech sdružených i marginálních četností dává hodnotu jedna.

T3 – Tabulka relativních četností

	y_1	y_2	…	y_s	p_{i+}
x_1	p_{11}	p_{12}	…	p_{1s}	p_{1+}
x_1	p_{21}	p_{22}	…	p_{2s}	p_{2+}
…	…	…	…	…	…
x_r	p_{r1}	p_{r2}	…	p_{rs}	p_{r+}
p_{+j}	p_{+1}	p_{+2}	…	p_{+s}	1

3.3.2 Pravděpodobnostní struktura vícerozměrných kontingenčních tabulek

Pravděpodobnostní vztahy ve vícerozměrných kontingenčních tabulkách jsou rozšířením vztahů dvourozměrných. V této kapitole jsou prezentovány na nejjednodušší vícerozměrné, tedy trojrozměrné, kontingenční tabulce.

Sdružené pravděpodobnosti $\pi_{ijk} = P(X = i, Y = j, Z = k)$ vyjadřují, s jakou pravděpodobností bude náhodné pozorování zařazeno do buňky (i, j, k).

Marginální pravděpodobnosti v trojrozměrné kontingenční tabulce se získají jako součet podmnožin sdružených pravděpodobností π_{ijk} při ignorování třetí proměnné. Na rozdíl od dvourozměrných kontingenčních tabulek, v trojrozměrné tabulce existují marginální pravděpodobnosti jedné a dvou proměnných. Například marginální pravděpodobnost jedné proměnné Y lze vyjádřit vztahem: $\pi_{+j+} = P(Y = j)$. Marginální pravděpodobnosti dvou proměnných Y a Z jsou značeny $\pi_{+jk} = P(Y = j, Z = k)$. Při neexistenci proměnné X by se jednalo o sdružené pravděpodobnosti dvourozměrné tabulky YZ.

Podmíněné rozdělení ve vícerozměrné tabulce získáme analogicky jako v dvourozměrné tabulce, tedy jako poměr sdružených a marginálních pravděpodobností. Protože v trojrozměrné tabulce existují dva typy marginálního rozdělení, existují zde také dva typy podmíněného rozdělení. První typ zjišťuje změnu sdruženého rozdělení např. proměnné Y při různých úrovních proměnných X a Z, tedy $\pi_{j|ik} = \frac{\pi_{ijk}}{\pi_{+j+}}$, kde $\sum_j \pi_{+j+} = 1$. Druhý typ zkoumá, jak se změní např. sdružené rozdělení proměnných Y a Z při různých úrovních proměnné X. Lze jej vyjádřit vztahem $\pi_{jk|i} = \frac{\pi_{ijk}}{\pi_{+jk}}$, přičemž platí, že $\sum_{jk} \pi_{+jk} = 1$.

3.4 Typy pravděpodobnostních rozdělení v kontingenčních tabulkách

Četnosti buněk v kontingenční tabulce jsou řízeny některým z pravděpodobnostních rozdělení představených v kapitole 2, především Poissonovým nebo multinomickým.

Multinomické rozdělení předpokládá pevně stanovenou celkovou četnost n, která je náhodně a nezávisle roztříděna do buněk kontingenční tabulky se známými sdruženými pravděpodobností π_{ij}. Náhodná veličina Y_{ij} představuje počet pozorování v i -tém řádku a j -tém sloupci a n_{ij} její pozorovanou hodnotu. Potom je pravděpodobnostní funkce četností buněk kontingenční tabulky rovna:

$$P(\boldsymbol{Y} = \boldsymbol{y} = n_{11}, \dots, n_{ij}) = \frac{n!}{n_{11}!, \dots, n_{ij}!} \prod_i \prod_j \pi_{ij}^{n_{ij}} \qquad (3.2)$$

kde $\boldsymbol{Y}$ je náhodný vektor a $\boldsymbol{y}$ je vektor pozorovaných hodnot. Výraz v rovnici (3.2) vyjadřuje pravděpodobnost, že v buňce $(1,1)$ bude právě n_{11} pozorování, až buňce (i,j) n_{ij} pozorování. Samotný zlomek tohoto výrazu vyjadřuje, kolika způsoby je možno získat n_{11} pozorování v buňce (1,1) až n_{ij} pozorování v buňce (i,j).

Poissonovo rozdělení modeluje četnosti buněk kontingenční tabulky Y_{ij} jako nezávislé náhodné veličiny Poissonova rozdělení s parametry μ_{ij}. Pravděpodobnostní funkce vyjadřující pravděpodobnost, že každá četnost v tabulce bude mít konkrétní hodnotu n_{ij} je dána vztahem (3.3).

$$P(\boldsymbol{Y} = \boldsymbol{y} = n_{11}, \dots, n_{ij}) = \frac{e^{-\mu_{ij}}}{n_{ij}!}\mu_{ij}^{n_{ij}} \tag{3.3}$$

3.4.1 Vztah mezi multinomickým a Poissonovým rozdělením

Vztah mezi multinomickým a Poissonovo rozdělením je dobře ilustrován na příkladu, který uveden v (AGRESTI, 2002). Výzkumníci plánovali analyzovat vztah mezi používáním bezpečnostních pásů (*ano, ne*) a úmrtností při automobilových haváriích na dálnicích (*smrtelná, bez ztráty životů*). Plánovali vytvořit seznam všech havárií, které nastanou během právě probíhajícího roku. Celkový počet havárií je v tomto případě neznámý, proto je vhodné použít Poissonovo rozdělení se čtyřmi náhodnými veličinami Y_{ij} s neznámými průměry $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$. Celková četnost $n = \sum Y_{ij}$ má také Poissonovo rozdělení s parametrem $\sum \mu_{ij}$. Pokud by výzkumníci získali 200 náhodně vybraných havárií z policejních záznamů za minulý rok, mohli by použít multinomické rozdělení s $n = 200$ nezávislými pokusy a sdruženými pravděpodobnostmi $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$.

Z výše uvedeného příkladu je zřejmé, že je-li stanovena celková četnost n, náhodné veličiny Y_{ij} již nebudou mít Poissonovo rozdělení, protože se stanou vzájemně závislé, neboť žádná jejich četnost n_{ij} nebude moci překročit celkovou četnost n. Z Poissonova rozdělení se tak stane rozdělení multinomické. Z pravděpodobnostního hlediska je tedy multinomické rozdělení Poissonovo rozdělení omezené podmínkou, že $\sum Y_i = n$. Mnoho kategoriálních analýz předpokládá multinomické rozdělení. V (AGRESTI, 2002) je dokázáno, že takové analýzy mají obvykle shodné odhady parametrů jako ty, které předpokládají Poissonovo rozdělení, a to díky podobnosti jejich věrohodnostních funkcí. Proto se pro většinu analýz volí Poissonovo rozdělení, u kterého není potřeba znalosti celkové četnosti n.

3.5 Typy nezávislosti v kontingenčních tabulkách

Při analyzování nezávislostí v kontingenčních tabulkách je nutné rozlišovat, zda se jedná o tabulky dvourozměrné nebo vícerozměrné. Také je dobré zdůraznit, že závislost znamená v kontextu kategoriálních dat *asociaci*, zatímco nezávislost je definována jako stav *bez asociace*.

3.5.1 Typy nezávislosti dvourozměrných kontingenčních tabulek

V případě dvourozměrných tabulek lze vzájemnou nezávislost dvou vysvětlovaných veličin X a Y prokázat v závislosti na typu proměnných pomocí:

- sdruženého rozdělení
- podmíněného rozdělení

Pomocí sdruženého rozdělení se ověřuje nezávislost dvou vysvětlovaných proměnných. Veličiny jsou nezávislé, pokud jsou všechny pravděpodobnosti sdruženého rozdělení π_{ij} rovny součinu svých marginálních pravděpodobností π_{i+} a π_{+j}, tedy:

$$\pi_{ij} = \pi_{i+}\pi_{+j} \tag{3.4}$$

Je-li kontingenční tabulka tvořena vysvětlující a vysvětlovanou proměnnou, pro definici nezávislosti stačí, jsou-li podmíněné pravděpodobnosti vysvětlované proměnné shodné pro každou úroveň vysvětlující proměnné.

$$\pi_{j|i} = \pi_{j|i'} \tag{3.5}$$

Například v tabulce T4, která popisuje míru v posmrtný život, se podmíněné pravděpodobnosti v obou řádcích téměř shodují. Je tak zřejmé, že víra v posmrtný život nezávisí na pohlaví.

T4 – Nezávislost v kontingenční tabulce; zdroj: (AGRESTI, 2002)

	Ano	Ne/Neví	Celkem
Ženy	509	116	625
Muži	398	104	502
Celkem	907	220	1127

	Ano	Ne/Neví	Celkem
Ženy	81%	19%	100%
Muži	79%	21%	100%
Celkem	80%	20%	100%

Jak je z tabulky T4 patrné, hodnoty pro muže a ženy se zcela neshodují. V praxi je téměř nemožné, aby se podařilo nalézt tabulku s dokonale nezávislými kategoriálními proměnnými. Proto se používají testy ověřující nezávislost v kontingenční tabulce, především testy založené na statistice chí-kvadrát, které budou představeny v kapitole 4.

3.5.2 Typy nezávislosti vícerozměrných kontingenčních tabulek

U vícerozměrných kontingenčních tabulek je situace složitější, existuje zde velké množství vztahů mezi proměnnými. V trojrozměrné kontingenční tabulce s proměnnými X, Y a Z existuje pět možných typů nezávislosti:

1. VZÁJEMNÁ NEZÁVISLOST (*mutual independence*)

   Všechny proměnné jsou vzájemně nezávislé. Lze ji možno vyjádřit pomocí pravděpodobností:

$$P(X = i, Y = j, Z = k) = P(X = i)\, P(Y = j)\, P(Z = k) \tag{3.6}$$

$$\pi_{ijk} = \pi_{i++}\, \pi_{+j+}\, \pi_{++k} \tag{3.7}$$

   Vzájemná nezávislost znamená automaticky nezávislost sdruženou. Obráceně tento vztah neplatí.

2. SDRUŽENÁ NEZÁVISLOST (*joint independence)*

   Jedná se o asociaci dvou proměnných nezávislých na třetí proměnné.

$$\pi_{ijk} = P(X = i, Y = j)\, P(Z = k) = \pi_{ij+}\, \pi_{++k} \tag{3.8}$$

3. MARGINÁLNÍ NEZÁVISLOST (*marginal independence*)

   Proměnné X a Z jsou marginálně nezávislé, jsou-li nezávislé v marginální tabulce. Proměnná Y není brána v úvahu.

$$\pi_{i+k} = P(X = i, Z = k) = \pi_{i++}\, \pi_{++k} \tag{3.9}$$

Jsou-li dvě proměnné sdruženě nezávislé, jsou zároveň marginálně nezávislé. Tento vztah je pouze jednosměrný, obráceně neplatí.

4. PODMÍNĚNÁ NEZÁVISLOST (*conditional independence*)

   Koncept podmíněné nezávislosti hraje velmi důležitou roli v mnoha statistických modelech, např. ve faktorové analýze. Jsou-li proměnné X a Y nezávislé při jednotlivých úrovních proměnné Z, říkáme o nich, že jsou podmíněně nezávislé.

$$\pi_{ijk} = P(Z = k)\, P(X = i, Y = j \mid Z = k) = \pi_{i++}\, \pi_{j|i}\, \pi_{k|i} \qquad (3.10)$$

   Podmíněná nezávislost neznamená automaticky nezávislost marginální a obráceně. Někdy má dokonce marginální a podmíněná asociace proměnných opačný směr. Tento jev se nazývá *Simpsonův paradox* (*Simpson's paradox*), který je detailně popsán např. v (AGRESTI, 2007).

5. HOMOGENNÍ ASOCIACE (*homogeneous association*)

   Homogenní asociace představuje zvláštní případ, při kterém jsou všechny páry proměnných podmíněně závislé při každé úrovni zbývající proměnné, na které jsou podmíněně nezávislé. Homogenní asociace je symetrická vlastnost, platí pro každý pár proměnných vzhledem k třetí proměnné. V případě, že tato vlastnost aspoň pro jeden pár proměnných neplatí, nejedná se o homogenní asociaci.

Kapitola 4
Testy založené na statistice chí-kvadrát a testování reziduí

Při analyzování vztahů v kontingenční tabulce se často používají testy založené na statistice chí-kvadrát. Testovými kritérii pak nejčastěji bývají Pearsonův chí-kvadrát test X^2 a statistika G^2 založená na věrohodnostním poměru Λ.

4.1 Chí-kvadrát test

„Pearsonův chí-kvadrát test se používá pro zjištění, zda vzorek dat odpovídá předpokládanému rozdělení.“ (OBITKO) Testovaná hypotéza H_0 svědčí pro předpokládané rozdělení. Testové kritérium, na jehož základě se testovaná hypotéza ověřuje, má tvar:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4.1}$$

kde O_i je pozorovaná frekvence a E_i je očekávaná frekvence. Toto testové kritérium je pak porovnáváno s příslušným kvantilem a $(I-1)(J-1)$ stupni volnosti chí-kvadrát rozdělení. Je-li hodnota testového kritéria vyšší než hodnota kvantilu, je hypotéza H_0 zamítnuta ve prospěch alternativní hypotézy. Například při ověřování nezávislosti ve čtyřpolní tabulce tak musí být hodnota testového kritéria vyšší než $X^2(1) = 3{,}841$, aby mohla být prokázána závislost proměnných na 95% hladině významnosti.

Významná vlastnost chí-kvadrát statistik je, že jejich součtem respektive rozdílem vznikají nové chí-kvadrát statistiky. Součet chí-kvadrát rozdělení s df_1 stupni a chí-kvadrát rozdělení s df_2 stupni volnosti má opět chí-kvadrát rozdělení s $df_1 + df_2$ stupni volnosti. Obdobně, jakékoliv chí-kvadrát rozdělení s počtem stupňů volnosti vyšším než jedna, lze rozdělit na několik chí-kvadrát rozdělení s menším počtem stupňů volnosti.

Chí-kvadrát test se používá k:

- testům dobré shody
    - ověření nezávislosti proměnných v kontingenční tabulce
    - testování shody modelu s napozorovanými daty
- testování významnosti změn při přidání/ubrání parametru do/z modelu

Testy dobré shody se používají k testování nezávislosti a ověření shody modelu s empirickými daty. Nezávislost proměnných v kontingenční tabulce se ověřuje stanovením nulové hypotézy: $H_0 = \pi_{i+}\pi_{+j}$. Pro dané n se pak vypočítají očekávané četnosti kontingenční tabulky: $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$. Ty se dále porovnávají s napozorovanými četnostmi. Protože teoretické pravděpodobnosti π se v praxi neznají, nahrazují se svými výběrovými protějšky. Odhady očekávaných četností se vypočítají pomocí vzorce: $\hat{\mu}_{ij} = np_{i+}p_{+j}$. Testové kritérium pak má následující tvar:

$$X^2 = \sum \frac{\left(n_{ij} - \hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}} \qquad (4.2)$$

Testování shody modelu s napozorovanými daty je analogické k testování nezávislosti. Jediný rozdíl je, že místo nulové hypotézy popisující nezávislost, popisuje tato hypotéza rozdělení za platnosti daného modelu.

Testování významnosti změn při přidání nebo ubrání parametru v modelu vychází ze zmíněné vlastnosti týkající se součtu a rozdílu chí-kvadrát rozdělení. Používá se v metodách, jako jsou *Backward elimination* nebo *Forward selection*, jejichž princip bude vysvětlen v kapitole 6.3.

4.2 Věrohodnostní poměr

Princip testu věrohodnostním poměrem spočívá v určení parametrů věrohodnostní funkce, které maximalizují věrohodnostní funkci za platnosti H_0 a za platnosti všech parametrů bez omezení, kdy hypotéza H_0 může být pravdivá, ale také nemusí. Oba parametry jsou pak porovnávány podle vztahu (4.3). Hodnoty blízké nule svědčí pro alternativní hypotézu.

$$\Lambda = \frac{H_{platnost\ nulové\ hypotézy}}{H_{platnost\ parametrů\ bez\ omezení}} \tag{4.3}$$

Samotný poměr Λ nemá zcela ideální vlastnosti. V (AGRESTI, 2007) je například uveden příklad, při kterém má maximalizovaná věrohodnostní funkce parametrů bez omezení mnohem vyšší hodnotu než v případě platnosti nulové hypotézy. Věrohodnostní poměr Λ se potom „příliš“ blíží nule. V takovém případě je obtížné kvantifikovat míru rozdílu nulové a alternativní hypotézy, např. pomocí p-hodnoty. Proto se používá logaritmická transformace (4.4), která tyto negativní vlastnosti nemá.

$$G^2 = -2 \ln \Lambda \tag{4.4}$$

Další výhodou této transformace je skutečnost, že výsledné kritérium G^2 má aproximativně chí-kvadrát rozdělení. U dvourozměrných kontingenčních tabulek předpokládajících multinomické rozdělení se používá zjednodušené kritérium (4.5).

$$G^2 = -2 \sum n_{ij} \ln\left(\frac{n_{ij}}{\mu_{ij}}\right) \tag{4.5}$$

Testové kritérium G^2 má obdobné vlastnosti jako statistika X^2. Se vzrůstajícím n konverguje jejich rozdíl k nule. Na rozdíl od X^2 se součty parciálních chí-kvadrát čtverců u G^2 přesně rovnají své původní hodnotě. Statistika X^2 tuto vlastnost nemá, součet parciálních chí-kvadrát čtverců se mírně odlišuje od původní hodnoty. Statistika X^2 naopak rychleji s rostoucím n konverguje k teoretickému rozdělení. Testová kritéria X^2 i G^2 tak představují alternativní způsoby ověřování nezávislosti v kontingenční tabulce. Obě mají chí-kvadrát rozdělení s $(I-1)(J-1)$ stupni volnosti. Jejich hodnoty se většinou příliš neliší.

4.3 Využití reziduí při testování nezávislosti

Pokud nejsou veličiny X a Y nezávislé, respektive neodpovídají předpokládanému modelu, je předmětem zkoumání, jaké buňky jsou příčinou tohoto jevu. K tomuto účelu slouží rezidua, která vyjadřují rozdíl mezi výběrovými a očekávanými četnostmi. Prostý rozdíl $n_{ij} - \mu_{ij}$ je však nedostatečný, protože u buněk, které mají vyšší očekávané četnosti, se také vyskytují větší absolutní rozdíly. Rozdíl $n_{ij} - \mu_{ij}$ proto vhodné vydělit jeho standardní chybou:

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\left[\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})\right]^{1/2}} \tag{4.6}$$

Výsledná rezidua se nazývají *adjustovaná* (*adjusted residuals*). Při větším rozsahu výběru konverguje jejich rozdělení k normovanému normálnímu rozdělení. Pak je možné porovnávat hodnoty reziduí s teoretickými kvantily tohoto rozdělení. Hodnota rezidua může při platnosti H_0 překročit kvantil 1,96 maximálně v 5 % případů.

Velmi často se rezidua nahrazují znaménkovým schématem, jako je tomu u tabulky T5.

T5 – Znaménkové schéma 1; zdroj: (OBITKO)

	– –	–	+	+ +
I	0	0	–	– –
II	0	+ +	0	0
III	+	0	0	0
IV	0	+	0	0

Rezidua jsou navíc rozdělena podle úrovně významnosti, které dosahují.

T6 – Znaménkové schéma 2; zdroj: (ACREA CR)

kvantil |μ|	hladina významnosti α	symbol
>=3,29	0,001	+ + + (– – –)
<2,58;2,39)	0,01	+ + (– –)
<1,96;2,58)	0,05	+ (–)
<1,96	x	0

Kapitola 5
Korespondenční analýza

Korespondenční analýza vychází z analyzování vztahů kategorií proměnných obsažených v kontingenční tabulce. Obdobně, jako faktorová analýza, která hledá nové smysluplné proměnné, si klade za cíl zjednodušení struktury dat vytvořením několika *latentních* proměnných, které představují osy redukovaného souřadnicového systému. Pomocí těchto proměnných je vysvětlena značná část variability sledovaných dat. Hlavním výstupem korespondenční analýzy je grafické zobrazení zvané *korespondenční mapa*. Jelikož je možné zobrazit maximálně trojrozměrný prostor, obvykle se nepoužívá vyšší počet latentních proměnných než tři. Většinou si však korespondenční analýza vystačí s dvourozměrným zobrazením. „Metoda je oblíbeným nástrojem zejména při zpracování rozsáhlejších kontingenčních tabulek, které obsahují mnohočetné kategorie, a kdy se grafické metody stávají ve srovnání s číselnými přehlednější." (HEBÁK, 2007)

Korespondenční analýza je spíše popisnou metodou, jejímž cílem je odhalit zákonitosti a souvislosti v datech. Nevznáší žádné hypotézy o zkoumaných datech. Často slouží jako východisko pro určení, které kategorie je vhodné sloučit v rámci zjednodušení analýzy.

Existují dva typy korespondenční analýzy. Jednoduchá korespondenční analýza a vícenásobná korespondenční analýza. Prvně jmenovaná metoda popisuje vztahy dvou kategoriálních proměnných, zatímco vícenásobná korespondenční analýza se zabývá vztahy třech a více kategoriálních proměnných. V této kapitole budou základní pojmy vysvětleny na jednoduché korespondenční analýze, v jejím závěru bude představeno rozšíření na vícenásobnou korespondenční analýzu.

5.1 Základní pojmy

5.1.1 Zátěže, profily

Východiskem korespondenční analýzy je *korespondenční matice* $\boldsymbol{P}$, která je získána jako podíl původní matice četností kontingenční tabulky $\boldsymbol{N}$ a celkové četnosti n. Jedná se vlastně o tabulku relativních četností představenou v kapitole 3.3.1.

$$\boldsymbol{P} = \boldsymbol{N}/n \tag{5.1}$$

Vydělením řádkových marginálních četností n_{i+} celkovou četností n, se získají řádkové zátěže r_i. Vektor řádkových zátěží se značí $\boldsymbol{r}$.

$$r_i = \frac{n_{i+}}{n} \tag{5.2}$$

Sloupcové zátěže c_j se získají jako podíl sloupcových marginálních četností c_j a celkové četnosti n. Výsledný vektor se značí $\boldsymbol{c}$.

$$c_j = \frac{n_{+j}}{n} \tag{5.3}$$

Při zkoumání struktury kontingenční tabulky nemá smysl porovnávat absolutní četnosti n_{ij}, protože se vztahují k různě vysokým řádkovým n_{i+} nebo sloupcovým n_{+j} marginálním součtům. Data je proto nutné transformovat na stejnou škálu. K tomuto účelu obvykle slouží řádkové $r_{j/i}$ a sloupcové profily $c_{i/j}$, které se počítají podle následujících vzorců:

$$r_{j/i} = n_{ij}/n_{i+} \tag{5.4}$$

$$c_{i/j} = n_{ij}/n_{+j} \tag{5.5}$$

Sloupcové a řádkové profily představují řádkově a sloupcově podmíněné relativní četnosti. Matice řádkových profilů je značena $\boldsymbol{R}$ a matice sloupcových profilů $\boldsymbol{C}$. Je možné je vypočítat také podle maticových vztahů (5.6) a (5.7).

$$\boldsymbol{R} = \boldsymbol{D}_{\boldsymbol{r}}^{-1}\boldsymbol{P} \tag{5.6}$$

$$\boldsymbol{C} = \boldsymbol{D}_{\boldsymbol{c}}^{-1}\boldsymbol{P}^{T} \tag{5.7}$$

kde $\boldsymbol{D}_{\boldsymbol{r}}$ je diagonální matice s prvky vektoru řádkových zátěží na diagonále a matice $\boldsymbol{D}_{\boldsymbol{c}}$ je diagonální matice s prvky vektoru sloupcových zátěží na diagonále.

Kontingenční tabulku je tak možné vyjádřit jako matici v následující podobě:

$$\begin{bmatrix} \boldsymbol{P} & \boldsymbol{r} \\ \boldsymbol{c}^{T} & 1 \end{bmatrix} \tag{5.8}$$

Nástrojem pro testování nezávislosti v kontingenčních tabulkách je chí-kvadrát test, který byl představen vztahem (4.2). Vzhledem k nově představené terminologii může být nyní tato statistika prezentována také v maticové podobě:

$$X^2 = n_{i+}(\boldsymbol{r}_i - \boldsymbol{c})^T \boldsymbol{D}_c^{-1}(\boldsymbol{r}_i - \boldsymbol{c}) \tag{5.9}$$

$$X^2 = n_{+j}\left(\boldsymbol{c}_j - \boldsymbol{r}\right)^T \boldsymbol{D}_r^{-1}(\boldsymbol{c}_j - \boldsymbol{r}) \tag{5.10}$$

Všechny varianty chí-kvadrát statistiky podávají stejné výsledky a jsou si rovnocenné. Nezávislost v kontingenční tabulce lze ověřovat i pomocí vztahů řádkových profilů a sloupcových zátěží. V případě nezávislosti jsou všechny řádkové profily shodné a navíc se shodují s vektorem sloupcových zátěží. Obdobný vztah platí pro nezávislost sloupcových profilů.

Zajímavé jsou i další vztahy mezi zátěžemi a profily. Například vektor sloupcových zátěží $\boldsymbol{c}$ lze vyjádřit jako vážený součet řádkových profilů $\boldsymbol{r}_i$ s vahami marginálních relativních četností p_{i+}.

$$\boldsymbol{c} = \sum_{i=1}^{r} p_{i+}\boldsymbol{r}_i \tag{5.11}$$

Vektor $\boldsymbol{c}$ je zároveň nazýván průměrným řádkovým profilem, tedy *centroidem* sloupcových profilů v s-rozměrném prostoru. Analogický vztah platí pro řádkové zátěže.

5.1.2 Vzdálenosti

Další oblastí, která hraje v korespondenční analýze důležitou roli, je problematika vzdáleností. K zakreslení řádkových popř. sloupcových profilů do korespondenční mapy, je potřeba znát vzájemné vzdálenosti jednotlivých profilů. Profily představují body umístěné v r-rozměrném resp. s-rozměrném prostoru. Cílem analýzy je jejich převedení nejlépe do dvourozměrného prostoru, ve kterém body odpovídají jednotlivým kategoriím.

Pro účely korespondenční analýzy se téměř výlučně používá *chí-kvadrát vzdálenost*. Jedná se o váženou euklidovskou vzdálenost. Pro vzdálenosti mezi řádkovými profily slouží jako váha prvky c_j průměrného sloupcového vektoru $\boldsymbol{c}^T$.

$$V(i,i') = \sqrt{\sum_{j=1}^{s} \frac{(r_{ij} - r_{i'j})^2}{c_j}} \tag{5.12}$$

Analogické vztahy platí pro vzdálenost sloupcových profilů, která je vyjádřena vzorcem (5.13).

$$V(j,j') = \sqrt{\sum_{i=1}^{r} \frac{(c_{ij} - c_{i'j})^2}{r_i}} \tag{5.13}$$

Z použití chí-kvadrát vzdálenosti v korespondenční analýze vyplývá mnoho jejích důležitých vlastností. Jednou z nejvýznamnějších je umožnění tvorby symetrické korespondenční mapy, která bude popsána v kapitole 5.2.

5.1.3 Inerce

Inerce (*inertia*) vyjadřuje rozptýlení bodů v mnohorozměrném prostoru, a zaujímá tak důležité místo v posuzování kvality modelu. „Termín inerce je převzatý z mechaniky, kde je definován jako součet součinu hmotnosti r a čtvercových vzdáleností (d^2) od centroidu všech částic fyzického objektu, tj. $I = rd^2$.“ (HEBÁK, 2007). Inerce tak roste se čtvercem vzdálenosti. Obdobnou funkci má i v korespondenční analýze, ve které jsou za hmotnost považovány relativní marginální četnosti p_{i+} resp. p_{+j} a za vzdálenosti chí-kvadrát vzdálenosti $V(i,i')$ resp.

$V(j,j')$ od svého centroidu. Chí-kvadrát vzdálenosti jsou už v základu vyjádřeny jako čtverec, nemusí se tedy mocnit dvěma.

Celková řádková inerce je pak definována jako vážený součet inercí všech řádků, ve kterém jsou vahami relativní marginální četnosti p_{i+}.

$$I = \sum_{i=1}^{r} p_{i+}(\boldsymbol{r}_i - \boldsymbol{c})^T \boldsymbol{D}_c^{-1} (\boldsymbol{r}_i - \boldsymbol{c}) \tag{5.14}$$

Obdobný vztah platí pro celkovou sloupcovou inerci:

$$I = \sum_{j=1}^{s} p_{+j}(\boldsymbol{c}_j - \boldsymbol{r})^T \boldsymbol{D}_r^{-1} (\boldsymbol{c}_j - \boldsymbol{r}) \tag{5.15}$$

Inerce je dána uspořádáním četností v tabulce. Podíl vysvětlené inerce sloupcovými a řádkovými profily se používá pro klasifikaci kvality modelu, přičemž vysoký podíl vysvětlené inerce svědčí pro kvalitní model.

5.2 Algoritmus korespondenční analýzy

Výpočetní algoritmy korespondenční analýzy snižují rozměr prostoru nejlépe na dvourozměrný při současném zachování co nejvíce informace z dat. Aby bylo možné zobrazit vícerozměrný prostor v dvourozměrném prostoru, je do něj vložena rovina, do které se všechny body promítají. Účelem korespondenční analýzy je najít takovou rovinu, která je v celkovém součtu nejblíže ke všem profilům.

Výpočty souřadnic bodů vycházejí z rozkladu matice normovaných reziduí $\boldsymbol{Z}$ vyjádřenou vztahem (5.16), pro jejíž prvky platí vztah (5.17).

$$\boldsymbol{Z} = \boldsymbol{D}_{\boldsymbol{r}}^{-\frac{1}{2}}(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}^T)\boldsymbol{D}_{\boldsymbol{c}}^{-\frac{1}{2}} \tag{5.16}$$

$$z_{ij} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \tag{5.17}$$

Používá se spektrální rozklad, nebo častěji jeho zobecnění, singulární rozklad vyjádřený jako:

$$\boldsymbol{Z} = \boldsymbol{U\Gamma V}^T \tag{5.18}$$

kde matice $\boldsymbol{U}$ je tvořena levými zobecněnými singulárními vektory, matice $\boldsymbol{\Gamma}$ je diagonální matice zobecněných singulárních hodnot a matice $\boldsymbol{V}$ je tvořena pravými zobecněnými singulárními vektory. Singulární rozklad lze aplikovat na jakoukoliv obdélníkovou matici, zatímco spektrální rozklad pouze na symetrickou čtvercovou matici. Nevýhodou singulárního rozkladu je jeho vyšší výpočetní náročnost.

Po realizaci singulárního rozkladu je nutné zvolit způsob normalizační metody, tedy způsob zobrazení bodů v korespondenční mapě. Jsou-li předmětem zájmu vztahy mezi řádkovými kategoriemi, je obvykle volena analýza řádkových profilů (*row principal*). Obdobně, při zkoumání vztahů mezi sloupcovými kategoriemi je volena analýza sloupcových profilů (*column principal*). Nejčastěji se však používá simultánní analýza řádkových i sloupcových profilů (*symmetrical normalization*). Volba jedné z metod nemá vliv na velikost singulárních hodnot, pouze dojde ke změně variability souřadnic, jejímž důsledkem jsou relativně odlišné výsledné korespondenční mapy.

Při analýze řádkových profilů najdeme normované souřadnice řádkových bodů (*principal coordinates*) v matici $\boldsymbol{F}$ vypočtené podle vztahu (5.19). „Souřadnice sloupcových kategorií jsou normovány tak, aby byl součet čtvercových vzdáleností od centroidu roven jedné.“ (HEBÁK, 2007) Jejich souřadnice lze vypočítat podle vztahu (5.20). Počet použitých sloupců matic normovaných souřadnic záleží na rozměru požadovaného řešení. Např. u dvourozměrného řešení matice $\boldsymbol{F}$ se použijí její první dva sloupce.

$$\boldsymbol{F} = \boldsymbol{D_r^{-1/2}U\Gamma} \tag{5.19}$$

$$\boldsymbol{Y} = \boldsymbol{D_c^{-1/2}V} \tag{5.20}$$

Obdobné vztahy platí i pro analýzu sloupcových profilů. Normované souřadnice sloupcových bodů (*principal coordinates*) se nacházejí ve sloupcích matice $\boldsymbol{G}$:

$$\boldsymbol{G} = \boldsymbol{D_c^{-1/2} V \Gamma} \tag{5.21}$$

Souřadnice řádkových bodů (*standard coordinates*) jsou ve sloupcích matice $\boldsymbol{X}$.

$$\boldsymbol{X} = \boldsymbol{D_r^{-1/2} U} \tag{5.22}$$

Pro zobrazení souřadnic řádkových i sloupcových profilů v jedné korespondenční mapě, je možné vycházet ze vzorce (5.19) pro řádkové body a vzorce (5.21) pro sloupcové body. Body řádkových a sloupcových kategorií v tomto případě náleží do odlišných prostorů, proto nemá smysl měření jejich vzájemných vzdáleností.

5.3 Korespondenční mapa

Hlavním cílem korespondenční analýzy je vytvoření korespondenční mapy, ve které jsou přehledně zobrazeny vztahy mezi kategoriemi proměnných.

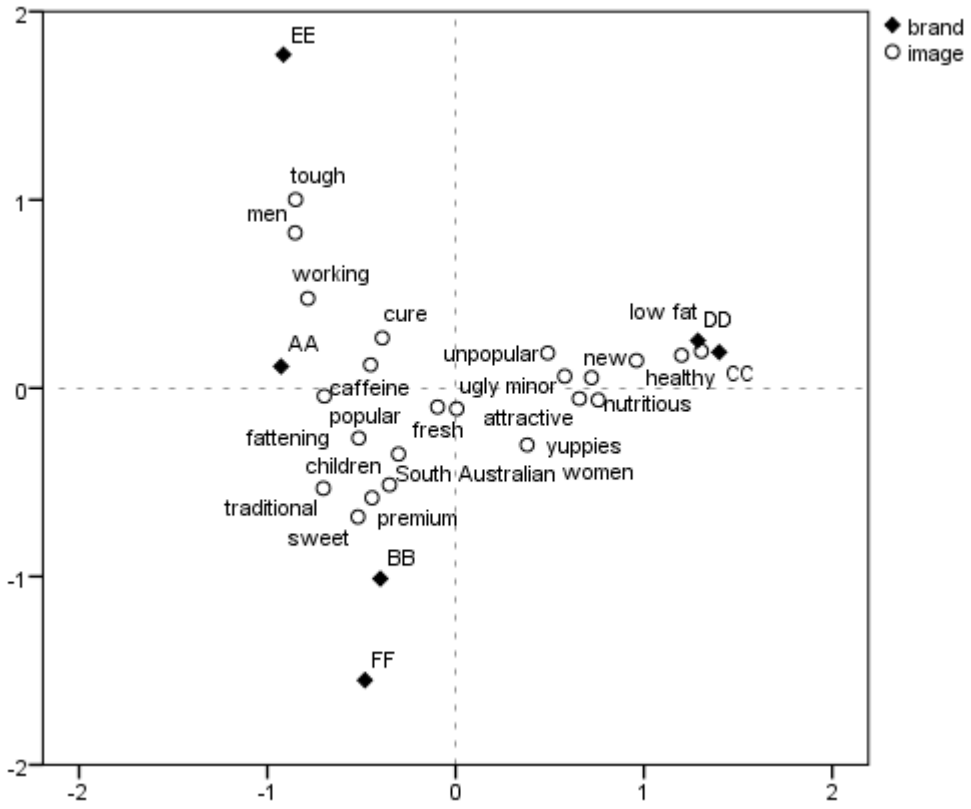
Na základě typu souřadnic vypočtených v kapitole 5.2 , mohou vzniknout tři korespondenční mapy:

- asymetrická korespondenční mapa řádkových profilů
- asymetrická korespondenční mapa sloupcových profilů
- symetrická korespondenční mapa řádkových a sloupcových profilů

Je-li zkoumána asymetrická mapa řádkových profilů, sloupcové body mají především interpretační význam vůči řádkovým bodům. Nazývají se *vrcholy* (*vertices*) nebo referenční body. Protože se nacházejí ve stejném prostoru jako řádkové body, mohou být s těmito body přímo poměřovány. Čím blíže se řádkové a sloupcové body nacházejí, tím jsou si dané kategorie podobnější. Naopak, velmi vzdálené body svědčí pro rozdílnost kategorií. Analogické vztahy platí také pro asymetrickou mapu sloupcových profilů. Nevýhodou asymetrických korespondenčních map je příliš velké „nahuštění“ bodů v jedné části korespondenční mapy, což je činí poměrně

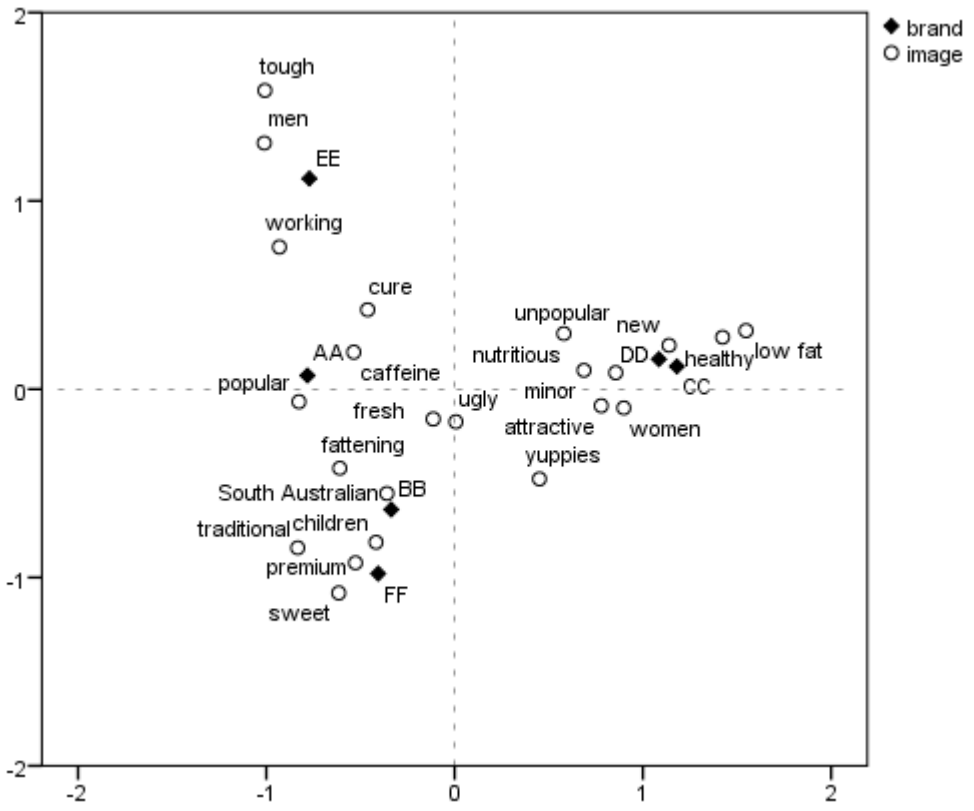
nepřehlednými. To je dobře znázorněno na grafu G1, ve kterém jsou body kategorií koncentrovány uprostřed korespondenční mapy.

G1 – Asymetrická korespondenční mapa; zdroj: (SPSS, Inc., 2012)



Z tohoto důvodu a také proto, že je často potřeba sledovat vliv řádkových i sloupcových kategorií současně, se většinou používají symetrické korespondenční mapy. Jak je možné pozorovat na grafu G2, rozmístění bodů v takové mapě je rovnoměrnější, a proto přehlednější. Body v symetrické mapě řádkových i sloupcových bodů se mohou interpretovat vzhledem k vzájemné poloze, ale také k poloze vůči hlavním osám. Měření vzdáleností mezi body řádkových a sloupcových kategorií zde však nemá smysl, protože se oba soubory bodů se nacházejí v odlišných prostorech.

G2 – Symetrická korespondenční mapa; zdroj: (SPSS, Inc., 2012)



Významnou roli při analyzování výstupů korespondenční analýzy hrají osy korespondenční mapy. Ty obvykle rozdělují soubor bodů podle určité věcné interpretace, podle které se pak pojmenovávají. Může to být např. věk, kdy budou např. mladší věkové kategorie vlevo od osy, zatímco starší budou zobrazeny napravo od ní.

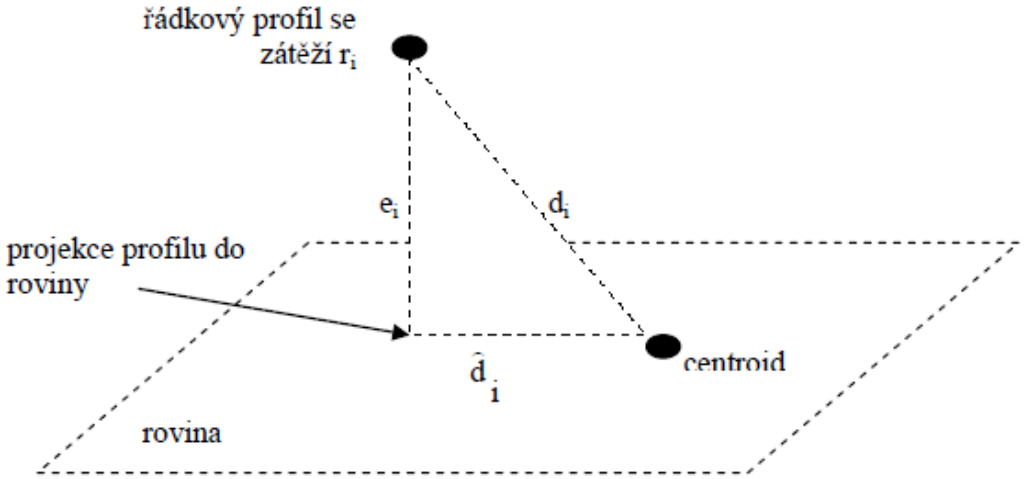
5.4 Hodnocení kvality modelu

Kvalita modelu je posuzována na základě vysvětlené inerce. Jak je vidět na grafu G3, celková inerce $\sum_i r_i d_i^2$ může být rozložena na inerci v rovině $\sum_i r_i \hat{d}_i^2$ a reziduální inerci $\sum_i r_i e_i^2$ podle Pythagorovy věty (5.23), která je detailněji popsaná v (KONRÁDOVÁ, 2009). Kvalita modelu je negativně ovlivňována pouze reziduální inercí.

$$\sum_i r_i d_i^2 = \sum_i r_i \hat{d}_i^2 + \sum_i r_i e_i^2 \qquad (5.23)$$

Reziduální inerce představuje vážený součet čtverců vzdáleností bodů od roviny a z hlediska kvality modelu je důležité ji minimalizovat. Jedná se o určitou analogii s regresní analýzou, u které se minimalizuje reziduální součet čtverců.

G3 – Promítnutí bodu do roviny; zdroj: (KONRÁDOVÁ, 2009)



Čím větší část reziduální inerce (dále jen inerce) je vysvětlena použitými dimenzemi řešení, tím je model lepší. Mezi celkovou inercí, statistikou X^2 a charakteristickými čísly symetrické matice $\boldsymbol{ZZ^T}$ nebo $\boldsymbol{Z^TZ}$ existuje jednoduchý vztah:

$$I = \frac{X^2}{n} = \sum_{i=1}^{r} \lambda_i^2 \qquad (5.24)$$

kde r odpovídá počtu singulárních hodnot a λ_i^2 jsou charakteristická čísla matice $\boldsymbol{ZZ^T}$ nebo $\boldsymbol{Z^TZ}$. Kvalita modelu se potom dá snadno určit podle vztahu (5.25), který porovnává vysvětlenou inerci v požadovaném řešení vůči celkové inerci.

$$\frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{r} \lambda_i^2} \qquad (5.25)$$

kde k je počet rozměrů požadovaného řešení a r vyjadřuje počet singulárních hodnot.

Další kritéria, která slouží k hodnocení kvality modelu, nebudou dále popisována pomocí vzorců, které jsou vypsány a vysvětleny např. v (HEBÁK, 2007), nýbrž budou představena na výstupech statistického softwaru SPSS. Nejvýznamnějšími hodnotícími kritérii kvality modelu jsou:

- příspěvky řádkových a sloupcových profilů k celkové inerci
- celková řádková (nebo sloupcová) inerce
- příspěvky řádkových (nebo sloupcových) bodů k inerci
- příspěvky os k reprodukci řádkových (nebo sloupcových) kategorií

Inerce může být sledována z pohledu řádkových nebo sloupcových bodů. Všechny níže uvedené příklady jsou představeny na řádkových bodech. Postup pro sloupcové body je analogický.

Příspěvky řádkových a sloupcových profilů k celkové inerci jsou vyjádřeny v tabulce T7 v části *Variance Accounted for Inertia*. Jde vlastně o aplikování vzorce (5.25). Samotná charakteristická čísla jsou části *Variance Accounted for Total*. Součástí této tabulky je i *Cronbachův koeficient alfa* (*Cronbach's Alpha*), který je mírou spolehlivosti (*reliability*) v datech. Používá se k měření vnitřní konzistence dat. Jeho hodnota vyjadřuje dolní hranici spolehlivosti v datovém souboru. Nabývá hodnot od nuly do jedné. Hodnoty do 0,5 se považují za nevyhovující, 0,7 a vyšší za přijatelné.

T7 – Souhrn modelu; zdroj: (SPSS, Inc., 2012)

		Variance Accounted For		
Dimension	Cronbach's Alpha	Total (Eigenvalue)	Inertia	% of Variance
1	,831	1,711	,856	85,560
2	,570	1,399	,699	69,933
Total		3,110	1,555	
Mean	,714	1,555	,777	77,747

Hodnoty inerce pro jednotlivé kategorie jsou znázorněny v tabulce T8 v části *Inertia*. Celková řádková inerce je rovna součtu inercí všech řádkových kategorií, v uvedeném příkladu tedy 0,804.

Příspěvky řádkových bodů k inerci dané dimenze jsou obsaženy pod názvem *Contribution Of Point to Inertia of Dimension*. Například kategorie *EE* způsobuje 47,7 % vysvětlené inerce druhé dimenze.

T8 – Přehled řádkových bodů; zdroj: (SPSS, Inc., 2012)

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
brand	Mass	1	2	Inertia	1	2	1	2	Total
AA	,217	-,927	,115	,127	,187	,003	,744	,004	,748
BB	,131	-,400	-1,012	,078	,021	,134	,135	,272	,407
CC	,185	1,400	,191	,193	,362	,007	,951	,006	,957
DD	,162	1,287	,253	,146	,267	,010	,928	,011	,939
EE	,152	-,915	1,771	,153	,127	,477	,420	,494	,914
FF	,153	-,482	-1,551	,107	,036	,369	,169	,550	,718
Active Total	1,000			,804	1,000	1,000			

Konečně, příspěvky os (tj. dimenzí) k reprodukci řádkových kategorií je možné nalézt pod názvem *Contribution Of Dimension to Inertia of Point*. Tyto příspěvky poskytují informaci, z kolika procent je inerce dané kategorie vysvětlena určitou dimenzí. Např. kategorie *CC* je vysvětlena první dimenzí z 95,1 % a druhou dimenzí pouze z 0,6 %. Jedná se o obdobu komunalit používaných ve faktorové analýze.

5.5 Vícenásobná korespondenční analýza

Jednoduchá korespondenční analýza řeší případy, když jsou předmětem zájmu vztahy kategorií dvou proměnných. Tato práce se však zabývá problematikou vícerozměrných kontingenčních tabulek, u kterých je počet analyzovaných proměnných mnohem vyšší. Pro takové tabulky se používá vícenásobná korespondenční analýza (*Multiple Correspondence Analysis*), která je rozšířením jednoduché korespondenční analýzy. Umožňuje společně analyzovat kategorie všech proměnných a umístit je do jedné korespondenční mapy, ve které se dají přehledně porovnávat.

Vícenásobná korespondenční analýza vychází z *matice indikátorů* **C** (*indicator matrix, design matrix*), která je představena v tabulce T9.

T9 – Matice indikátorů; zdroj: (STATSOFT, 2012)

	Přežití		Věk			Lokalita		
Pozorování	Ne	Ano	<50	50-69	69<	Tokio	Boston	Glamorgan
1	0	1	0	1	0	0	0	1
2	1	0	1	0	0	1	0	0
3	0	1	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1
…	.	.	.	.	.	.	.	.
…	.	.	.	.	.	.	.	.
…	.	.	.	.	.	.	.	.
762	1	0	0	1	0	1	0	0
763	0	1	1	0	0	0	1	0
764	0	1	0	1	0	0	0	1

Každý řádek matice indikátorů $\boldsymbol{C}$ se týká právě jednoho pozorování. U každé kategorie je zobrazena hodnota jedna nebo nula podle toho, zda sledované pozorování do této kategorie spadá nebo ne. Matice indikátorů je tak tvořena tolika sloupci, kolik je celkem kategorií pro všechny proměnné, a tolika řádky, kolik je celková četnost pozorování. Hlavní nevýhodou této matice jsou její rozměry. Už poměrně jednoduchá tabulka T9 zaujímá se svými rozměry 764 x 8 celkem 6112 polí. Proto indikátorová matice slouží především jako východisko k tvorbě *Burtovy matice* $\boldsymbol{B}$, ze které je vícenásobná korespondenční analýza většinou konstruována.

T10 – Burtova matice; zdroj: (STATSOFT, 2012)

		Přežití		Věk			Lokalita		
		Ne	Ano	<50	50-69	69<	Tokio	Boston	Glamorgan
Přežití	Ne	210	0	68	93	49	60	82	68
	Ano	0	554	212	258	84	230	171	153
Věk	<50	68	212	280	0	0	151	58	71
	50-69	93	258	0	351	0	120	122	109
	69<	49	84	0	0	133	19	73	41
Lokalita	Tokio	60	230	151	120	19	290	0	0
	Boston	82	171	58	122	73	0	253	0
	Glamorgan	68	153	71	109	41	0	0	221

Burtova matice (*Burt matrix*, *Burt table*) je oproti matici indikátorů kompaktnější, má tvar symetrické kontingenční tabulky. Vztah mezi maticí indikátorů a Burtovou maticí je vyjádřen jako:

$$\boldsymbol{B} = \boldsymbol{C}^T\boldsymbol{C} \tag{5.26}$$

Burtova matice je složena ze submatic popisujících vztahy mezi všemi kombinacemi kategorií proměnných. Tabulka T9 obsahuje tři kategoriální proměnné, proto je v Burtově matice obsaženo $3\,x\,3 = 9$ submatic. Na hlavní diagonále obsahuje tzv. *diagonální submatice*, které vznikají konfrontací dvou stejných kategoriálních proměnných. Tyto submatice mají obsazeny pouze diagonální prvky, jejichž součet vždy dává celkovou četnost n. Ostatní prvky jsou nulové.

„Vícenásobnou korespondenční analýzu lze chápat jako použití jednoduché korespondenční analýzy na Burtovu matici.“ (KONRÁDOVÁ, 2009) Obdobně je možné postupovat i z matice indikátorů $\boldsymbol{C}$, což však není příliš běžné. Výsledky obou metod jsou velice podobné. Výrazněji se liší pouze u hodnoty inerce, která je u Burtovy matice vyšší kvůli přítomnosti diagonálních submatic. Ty samy o sobě nepřinášejí žádnou informaci, ale zato významným způsobem zvyšují inerci. Obvykle je tento problém řešen výpočetními optimalizacemi, které jsou založeny na úpravě charakteristických čísel Burtovy matice.

Samotné hodnocení kvality modelu u vícenásobné korespondenční analýzy je obdobné jako v případě jednoduché korespondenční analýzy. Ukazatele kvality modelu jsou založeny na míře vysvětlené inerce a dále na mírách s ní spojených, jako jsou příspěvky jednotlivých os k reprodukci kategorií a příspěvky jednotlivých bodů k inerci. Interpretace korespondenční mapy je opět analogická s popisem korespondenční mapy jednoduché korespondenční analýzy v kapitolách 5.2 a 5.3. Jediný rozdíl je ve skutečnosti, že u vícenásobné korespondenční analýzy existuje více vztahů mezi kategoriemi jednotlivých proměnných.

Kapitola 6
Loglineární modely

Loglineární modely se používají pro modelování četností v kontingenčních tabulkách a pro určování asociace mezi kategoriálními proměnnými. Poskytují také komplexní přehled o asociacích mezi kategoriemi kategoriálních proměnných.

Rovnice loglineárních modelů vykazuje určitou podobnost s rovnicí lineární regrese. Na levé straně jsou modelované četnosti, na pravé straně je lineární kombinace parametrů v multiplikativní formě. Prvky této rovnice se obvykle transformují na přirozené logaritmy. Tak je docíleno aditivního vztahu, který je výhodnější z důvodu výpočetní náročnosti, přičemž na výsledky nemá tato úprava žádný vliv. Výsledné parametry loglineárních modelů musí být pro účely věcné interpretace odlogaritmovány.

Na rozdíl od klasické terminologie *GLM*, která rozlišuje proměnné na vysvětlující a vysvětlované, loglineární modely zacházejí se všemi proměnnými jako s vysvětlovanými. Zkoumají tak asociace mezi všemi proměnnými včetně těch, u kterých z hlediska věcné interpretace nemají žádný význam.

Rozdělení četností v kontingenční tabulce není normální. Loglineární modely používají buď multinomické, nebo Poissonovo rozdělení. Pro úsudky prováděné na základě četností v kontingenční tabulce se v praxi většinou používá aproximace těchto nespojitých rozdělením normálním. Přesné postupy s využitím nespojitých rozdělení se teprve studují.

Loglineární analýza si klade za otázku, „zdali a jak se efekt jedné kategorie proměnné liší od jiné kategorie té samé proměnné.“ (PETŘÍKOVÁ, 2009) Snaží se tak zjistit, zda existuje mezi dvěma kategoriemi dané proměnné významný rozdíl, a pokud ano, kolikrát je vyšší šance, že nastane určitá kategorie oproti druhé kategorii.

Šance (*odds*) Ω je v problematice loglineárních modelů důležitou statistikou. Šance, že daná veličina nabude v i-tém řádku hodnoty j a ne j', je vyjádřena jako poměr dvou podmíněných relativních četností:

$$\Omega = \frac{p_{j/i}}{p_{j'/i}} = \frac{p_{ij}}{p_{ij'}} = \frac{n_{ij}}{n_{ij'}} \tag{6.1}$$

Může tak být např. vyjádřena šance jedna ku deseti, že se muž stane alkoholikem.

Jsou-li srovnávány dvě šance v i-tém a i'-tém řádku, výsledkem je poměr šancí (*odds ratio*), který je definován vztahem (6.2). Nabývá nezáporných hodnot. Je-li roven jedné, vyjadřuje nezávislost sledovaných proměnných.

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{ij}/\pi_{ij'}}{\pi_{i'j}/\pi_{i'j'}} \cong \frac{n_{ij}/n_{ij'}}{n_{i'j}/n_{i'j'}} \tag{6.2}$$

Čím více se hodnoty vzdalují od jedné (ať už jakýmkoliv směrem), tím větší je asociace mezi proměnnými. Pokud se aspoň jedna pravděpodobnost (nebo četnost) ve vztahu (6.2) rovná nule, pak poměr šancí nabývá hodnot nula resp. nekonečno v závislosti na tom, jestli se nulová pravděpodobnost vyskytuje v čitateli nebo jmenovateli. Pomocí poměru šancí je možné zkoumat, kolikrát je vyšší šance, že se stane alkoholikem muž, než že se jím stane žena. Obdobně je možné postupovat pro případ vícerozměrných interakcí. Pak lze sledovat například šance, že se stane alkoholikem vysokoškolsky vzdělaný muž oproti středoškolsky vzdělané ženě.

6.1 Model nezávislosti

Četnosti v kontingenční tabulce mají za podmínky nezávislosti Poissonovo rozdělení se strukturou vyjádřenou následujícím vztahem:

$$\mu_{ij} = \mu\alpha_i\beta_j \tag{6.3}$$

μ značí průměrnou četnost a parametry α_i a β_j vlivy řádkových a sloupcových kategorií, přičemž $\sum_i \alpha_i = \sum_j \beta_j = 1$.

Vztah (6.3) představuje model pro nezávislost v dvourozměrné kontingenční tabulce v multiplikativním vztahu. Z hlediska výpočetní náročnosti je výhodnější používat aditivní vztah, jehož je docíleno transformací pomocí přirozených logaritmů:

$$\ln \mu_{ij} = \lambda + \alpha_i^* + \beta_j^* \tag{6.4}$$

kde $\lambda = \ln \mu$, $\alpha_i^* = \ln \alpha_i$ a $\beta_j^* = \ln \beta_j$. Pro účely loglineárních modelů se vztah (6.4) zjednodušuje na:

$$\ln \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \tag{6.5}$$

kde λ_i^X vyjadřuje efekt klasifikace řádkových proměnných a stará se o to, že součet všech očekávaných četností se rovná četnostem výběrovým. $\sum_j \mu_{ij} = \mu_{i+} = n_{i+}$. Obdobně lze interpretovat efekt sloupcových proměnných λ_j^Y. Model nezávislosti pro trojrozměrnou tabulku by se vytvořil přidáním efektu třetí proměnné λ_k^Z do vztahu (6.5). Způsob zápisu loglineárních modelů připomíná zápis faktorů v analýze rozptylu.

6.2 Saturovaný model

Jsou-li proměnné X a Y závislé, loglineární model (6.5) obsahuje navíc ještě člen λ_{ij}^{XY} vyjadřující jejich interakci. Tento člen obsahuje všechny odchylky od platnosti nezávislosti, tedy určuje hodnotu poměru šancí, který je v modelu nezávislosti vždy roven jedné. Z výše uvedeného plyne, že model nezávislosti je u dvourozměrné tabulky zvláštním případem saturovaného modelu za podmínky, že $\lambda_{ij}^{XY} = 0$.

$$\ln \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{6.6}$$

V saturovaném modelu je právě tolik parametrů, kolik je ve sledované kontingenční tabulce buněk s četnostmi. Z toho vyplývají dva důležité závěry. Saturovaný model se vždy perfektně shoduje s napozorovanými daty. Všechny výběrové četnosti se tak shodují s očekávanými. Zadruhé, počet stupňů volnosti je u saturovaného modelu roven nule. K namodelování četností kontingenční tabulky bylo použito maximální množství parametrů.

V praxi se, pokud je to možné, dává přednost nesaturovaným modelům. Ať už z důvodu určitého „zobecnění“ sledovaných zákonitostí nebo kvůli jednoduššímu modelování dat.

Loglineární modely se zapisují do závorek, ve kterých jsou zobrazeny jednotlivé hlavní efekty a interakce, které obsahují. Model pak může vypadat např. takto:

$$(A, B, C, AB, BC, AC, ABC) \qquad (6.7)$$

Jedná se o zápis saturovaného modelu, který obsahuje tři proměnné (A, B a C) a všechny jejich možné interakce. Jednotlivá písmena značí efekty kategorií jednotlivých proměnných, dvojice či trojice písmen značí efekty jejich interakcí. Interakce proměnných AB , BC a AC odpovídají marginálním kontingenčním tabulkám, které jsou konstruovány z trojrozměrné kontingenční tabulky. Jejich modelové četnosti se od pozorovaných liší (kromě saturovaného modelu), ale marginální součty se shodují. Celková četnost n v modelové i pozorované marginální kontingenční tabulce je tedy shodná. Toho se využívá při ověřování shody modelových a pozorovaných dat, k němuž se používají testy založené na chí-kvadrát statistice X^2 a G^2 představené v kapitole 4.

6.3 Nesaturovaný model

Cílem statistického modelování je najít kompromis mezi přesností a úsporností modelu, aby dobře vynikly důležité zákonitosti v datech a zároveň byl model výpočetně nenáročný a snadno interpretovatelný. Cílem je najít takový nesaturovaný model, který má menší počet parametrů než model saturovaný, ale zároveň zachovává původní strukturu dat a vztahy mezi proměnnými.

Nesaturované modely se rozlišují na *hierarchické* a *nehierarchické*. Hierarchické modely jsou definovány tak, že obsahují-li interakci vyššího řádu, pak obsahují všechny interakce nižších řádů. Lze je zapsat ve zjednodušené formě, kdy zapíšeme pouze nejvyšší interakci obsaženou v modelu. Například hierarchický saturovaný model ze vztahu (6.7) lze zapsat následovně:

$$(ABC) \qquad (6.8)$$

Naopak, nehierarchické modely obsahují interakce vyšších řádů, aniž by obsahovaly všechny interakce řádů nižších. Tyto typy modelu se často nevyskytují, protože „není vždy snadné odhadnout jejich modelové četnosti, a jednak proto, že jsou obtížně interpretovatelné.“ (PETŘÍKOVÁ, 2009) Kvůli těmto nepříznivým vlastnostem se ve většině případů používají modely hierarchické. Ty navíc umožňují výběr optimálního modelu pomocí výpočetních algoritmů *Backward elimination* nebo *Forward selection*, které se uplatňují především u vícerozměrných tabulek s mnoha možnými interakcemi.

Metoda *Backward elimination* standardně začíná saturovaným modelem a testuje jeho platnost tak, že ubere interakce nejvyššího řádu. Změna ve velikosti statistiky G^2 je pak podrobena samotnému testu. Jako mezní je většinou považována p-hodnota rovna 0,05. Jeli skutečná p-hodnota vyšší, změna G^2 není významně odlišná od nuly, a zkoumanou interakci tak můžeme vypustit. V dalších krocích se obdobně testují interakce nižších řádů. Interakce, které mají p-hodnotu nižší než 0,05, jsou v modelu ponechány, interakce s p-hodnotou vyšší jsou odstraněny. Tímto způsobem se pokračuje do doby, než zůstanou pouze interakce s p-hodnotami nižšími než 0,05.

Metoda *Forward selection* je opakem *Backward elimination*. Začíná většinou modelem nezávislosti a postupně přidává stále složitější interakce, u kterých testuje, zda je přírůstek G^2 významně odlišný od nuly. Pokud tomu tak není, je dosavadní model považován za optimální.

6.4 Dummy a effect kódování v loglineárních modelech

Loglineární modely se zapisují pomocí indikátorových proměnných typu *effect* nebo *dummy*. Výběr určitého typu kódování ovlivňuje způsob interpretace parametrů modelu. Častěji se v praxi používá *effect* kódování, *dummy* kódování má zase výhodné vlastnosti ohledně věcné interpretace parametrů. V této kapitole budou popsány oba způsoby podrobněji.

6.4.1 Dummy kódování

K identifikaci parametrů se v případě *dummy* proměnných používá kombinace binárních proměnných obsahujících nuly a jedničky. Pro vyjádření k kategorií postačí pouze $(k-1)$ *dummy* proměnných, které poskytují úplnou informaci o tom, jaké pozorování patří do jaké skupiny. V případě použití k *dummy* proměnných, by byly jednotlivé řady lineárně závislé a parametry modelu by nebylo možné odhadnout.

Dummy proměnné slouží k vyjádření informace, jaké efekty se v daném řádku modelu používají. Jako výchozí je zvolen například saturovaný model (6.6) představený v kapitole 6.2.

Loglineární model pro řádkovou proměnnou X se čtyřmi kategoriemi je definován jako model se třemi indikátorovými proměnnými: D_1, D_2 a D_3:

$$\ln \mu_{1j} = \lambda + D_1\lambda_1^X + D_2\lambda_2^X + D_3\lambda_3^X + \lambda_j^Y + \lambda_{1j}^{XY} \tag{6.9}$$

Pomocí jedniček a nul je možné vytvořit čtyři odlišné kombinace:

$$\begin{aligned}
\ln \mu_{1j} &= \lambda + 1\cdot\lambda_1^X + 0\cdot\lambda_2^X + 0\cdot\lambda_3^X + \lambda_j^Y + \lambda_{1j}^{XY} \\
\ln \mu_{2j} &= \lambda + 0\cdot\lambda_1^X + 1\cdot\lambda_2^X + 0\cdot\lambda_3^X + \lambda_j^Y + \lambda_{2j}^{XY} \\
\ln \mu_{3j} &= \lambda + 0\cdot\lambda_1^X + 0\cdot\lambda_2^X + 1\cdot\lambda_3^X + \lambda_j^Y + \lambda_{3j}^{XY} \\
\ln \mu_{4j} &= \lambda + 0\cdot\lambda_1^X + 0\cdot\lambda_2^X + 0\cdot\lambda_3^X + \lambda_j^Y + \lambda_{4j}^{XY}
\end{aligned} \tag{6.10}$$

V každém řádku modelu je zobrazena jednička pouze u členu, který se dané proměnné týká. Poslední kategorie má u všech členů nuly. Tato kategorie se nazývá referenční. V modelu je nahrazena konstantou, se kterou jsou konfrontovány efekty ostatních kategorií. Referenční může být libovolná kategorie. Při její změně se efekty modelu sice změní, ale vztahy a závislosti mezi proměnnými a kategoriemi zůstanou beze změny.

Výhodou *dummy* kódování je, že umožňuje vyjádření parametrů loglineárního modelu jako poměrů šancí mezi stanovenou a referenční kategorií. Lze tak například zkoumat, kolikrát je vyšší šance, že si rodina pořídí automobil ve druhém a ne ve čtvrtém čtvrtletí.

6.4.2 Effect kódování

Druhý způsob, který se používá k zápisu loglineárních modelů, je kódování pomocí indikátorových proměnných typu *effect*, které značíme E_1, E_2 a E_3. Parametry takového modelu představují odchylky efektů kategorií od celkového průměrného efektu. Jednoznačného přiřazení parametrů je docíleno pomocí hodnot jedna, nula a mínus jedna. Zápis shodného modelu (6.6) při použití *effect* kódování vypadá takto:

$$\begin{aligned}
\ln \mu_{1j} &= \lambda + 1 \cdot \lambda_1^X + 0 \cdot \lambda_2^X + 0 \cdot \lambda_3^X + \lambda_j^Y + \lambda_{1j}^{XY} \\
\ln \mu_{2j} &= \lambda + 0 \cdot \lambda_1^X + 1 \cdot \lambda_2^X + 0 \cdot \lambda_3^X + \lambda_j^Y + \lambda_{2j}^{XY} \\
\ln \mu_{3j} &= \lambda + 0 \cdot \lambda_1^X + 0 \cdot \lambda_2^X + 1 \cdot \lambda_3^X + \lambda_j^Y + \lambda_{3j}^{XY} \\
\ln \mu_{4j} &= \lambda + (-1) \cdot \lambda_1^X + (-1) \cdot \lambda_2^X + (-1) \cdot \lambda_3^X + \lambda_j^Y + \lambda_{4j}^{XY}
\end{aligned} \tag{6.11}$$

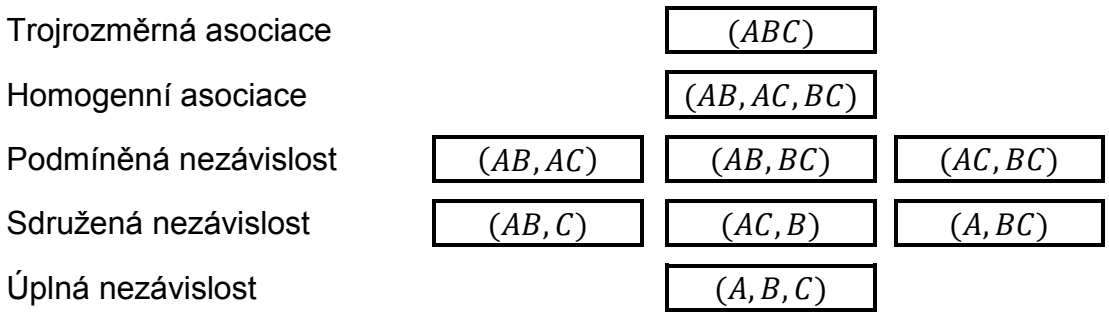
Hodnoty mínus jedna jsou použity pouze u poslední (tj. referenční) kategorie. Tam slouží k vyrovnání efektů předchozích kategorií. To znamená, že součet všech indikátorových proměnných v daném sloupci je roven nule.

Pro správnou interpretaci parametrů je nezbytné mít představu o použitých indikátorech. Např. v softwaru SPSS existují dvě procedury loglineárních modelů, přičemž každá je založena na jiných indikátorech.

6.5 Loglineární modely pro vícerozměrné tabulky

Teorie loglineárních modelů byla primárně vytvořena pro analýzu vícerozměrných tabulek, protože pro potřeby identifikování dvourozměrných asociací postačovaly míry, jako například *Pearsonův koeficient kontingence* a příslušné testy. Loglineární modely pro kontingenční tabulky o třech a více rozměrech vznikají rozšířením modelů pro dvourozměrné tabulky. Zatímco u dvourozměrných tabulek existuje pouze závislost, nebo nezávislost, ve vícerozměrných tabulkách se vyskytují různé typy asociačních vztahů. Základní typy modelů, které se mohou vyskytovat v trojrozměrné kontingenční tabulce, jsou přestaveny na schématu G4, ve kterém jsou seřazeny od nejkomplexnějšího k nejjednoduššímu.

G4 – Typy vztahů v trojrozměrné kontingenční tabulce; zdroj: (AGRESTI, 2002)



Vztahy popsané v této kapitole platí i pro vyšší než trojrozměrné tabulky. S růstem počtu rozměrů roste rychle i počet možných parametrů, a to podle vztahu 2^n. To znamená, že saturovaný model pro trojrozměrnou tabulku obsahuje 8 členů, pro čtyřrozměrnou 16, pětirozměrnou 32, atd. Při konstrukci modelu je vhodné používat co nejméně faktorové členy. Je-li v modelu použitý alespoň jeden člen vyššího řádu, není možné interpretovat členy nižších řádů samostatně, ale vždy s přihlédnutím k tomuto členu. Věcná interpretovatelnost výsledků tak s rostoucím počtem faktorů výrazně klesá.

6.6 Výběr optimálního modelu

Doposud byla představena především struktura loglineárních modelů a typy vztahů, které je modelují. Často dochází k situaci, kdy je nutné vybrat jeden z více odpovídajících modelů. Jak vybrat ten nejlepší? Běžně se používají dva způsoby volby optimálního modelu:

- pomocí statistik X^2 a G^2
- pomocí testových kritérií *AIC* a *BIC*

První možnost vychází z porovnávání hodnot chí-kvadrát statistik vzhledem k počtu stupňů volnosti. Hledá se takový model, který při co nejvyšším počtu stupňů volnosti vysvětlí co nejvíce inerce. Jedná se o zjevně si odporující požadavky, výsledkem je vždy určitý kompromis. Často rozhodování závisí na zkušenostech řešitele. Dále je nutné brát zřetel na rozměr interakcí v modelu. Při rozhodování ze dvou modelů s přibližně stejným součtem chí-kvadrát čtverců a podobnými stupni volnosti, je vhodné zvolit model s jednoduššími interakcemi, který má jednodušší věcnou interpretaci.

Na principu porovnávání chí-kvadrát statistik jsou založeny výše popsané metody *Backward elimination* a *Forward selection*. Ty se na základě automatických algoritmů snaží docílit co nejjednoduššího modelu, který se zároveň dostatečně shoduje s daty.

V posledních letech zažívá rozvoj porovnávání modelů pomocí testových kritérií *AIC* (*Akaikeho informační kritérium*) nebo *BIC* (*Bayesovské informační kritérium*). Jejich filozofie vychází z určení míry informace, kterou model o zkoumané realitě poskytuje. Vypočítají se podle vztahů:

$$AIC = G^2 - 2df \tag{6.12}$$

$$BIC = G^2 - df(\ln n) \tag{6.13}$$

kde n značí celkovou četnost a df je počet stupňů volnosti. Čím menší je hodnota testového kritéria, tím je model lepší. Vyjde-li u testového kritéria kladná hodnota, model neposkytuje dobrou shodu s daty. V takovém případě se volí saturovaný model.

V praxi se často používá určitý kompromis obou výše uvedených přístupů. U vzorků s malým počtem pozorování se spíše volí metody založené na porovnávání významnosti modelu pomocí testování chí-kvadrát statistik. U modelů s velkým počtem pozorování se dá téměř každá změna chí-kvadrát statistiky považovat za statisticky významnou, proto se u nich dává přednost testovým kritériím *AIC* nebo *BIC*.

V případech, ve kterých nevychází žádný nesaturovaný model významný, je vhodné prozkoumat rezidua těchto modelů. Nevýznamnost modelu mohou často způsobit extrémní četnosti pouze v několika buňkách kontingenční tabulky. Tyto četnosti mohou být odhaleny pomocí analýzy reziduí představené v kapitole 4.3. Případné pokračování v analýze poté záleží na zkušenostech výzkumníka.

Kapitola 7
Analýza názorů na podnikatelské prostředí v ČR

7.1 Použitá data

Jak už bylo zmíněno v úvodu, jedním z cílů této práce je analyzovat názory veřejnosti na podnikatelské prostředí v ČR a porovnat postupy a výsledky dvou odlišných metod, korespondenční analýzy a loglineárních modelů. K tomuto účelu jsem použil data od agentury GfK Czech s.r.o., která pocházejí z výzkumu zabývajícím se znalostí soutěže Podnikatel roku.

Data byla založena na kvótním výběru, který požadoval 25% zastoupení podnikatelů a 75% zastoupení populace. Celkem bylo pořízeno 1349 dotazníků.

Pro analýzu názoru veřejnosti na podnikatelské prostředí v ČR jsem vybral dvě otázky:

- Je snadné začít v České republice podnikat? (*Q6*)
- Vláda nastavuje všem stejné podmínky pro podnikání v ČR (*Q7*)

Zatímco otázka Q6 pokrývá oblast začátků podnikání, otázka *Q7* se věnuje jak začátkům, tak průběhu podnikání. Na prvně jmenovanou otázku odpovídali respondenti na pětibodové škále, u otázky *Q7* na škále jedenáctibodové. Proměnné tak nabývají velkého počtu kategorií, který přímo souvisí s vysokým počtem parametrů modelů, a tedy s jejich nepřehledností. Aby byla míra složitosti výsledných modelů udržena na přijatelné úrovni, rozhodl jsem se sloučit kategorie vysvětlovaných i vysvětlujících proměnných do menšího počtu kategorií. Také z hlediska interpretace výsledků je výhodné používat menší počet více diferenciovaných kategorií.

Vysvětlované proměnné jsou překódované v tabulkách T11 a T12.

T11 – Překódování proměnné Q7 (Vláda nastavuje všem stejné podmínky pro podnikání v ČR)

pův. hodnota	interpretace	nová hodnota	interpretace
5	Nesnadné	1	Nesouhlas
4	Spíše nesnadné		
3	Ani snadné ani nesnadné	2	Nevyhraněn
2	Spíše snadné	3	Souhlas
1	Snadné		

T12 – Překódování proměnné Q6 (Je snadné začít v České republice podnikat?)

pův. hodnota	nová hodnota	interpretace
0 až 3	1	Nesouhlas
4 až 6	2	Nevyhraněn
7 až 10	3	Souhlas

Zajímá mě, jak budou ovlivněny odpovědi na otázky *Q6* a *Q7* v závislosti na vysvětlujících proměnných:

- Ekonomická aktivita (*S6*)
- Politická orientace (*Q9*)
- Pohlaví (*S1*)

Obdobně jako u vysvětlovaných proměnných, muselo dojít k upravení škály z výše uvedených důvodů i u vysvětlujících proměnných. Ty jsou překódovány v tabulkách T13 a T14. Proměnná *Ekonomická aktivita* byla upravena následovně:

T13 – Překódování proměnné S6 (Ekonomická aktivita)

pův. hodnota	interpretace	nová hodnota	interpretace
1	podnikatel bez zaměstnanců	1	Podnikatel
2	podnikatel se zaměstnanci		
3	zaměstnanec - řadový pracovník	2	Zaměstnaný
4	zaměstnanec - vedoucí pracovník		
5	nezaměstnaný	3	Nepracující
6	nepracující důchodce		
7	v domácnosti, rodičovská		
8	student, žák, učeň		
9	jiné		

V tabulce T13 došlo ke sloučení podnikatelů bez zaměstnanců s podnikateli se zaměstnanci do kategorie *Podnikatel*, dále k sloučení řadových a vedoucích pracovníků do kategorie *Zaměstnaný*. Největší kompromisy bylo potřeba udělat v poslední skupině, která sdružuje nezaměstnané, nepracující důchodce, osoby na mateřské a rodičovské dovolené a studenty. Tyto skupiny byly sloučeny do jediné nadskupiny nazvané *Nepracující.*

Změny proměnné *Politická orientace* jsou vyjádřeny v tabulce T14. Respondenti, kteří nedokázali odpovědět, kterou stranu by právě volili, nebyli do nové proměnné zařazeni.

T14 – Překódování proměnné Q9 (Politická orientace)

pův. hodnota	interpretace	nová hodnota	interpretace
1	ČSSD	1	Levice
2	KSČM		
5	Strana práv občanů - Zemanovci		
3	KDU-ČSL	2	Střed
6	Strana zelených		
9	Věci veřejné		
10	Jiná strana, nezávislí		
4	ODS	3	Pravice
7	Suverenita		
8	TOP 09		
11	Neví	x	

Jelikož otázka *Politická orientace* v datech nebyla, musela být překódována z otázky: *Preferovaná strana v současnosti*. Samotné rozdělení stran na levici, pravici a střed je velmi zjednodušující, ale jedná se pravděpodobně o jediné řešení, které spojuje jednoduchou datovou strukturu s všeobecným vnímáním politické orientace veřejností. Respondenty, kteří nevěděli, jakou politickou stranu by právě zvolili, jsem do nové proměnné nezařadil.

Poslední typ proměnných jsou tzv. *doplňkové proměnné* (*supplementary variables*). Tyto proměnné slouží v analýze pouze k porovnávání s ostatními proměnnými. Samy však součástí výpočetní části analýzy nejsou. V praktické části hodlám použít pouze jednu doplňkovou proměnnou, a to proměnnou *Věk* (*S2b*). Také u proměnné *Věk* došlo k redukci počtu kategorií, viz tabulka T15.

T15 – Překódování proměnné S2b (Věk)

pův. hodnota	interpretace	nová hodnota	interpretace
1	18-25	1	18-25
2	26-35	2	26-55
3	36-45		
4	46-55		
5	56-65	3	56-65

7.2 Popis proměnných

Před samotnými analýzami je vhodné se podívat na strukturu dat podle jednotlivých proměnných. Základní statistiky jsou představeny v tabulce T16. Všechny proměnné obsahují 1349 pozorování kromě proměnné *Q9* (*Politická orientace*), která jich obsahuje 1120. Do této proměnné nebylo zařazeno 229 respondentů, u kterých nemohla být identifikována politická orientace. Pro analýzy budu v souladu s metodou *listwise deletion*, která vyřadí každý záznam s aspoň jednou chybějící hodnotou, používat pouze 1120 pozorování u všech proměnných.

T16 – Popisná statistika

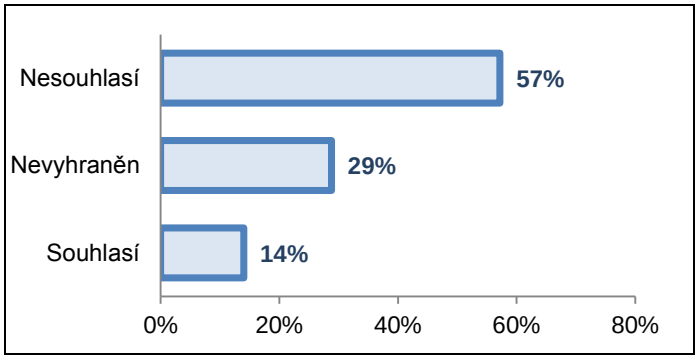
	Q7	Q6	S1	S6	Q9	S2b
N	1349	1349	1349	1349	1120	1349
Modus	1	1	1	2	3	2
Medián	1	1	x	x	x	2
Nomvar	x	x	0,997	0,919	0,978	x
Dorvar	0,731	0,811	x	x	x	0,560
Entropie[1]	0,712	0,746	0,998	0,925	0,979	0,341

Proměnná *Q7* je ordinální, obsahuje 1349 pozorování. Její první kategorie je modální i mediánová zároveň, což vyjadřuje, že nejvíce a zároveň více než 50 % respondentů nesouhlasí s výrokem, že vláda nastavuje všem stejné podmínky pro podnikání. To potvrzují i data v tabulce T16. Všechny míry variability nabývají maximální hodnoty, jsou-li všechny kategorie shodně zastoupeny. Míra ordinální variability (*dorvar*) a entropie podávají velmi obdobné výsledky (0,731 a 0,712), které svědčí o poměrně vysoké variabilitě kategorií této proměnné. Četnosti proměnné *Q7* jsou vyjádřeny v následující tabulce:

[1] u ordinálních proměnných počítána podle upraveného vzorce

T17 – Četnosti proměnné Q7

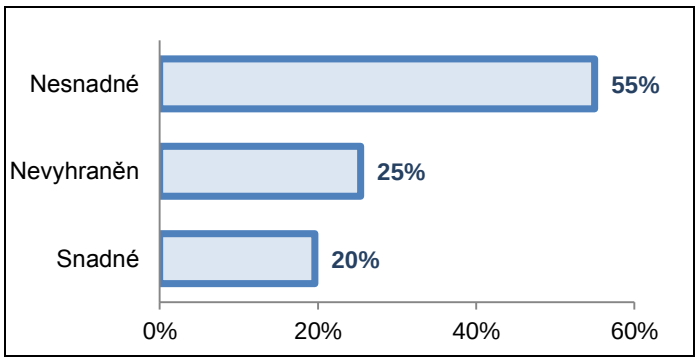
	Počet	v %
Nesouhlasí	771	57,2
Nevyhraněn	389	28,8
Souhlasí	189	14,0
Celkem	1349	100,0



Ordinální proměnná *Q6* zjišťuje, zda je v České republice snadné podnikat. Všechny statistiky i četnosti jednotlivých kategorií jsou velmi podobné proměnné *Q7*. Také u ní existuje výrazná převaha negativních odpovědí. Dobře je to možné posoudit srovnáním tabulek T17 a T18.

T18 – Četnosti proměnné Q6

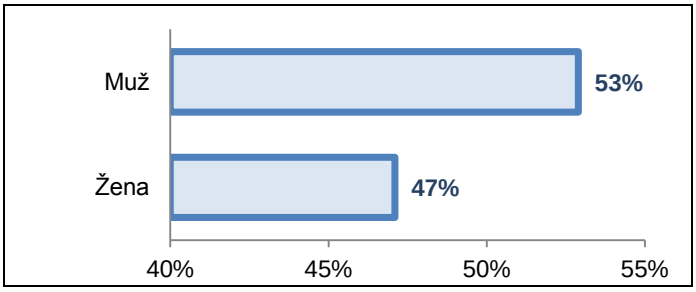
	Počet	v %
Nesnadné	742	55,0
Nevyhraněn	342	25,4
Snadné	265	19,6
Celkem	1349	100,0



Proměnná *S1* je nominální proměnná vyjadřující pohlaví respondenta. Z hlediska měr polohy má u nominálních proměnných význam sledovat pouze modus a nikoli medián. Více bylo ve výběru zastoupeno mužů, konkrétně 53 %. Míra nominální variability (*nomvar*) činí 0,997, což rovněž svědčí pro vyrovnanost kategorií mužů a žen.

T19 – Četnosti proměnné S1

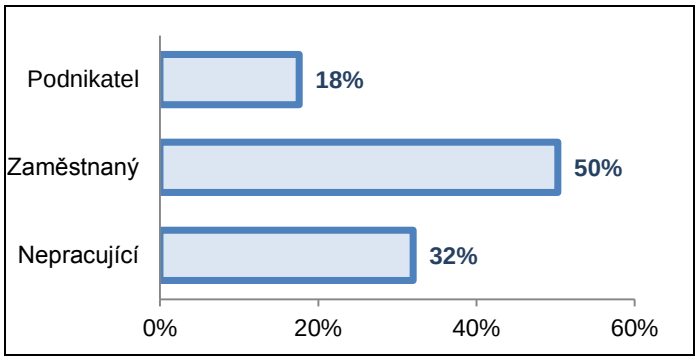
	Počet	v %
Muž	714	52,9
Žena	635	47,1
Celkem	1349	100,0



Proměnná *S6* představuje ekonomickou aktivitu respondenta. Nejvíce lidí, konkrétně 50 % sebe samé označilo za zaměstnané. Variabilita této proměnné je vysoká, je rovna hodnotě 0,919.

T20 – Četnosti proměnné S6

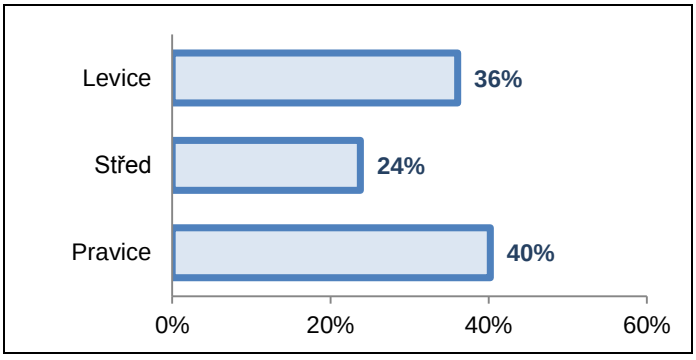
	Počet	v %
Podnikatel	238	17,6
Zaměstnaný	679	50,3
Nepracující	432	32,0
Celkem	1349	100,0



Rozložením voličů podle politické orientace se zabývá proměnná *Q9*. Zastoupení pravicově a levicově orientovaných voličů je v souboru poměrně rovnoměrné. To potvrzuje i míra variability této proměnné o velikosti 0,978. Příznivců levice je 36 %, příznivců pravice 40 % a zbytek tvoří voliči středových stran, tedy politicky nevyhranění voliči.

T21 – Četnosti proměnné Q9

	Počet	v %
Levice	404	36,1
Střed	266	23,8
Pravice	450	40,2
Celkem	1120	100,0

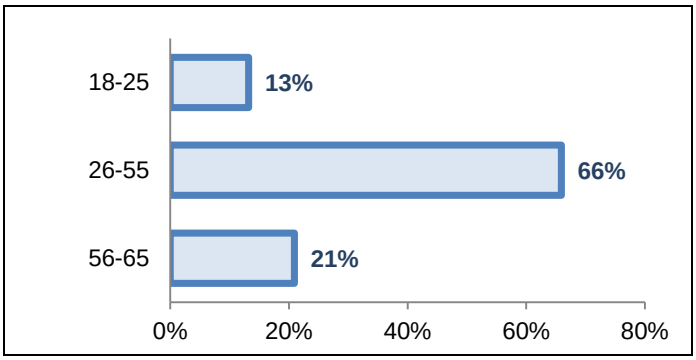


Doplňková proměnná *S2b* je tvořena věkovými intervaly rozdělujícími lidský život na předproduktivní, produktivní a poproduktivní, jsou značně nevyvážené. To je jeden z důvodu, proč jsem se rozhodl nezařadit tuto proměnnou do výpočtů analýzy. Dalším důvodem byl fakt, že by přidání další proměnné zvýšilo počet vztahů v modelu, a tím zkomplikovalo jeho interpretaci. Ve věkové kategorii 18 – 25 let odpovídalo 13 % respondentů. Největší část tvořili lidé v produktivním věku 26 – 55, celkem 66 %. Tento věkový interval je zároveň i mediánovou kategorií. Respondentů starších 56 let bylo 21 %. Míra ordinální variability je u této proměnné nejnižší ze zde sledovaných

proměnných, a to 0,560. Pokud bychom použili jako statistiku entropii, bylo by to dokonce jen 0,341. Tato nízká hodnota je způsobena dominancí prostředního věkového intervalu.

T22 – Četnosti proměnné S2b

	Počet	v %
18-25	178	13,2
26-55	889	65,9
56-65	282	20,9
Celkem	1349	100,0



Data budou analyzována v softwaru SPSS. Všechny použité úpravy dat a analýzy jsou obsaženy ve formě syntaxe v příloze 1.

Kapitola 8
Aplikace korespondenční analýzy

8.1 Nastavení korespondenční analýzy

Vícenásobná korespondenční analýzu umožňuje získat představu o struktuře zkoumaných dat a vztahy mezi proměnnými pomocí přehledné korespondenční mapy. Na základě jejích výsledků budu pokračovat v analýze kategoriálních dat pomocí loglineárních modelů.

K samotné analýze jsem zvolil software SPSS, u kterého lze tuto metodu najít v menu pod položkou *Optimal scaling*. Pod tímto názvem se skrývají hned tři metody, a sice:

- *Vícenásobná korespondenční analýza* (*Multiple Correspondence Analysis*)
- *Metoda kategoriálních hlavních komponent* (*Categorical Principal Components*)
- *Nelineární kanonická korelace* (*Nonlinear Canonical Correlation*)

Konkrétní výběr metody závisí na struktuře analyzovaných dat. Vícenásobná korespondenční analýza vyžaduje, aby všechny analyzované proměnné vycházely z jedné sady dat a mohly být považovány za nominální. Zkoumaná data tyto požadavky splňují, a proto je vícenásobná korespondenční analýza vhodnou volbou.

Nastavení vícenásobné korespondenční analýzy v SPSS je poměrně jednoduché. Do pole *Analysis Variables* vložím proměnné, u nichž mě zajímá, jak spolu korespondují, tedy proměnné *S6*, *Q6*, *S1*, *Q9* a *Q7*. Proměnnou *S2b*, u které nechci, aby byla součástí samotné analýzy, ale potřebuji, aby byla zahrnuta ve výsledných výstupech, umístím do pole *Supplementary Variables*. Nakonec je nutné zvolit rozměr výsledného řešení. Zvolím dvourozměrné řešení, které ve většině případů splňuje požadavky na jednoduchou a přitom snadnou interpretaci výsledků.

Dále je nutné nadefinovat proměnné, které budou součástí grafického řešení korespondenční analýzy, tedy korespondenční mapy. Vybral jsem všechny analyzované proměnné včetně doplňkové proměnné *S2b*.

SPSS nabízí tři způsoby, jak zacházet s chybějícími pozorováními:

- vyloučení chybějících hodnot z analýzy (*passive treatment*)
- dopočítání chybějících hodnot (*active treatment*)
- vyloučení pozorování s chybějícími hodnotami (*listwise deletion*)

Jak jsem již zmínil v kapitole 7.2, použiji metodu *listwise deletion*, při níž se do analýzy nezapočítává celý řádek, ve kterém se vyskytlo aspoň jedno chybějící pozorování. Budou tak použity pouze kompletní dotazníky a data tak nebudou žádným způsobem zkreslena. Po vyřazení 229 pozorování s chybějícími hodnotami, zůstane celková četnost $n = 1120$ dostatečná pro požadované analýzy.

8.2 Výstupy korespondenční analýzy

Po zadání nastavení parametrů analýzy uvedených v předchozí kapitole, byla provedena korespondenční analýza s těmito výstupy:

T23 – Souhrn modelu

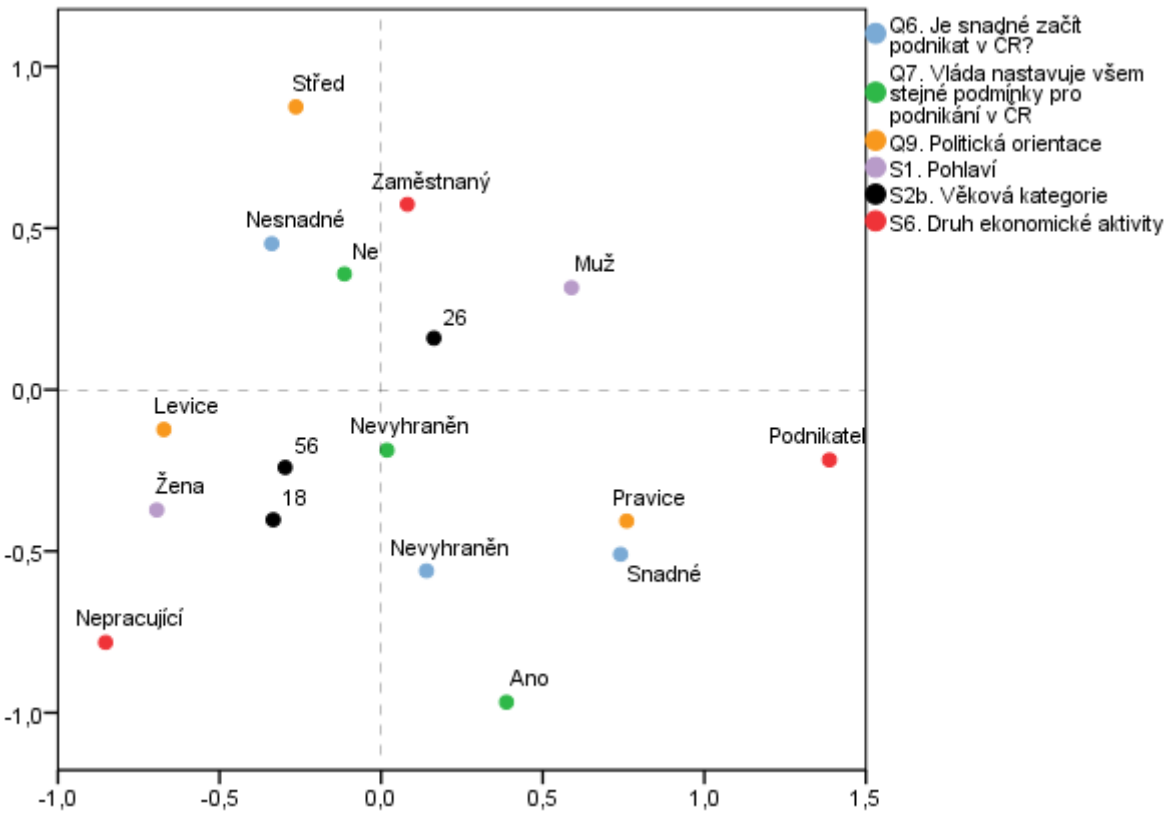
Dimension	Cronbach's Alpha	Variance Accounted For		
		Total (Eigenvalue)	Inertia	% of Variance
1	,465	1,592	,318	31,847
2	,217	1,210	,242	24,203
Total		2,803	,561	
Mean	,358[a]	1,401	,280	28,025

Tabulka T23 zobrazuje základní informace o kvalitě provedené korespondenční analýzy. Ve sloupci *Inertia* je obsažena informace, že se modelem podařilo vysvětlit 56 % celkové variability, z čehož připadá 32 % na první dimenzi a 24 % na dimenzi druhou. Tyto hodnoty nejsou nijak vysoké, ale ve vícerozměrné kategoriální analýze celkem běžné. Cronbachův koeficient alfa, nabývá rovněž velmi nízkých hodnot, které značí velmi slabou konzistenci dat. Opět je nutné brát na zřetel, že v pětirozměrné kontingenční tabulce jsou vysoké hodnoty tohoto koeficientu v podmínkách marketingového průzkumu trhu nereálné.

Samotná korespondenční mapa je zobrazena na grafu G5. Jelikož je proměnná *S2b* (*Věk*) doplňková proměnná, byly souřadnice jejích kategorií do korespondenční mapy dopočteny až po skončení konvergenčních výpočtů souřadnic ostatních proměnných.

Tím bylo zabezpečeno, že její kategorie nijak neovlivnily rozmístění bodů ostatních kategorií na mapě.

G5 – Korespondenční mapa



Z korespondenční mapy lze vyčíst spoustu zajímavých souvislostí mezi analyzovanými proměnnými. První a druhý kvadrant obsahují převážně zaměstnance v produktivním věku. Z blízkosti negativní odpovědi otázky *Q7* a kategorie *Zaměstnaný* u otázky *Q6* ve druhém kvadrantu lze usuzovat, že právě zaměstnanci budou spíše nesouhlasit, že vláda ČR nastavuje všem stejné podmínky a že je snadné v ČR podnikat. Ve čtvrtém kvadrantu korespondenční mapy jsou koncentrovaní pravicoví voliči, podnikatelé a pozitivní odpovědi na obě sledované otázky. Ve třetím kvadrantu se naopak nacházejí levicoví voliči a nepracující. Vzhledem k tomuto rozdělení jsem pojmenoval vertikální osu jako *Politická orientace*, protože odděluje pravicově a levicově smýšlející lidi. Horizontální osa odděluje muže a ženy, zaměstnané a nepracující, ale třeba také lidi v produktivním věku od lidí, kteří ještě nebo už nepracují. Proto horizontální osu nazývám *Sociální status*.

Míry diskriminace (*Discrimination Measures*) mohou být pokládány za obdobu druhých mocnin komponentních zátěží používaných v metodě hlavních komponent. V korespondenční analýze představují druhé mocniny korelací kvantifikovaných

proměnných se svými souřadnicemi (*object scores*). Zároveň vyjadřují míru variability kvantifikované proměnné v jednotlivých dimenzích. Jejich maximální hodnota činí jedna.

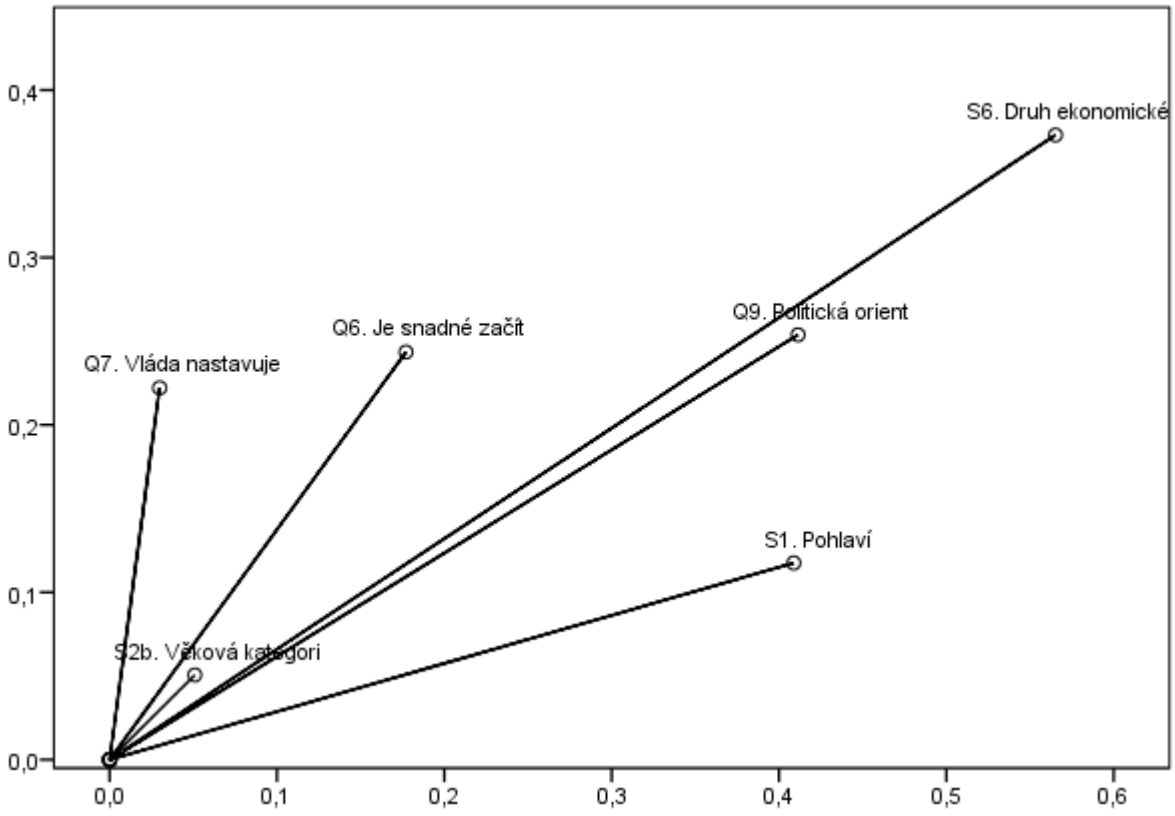
Tabulka T24 zobrazuje míry diskriminace pro všechny proměnné s oběma dimenzemi. Se souřadnicemi v obou dimenzích nejvíce koreluje proměnná *S6* (*Druh ekonomické aktivity*). Vysvětlující proměnné *S6*, *Q9* a *S1* korelují více se svými souřadnicemi v první dimenzi, naopak, vysvětlované proměnné *Q6* a *Q7* korelují více se souřadnicemi v druhé dimenzi.

T24 – Míry diskriminace 1 (upravený výstup SPSS)

	Dimension		Mean
	1	2	
Q7. Vláda nastavuje všem stejné podmínky pro podnikání v ČR	0,030	0,222	0,126
S1. Pohlaví	0,409	0,118	0,263
S6. Druh ekonomické aktivity	0,565	0,373	0,469
Q9. Politická orientace	0,411	0,254	0,333
Q6. Je snadné začít podnikat v ČR?	0,177	0,244	0,210
S2b. Věková kategorie	0,051	0,051	0,051
Celkem	1,592	1,210	1,401
% variability	31,847	24,203	28,025

Variabilitu proměnných v jednotlivých dimenzích lépe znázorňuje graf G6, ve kterém jsou druhé mocniny korelací mezi proměnnými a souřadnicemi vyjádřeny formou úseček s počátkem v bodě $[0;0]$ a koncem v bodě [dimenze1;dimenze2] z tabulky T24. Délka úseček vyjadřuje míru variability dané proměnné. Čím menší úhel svírá úsečka proměnné s danou dimenzí, tím více variability je obsaženo právě v této dimenzi. Variabilita proměnné *S6* je velmi silná a je rozložena mezi obě dimenze rovnoměrně. Téměř celá variabilita proměnné *Q7* je vyjádřena druhou dimenzí. Nejnižší variabilitu vykazuje proměnná *S2b*.

G6 – Míry diskriminace 2



Kapitola 9
Aplikace loglineárních modelů

V SPSS se loglineární modely nacházejí pod nabídkou *Loglinear*, kde se nalézají hned dva způsoby tvorby loglineárních modelů. Obecný loglineární model (*General loglinear*), který je konstruován pomocí *dummy* proměnných, má výhodné podmínky ohledně interpretace parametrů. Druhou možností je hierarchický loglineární model (*Model selection*). Tento typ loglineárního modelu vychází z indikátorových proměnných typu *effect*. Jeho výhodou je možnost použití metody *Backward elimination*. Ta umožňuje výběr co nejjednoduššího modelu při zachování dostatečné shody s původními daty. Parametry tohoto modelu jsou ovšem obtížně věcně interpretovatelné.

Obecný i hierarchický model se stejným předpisem vykazují shodné hodnoty statistiky G^2. Toho jsem se rozhodl využít pro účely své analýzy. K určení vhodného modelu použiji metodu *Backward elimination* obsaženou v hierarchickém modelu a výsledný model zadám do obecného loglineárního modelu, z jehož výstupu získám věcně interpretovatelné parametry.

Tabulka T25 představuje pětirozměrnou tabulku v podobě, v jaké ji chci analyzovat. V řádcích jsou vysvětlující proměnné *Pohlaví*, *Ekonomická aktivita* a *Politická orientace*, ve sloupcích jsou vysvětlované proměnné představující odpovědi na oba výroky týkající se ekonomické situace v ČR.

T25 – Data pro loglineární model

			Q7. Vláda nastavuje všem stejné podmínky pro podnikání v ČR.			Q6. Je snadné začít v ČR podnikat?		
Pohlaví	Ekonomická aktivita	Politická orientace	Ne	Nevyhraněn	Ano	Snadné	Nevyhraněn	Nesnadné
Muž	Podnikatel	Levice	11	4	0	2	2	11
		Střed	21	5	2	8	8	12
		Pravice	57	29	21	38	31	38
	Zaměstnaní	Levice	78	33	18	22	29	78
		Střed	48	30	4	12	17	53
		Pravice	64	44	29	33	43	61
	Nepracující	Levice	26	14	9	12	11	26
		Střed	11	7	7	2	4	19
		Pravice	15	13	6	10	9	15
Žena	Podnikatel	Levice	8	0	0	2	2	4
		Střed	6	3	0	3	1	5
		Pravice	12	8	3	5	10	8
	Zaměstnaný	Levice	47	18	8	9	18	46
		Střed	40	19	7	11	9	46
		Pravice	45	21	14	18	21	41
	Nepracující	Levice	78	36	16	19	39	72
		Střed	30	19	7	8	11	37
		Pravice	28	24	17	10	22	37

9.1 Hierarchický model

Analýzu začnu hierarchickým modelem s proměnnými typu *effect*, do kterého zahrnu všechny proměnné, které chci zkoumat, tedy *Q7*, *Q6*, *S1*, *S6* a *Q9*. Z důvodu hierarchické struktury modelu, je nutné nastavit u každé proměnné rozpětí hodnot, kterých může nabývat. Všechny proměnné kromě *S1* nabývají hodnot od jedné do tří, proměnná *S1* od jedné do dvou. Dále nastavím tvorbu výsledného modelu pomocí iterativní metody *Backward elimination*, která umožňuje přehlednou orientaci v nepřeberném množství proměnných a jejich interakcí. Protože původní maximální počet iterací nastavený na hodnotu deset se mi osvědčil jako nedostatečný, zvýšil jsem jeho hodnotu na dvacet. Vstupní model pro analýzu ponechám výchozí nastavený, tedy saturovaný. Pokud bych vybral některý z jednodušších modelů, metoda *Backward elimination* by započala tímto modelem a pokračovala by dále směrem k jednodušším modelům.

Pro získání představy o závislostech ve sledované tabulce je dobré provést proceduru, která je v SPSS známá pod názvem *K-Way and Higher-Order Effects*, jejíž výstup je zobrazen v tabulce T26. Ta obsahuje dva testy (*K-way and Higher Order Effects* a *K-way Effects*), které testují nulovou hypotézu, že prvky daného rozměru K (hlavní efekty nebo faktory) jsou v modelu nepotřebné. Oba testy spolu úzce souvisí. Zatímco první testuje hypotézu, že daný rozměr a všechny vyšší jsou nepotřebné, druhý testuje hypotézu o významnosti právě testovaného rozměru. Jako testová kritéria jsou použita Pearsonova X^2 statistika a věrohodnostní poměr G^2.

T26 – K-Way and Higher-Order Effects

			Likelihood Ratio		Pearson		Number of
	K	df	Chi-Square	Sig.	Chi-Square	Sig.	Iterations
K-way and Higher Order Effects	1	161	1241,024	,000	1564,571	,000	0
	2	152	479,715	,000	573,384	,000	2
	3	120	116,960	,561	106,052	,814	6
	4	64	55,288	,773	47,272	,942	4
	5	16	15,554	,485	13,248	,655	4
K-way Effects	1	9	761,309	,000	991,188	,000	0
	2	32	362,755	,000	467,332	,000	0
	3	56	61,672	,280	58,780	,374	0
	4	48	39,734	,796	34,024	,936	0
	5	16	15,554	,485	13,248	,655	0

Hodnota 1241,024 v prvním řádku tabulky T26 představuje hodnotu statistiky G^2, pokud by neplatily žádné parametry, a jediným parametrem by tak byl pouze průměr. Za ten by byl odečten jeden stupeň volnosti, takže počet zbývajících stupňů volnosti df je roven 161. Hodnota testového kritéria pro model s efekty prvních a vyšších řádů činí 479,715, pro efekty druhých a vyšších řádů 116,960, atd. Na základě konfrontace těchto modelů se saturovaným modelem jsou vypočteny jejich p-hodnoty. Je-li p-hodnota vyšší než 0,05, pak nebyl K-tý a vyšší rozměr prokázán jako významný.

Druhá část této tabulky testuje významnost efektů konkrétního rozměru K. Počet chí-kvadrát čtverců pro první rozměr se získá jako rozdíl prvních dvou řádků v předchozím testu ($1241{,}024 - 479{,}715 = 761{,}309$). Tato hodnota vyjadřuje, o kolik se model zlepšil po zahrnutí efektů prvního řádu. Podle p-hodnoty v prvním

řádku se určí, zda je zlepšení modelu po přidání efektů prvního rozměru statisticky významné. Obdobným způsobem se postupuje pro efekty ostatních řádů.

V tabulce T26 jsou od třetího řádu všechny p-hodnoty vyšší než 0,05. Efekty třetího a vyššího řádu tak nebyly prokázány jako důležité. Ve výsledném loglineárním modelu bych si tak měl vystačit s maximálně dvoufaktorovými členy.

T27 – Backward elimination

Step			Effects	Chi-Square	df	Sig.	Number of Iterations
16	Generating Class		Q6*Q7*S1, Q6*Q9, Q7*Q9, Q9*S6, S1*S6, Q6*S6, Q7*S6	105,071	118	,797	
	Deleted Effect	1	Q6*Q7*S1	13,859	4	,008	5
		2	Q6*Q9	22,652	4	,000	5
		3	Q7_i4*Q9	25,527	4	,000	4
		4	Q9*S6	103,400	4	,000	4
		5	S1*S6	149,407	2	,000	5
		6	Q6*S6	10,762	4	,029	5
		7	Q7*S6	15,956	4	,003	4
17	Generating Class		Q6*Q7*S1, Q6*Q9, Q7*Q9, Q9*S6, S1*S6, Q6*S6, Q7*S6	105,071	118	,797	

Pro výběr optimálního modelu byla použita metoda *Backward elimination*. V mém případě bylo k nalezení optimálního modelu potřeba 17 kroků. Tabulka T27 zobrazuje poslední dva kroky tohoto procesu, přičemž výsledný model je obsažen v posledním řádku:

$$(Q6\ Q7\ S1, Q6\ Q9, Q7\ Q9, Q9\ S6, S1\ S6, Q6\ S6, Q7\ S6)$$

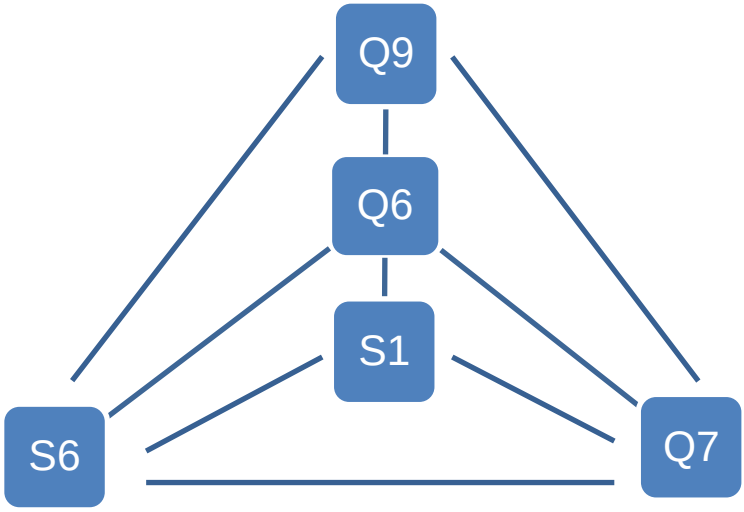
Navrhovaný model obsahuje jeden trojfaktorový člen, ostatní členy jsou dvoufaktorové. Má 118 stupňů volnosti při velikosti statistiky chí-kvadrát 105,071. K namodelování vztahů v kontingenční tabulce stačí $(162 - 118 = 44)$ parametrů. Z existence trojfaktorového členu vyplývá, že je nutné posuzovat vztahy mezi dvoufaktorovými členy vždy k členu trojfaktorovému. Model odhalil vztah mezi proměnnými *Q6* (*Je snadné začít v České republice podnikat?*), *Q7* (*Vláda nastavuje všem stejné podmínky pro podnikání v ČR*) a *S1* (*Pohlaví*). *Politická orientace* (*Q9)* souvisí s oběma vysvětlovanými proměnnými *Q6* i *Q7*. Obdobně, proměnná *S6* (*Ekonomická aktivita*) souvisí také s proměnnými *Q6* a *Q7*. Dále byly prokázány

vztahy mezi vysvětlujícími proměnnými *Pohlaví* (*S1*) a *Politická orientace* (*S6*), jejichž vzájemný vztah jsem sice primárně nezkoumal, ale je součástí loglineárního modelu a vypustit jej není možné. To samé platí pro proměnné *Q9* (*Politická orientace*) a *S6* (*Ekonomická aktivita*). Velké množství vzájemných vztahů přehledně zobrazuje graf nezávislosti.

Graf nezávislosti (*independence graph*) slouží k zobrazení asociací v loglineárním modelu. Je složen z *vrcholů* (*vertices*), které představují proměnné v kontingenční tabulce, a *hran* (*edges*), které značí podmíněnou závislost určitých párů proměnných. Chybí-li mezi dvěma proměnnými spojení pomocí hrany, jsou podmíněně nezávislé. Dvě proměnné, které nejsou vzájemně propojeny hranou, mohou být propojitelné posloupností několika hran, tzv. *cestou* (*path*). Takové proměnné jsou podmíněně nezávislé za podmínky daných hodnot proměnných, kterými je tvořena cesta.

V grafu G7 je dobře vidět, že všechny proměnné vzájemně podmíněně závislé, kromě páru vysvětlujících proměnných *S1* a *Q9*. Tyto proměnné jsou podmíněně nezávislé za podmínky dané hodnoty *Q6*. Všechny ostatní páry proměnných jsou přímo spojeny hranou, existuje mezi nimi tedy podmíněná závislost.

G7 – Graf nezávislosti



T28 – Test dobré shody

	Chi-Square	df	Sig.
Likelihood Ratio	105,071	118	,797
Pearson	95,415	118	,937

V tabulce T28 jsou prezentovány testy dobré shody, které testují nulovou hypotézu, že se mnou zvolený model shoduje s výběrovými daty. Z výsledků je zjevné, že nulová hypotéza nebyla zamítnuta. Model je tedy vyhovující.

9.2 Obecný loglineární model

Nyní zadám získaný hierarchický model do procedury *General loglinear*, která používá *dummy* kódování. Pravděpodobnostní rozdělení četností zvolím Poissonovo.

Z tabulky T29 je zřejmé, že výsledný model je totožný s modelem pořízeným v *Model Selection*. Součty čtverců obou chí-kvadrát statistik se shodují a také počty stupňů volnosti jsou shodné. Z tohoto srovnání plyne závěr, že přestože mají porovnávané modely odlišné parametry, oba docházejí ke stejným závěrům. Odlišnost parametrů je totiž způsobena pouze použitím indikátorů jiného typu.

T29 – Test dobré shody 2

	Value	df	Sig.
Likelihood Ratio	105,071	118	,797
Pearson Chi-Square	95,411	118	,937

Odhady parametrů pro první sadu proměnných jsou v tabulce T30, kompletní seznam lze nalézt v příloze 2.

T30 – Odhady parametrů

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	1,476	,288	5,122	,000	,911	2,040
[Q7 = 1] * [q6 = 1] * [s1 = 1]	,307	,346	,887	,375	-,372	,987
[Q7 = 1] * [q6 = 1] * [s1 = 2]	1,273	,321	3,969	,000	,644	1,902
[Q7 = 1] * [q6 = 2] * [s1 = 1]	-,174	,369	-,470	,638	-,897	,550
[Q7 = 1] * [q6 = 2] * [s1 = 2]	,665	,342	1,944	,052	-,005	1,336
[Q7 = 1] * [q6 = 3] * [s1 = 1]	-,709	,344	-2,058	,040	-1,383	-,034
[Q7 = 1] * [q6 = 3] * [s1 = 2]	,426	,317	1,346	,178	-,194	1,047
[Q7 = 2] * [q6 = 1] * [s1 = 1]	-,037	,361	-,102	,919	-,745	,671
[Q7 = 2] * [q6 = 1] * [s1 = 2]	,953	,334	2,858	,004	,300	1,607
[Q7 = 2] * [q6 = 2] * [s1 = 1]	-,108	,373	-,289	,772	-,839	,623
[Q7 = 2] * [q6 = 2] * [s1 = 2]	,836	,343	2,435	,015	,163	1,509
[Q7 = 2] * [q6 = 3] * [s1 = 1]	-,424	,354	-1,200	,230	-1,118	,269
[Q7 = 2] * [q6 = 3] * [s1 = 2]	-,610	,400	-1,526	,127	-1,394	,174
[Q7 = 3] * [q6 = 1] * [s1 = 1]	-,171	,337	-,507	,612	-,830	,489
[Q7 = 3] * [q6 = 1] * [s1 = 2]	,538	,321	1,677	,094	-,091	1,167
[Q7 = 3] * [q6 = 2] * [s1 = 1]	-,808	,391	-2,063	,039	-1,575	-,040
[Q7 = 3] * [q6 = 2] * [s1 = 2]	,117	,366	,321	,748	-,599	,834
[Q7 = 3] * [q6 = 3] * [s1 = 1]	-,329	,320	-1,028	,304	-,955	,298
[Q7 = 3] * [q6 = 3] * [s1 = 2]	0[a]	.	.	.	.	.

V tabulce T30 jsou uvedeny odhady parametrů pro první sadu proměnných modelu, tedy proměnných *Q7*, *Q6* a *S1*, přičemž proměnné *Q7* a *Q6* vyjadřují odpovědi na obě otázky kladené respondentům a proměnná *S1* značí pohlaví respondenta. Všechny parametry, včetně standardních chyb, jsou zobrazeny jako přirozené logaritmy. Chceme-li je věcně interpretovat, musíme použít jejich odlogaritmované hodnoty. V této tabulce tedy jsou porovnávány všechny kombinace s ženami, které odpověděly na obě položené otázky pozitivně, tzn., že je snadné v ČR podnikat a že vláda nastavuje všem stejné podmínky pro podnikání.

Porovnám-li muže a ženy, zjistím, že ženy odpovídaly na položené otázky pozitivněji než muži. Šance, že muž odpoví na obě otázky kladně je 0,7. Žena tedy na tyto otázky odpoví kladně 1,4krát spíše ($1/0{,}7 = 1{,}4$). Za zajímavý dále považuji výrazný rozdíl mezi zastoupením mužů a žen v podnikání. Šance, že se podnikatelem stane muž, je téměř devětkrát vyšší ($\exp 1{,}181 = 8{,}86$), než je tomu u žen.

Z hlediska levicově-pravicového rozdělení je nespokojenost levicových voličů vůči překážkám v podnikání 1,3krát ($\exp 0{,}217 = 1{,}286$) vyšší než u voličů pravicových.

Na otázku, zda je v ČR snadné podnikat se výrazně liší názory podnikatelů a zbytku populace. U podnikatelů je téměř dvakrát větší šance, že odpoví kladně, než lidé, kteří nepodnikají. Pravděpodobně to bude způsobeno skutečností, že se podnikatelé v administrativě podnikání mnohem lépe orientují. Tento fakt souvisí i názorem podnikatelů na rovnost podnikání, u nichž je více než 2,5krát větší šance souhlasu s výrokem, že vláda nastavuje všem stejné podmínky pro podnikání oproti zbytku populace.

9.2.1 Kvalita výsledného modelu

Tabulka T31 obsahuje hodnoty napozorovaných a očekávaných dat pro část kompletní tabulky. Ve sloupci *Residual* jsou obsaženy jejich rozdíly. Aby bylo možné identifikovat, které rozdíly značí významné odchylky modelu od skutečnosti, používají se adjustovaná rezidua, která jsou porovnávána s kvantily normálního rozdělení. Tato identifikace se používá především pro případy, že jinak dobrý model nevykazuje dobrou shodu s daty kvůli několika extrémním četnostem v tabulce.

T31 – Výběrové, očekávané četnosti a rezidua

Q7	Q6	S6	Q9BB	S1	Observed Count	Observed %	Expected Count	Expected %	Residual	Adjusted Residual
ne	snadné	podnikatel	levice	Muž	1	,1%	2,752	,2%	-1,752	-1,201
				Žena	2	,2%	,966	,1%	1,034	1,111
			střed	Muž	5	,4%	4,116	,4%	,884	,515
				Žena	2	,2%	1,446	,1%	,554	,497
			pravice	Muž	13	1,2%	15,777	1,4%	-2,777	-1,027
				Žena	2	,2%	5,540	,5%	-3,540	-1,845
		zaměstnaný	levice	Muž	8	,7%	10,463	,9%	-2,463	-,972
				Žena	7	,6%	8,541	,8%	-1,541	-,652
			střed	Muž	6	,5%	6,836	,6%	-,836	-,387
				Žena	8	,7%	5,580	,5%	2,420	1,208
			pravice	Muž	16	1,4%	11,486	1,0%	4,514	1,712
				Žena	10	,9%	9,376	,8%	,624	,254
		nepracující	levice	Muž	6	,5%	3,860	,3%	2,140	1,226
				Žena	15	1,3%	12,006	1,1%	2,994	1,174
			střed	Muž	0	,0%	1,558	,1%	-1,558	-1,329
				Žena	4	,4%	4,846	,4%	-,846	-,456
			pravice	Muž	4	,4%	2,153	,2%	1,847	1,356
				Žena	5	,4%	6,698	,6%	-1,698	-,798

V zobrazované části tabulky není žádná hodnota ve sloupci *Adjusted Residual* v absolutní hodnotě větší než hodnota 95% kvantilu normovaného normálního rozdělení, tedy 1,96. V celé tabulce, která je obsažena v příloze 3, se vyskytují pouze dvě významné odchylky modelu od napozorovaných dat, které jsou prezentovány v tabulce T32. Tyto odchylky jsou označeny na grafu G8 pomocí křížků.

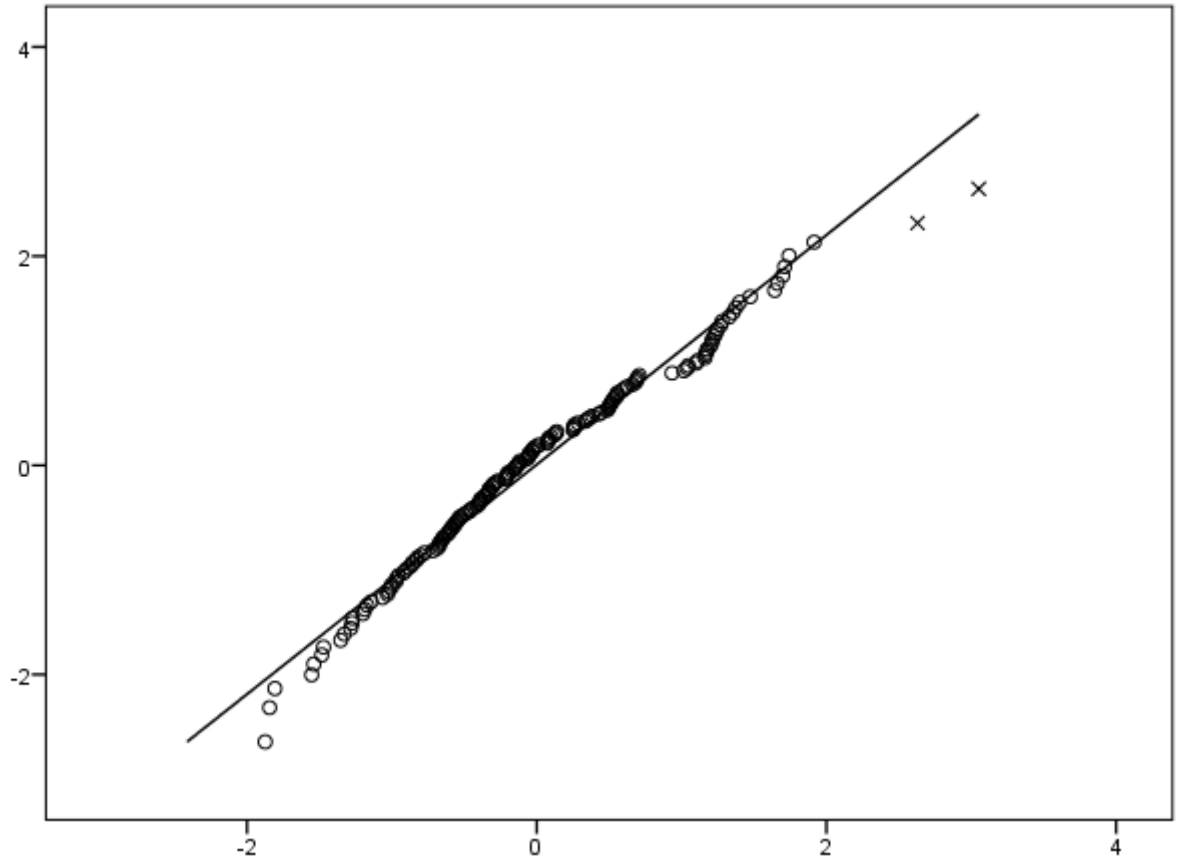
T32 – Odlehlé hodnoty

Q7	Q6	S6	Q9	S1	Adjusted Residual
ano	snadné	podnikatel	pravice	muž	2,628
ano	nesnadné	nepracující	střed	muž	3,052

Graf Q-Q (*quantile-quantile*) porovnává dvě pravděpodobnostní rozdělení tak, že staví hodnoty uspořádaných kvantilů proti sobě. Dvě rozdělení je možné považovat za shodná, jsou-li body v grafu uspořádány podle funkce $y = x$, tedy pod úhlem 45°. V této práci použiji *Q-Q* graf k porovnání rozdělení adjustovaných reziduí s kvantily normovaného normálního rozdělení. Test normality reziduí je významným diagnostickým nástrojem kvality modelu. Jestliže model dostatečně vysvětluje vztahy mezi proměnnými, náhodná složka (tedy přeneseně rezidua) obsahuje pouze bílý

šum, který má normované normální rozdělení. Proto je vlastnost normality reziduí považována za významný indikátor kvality modelu.

G8 – Q-Q graf

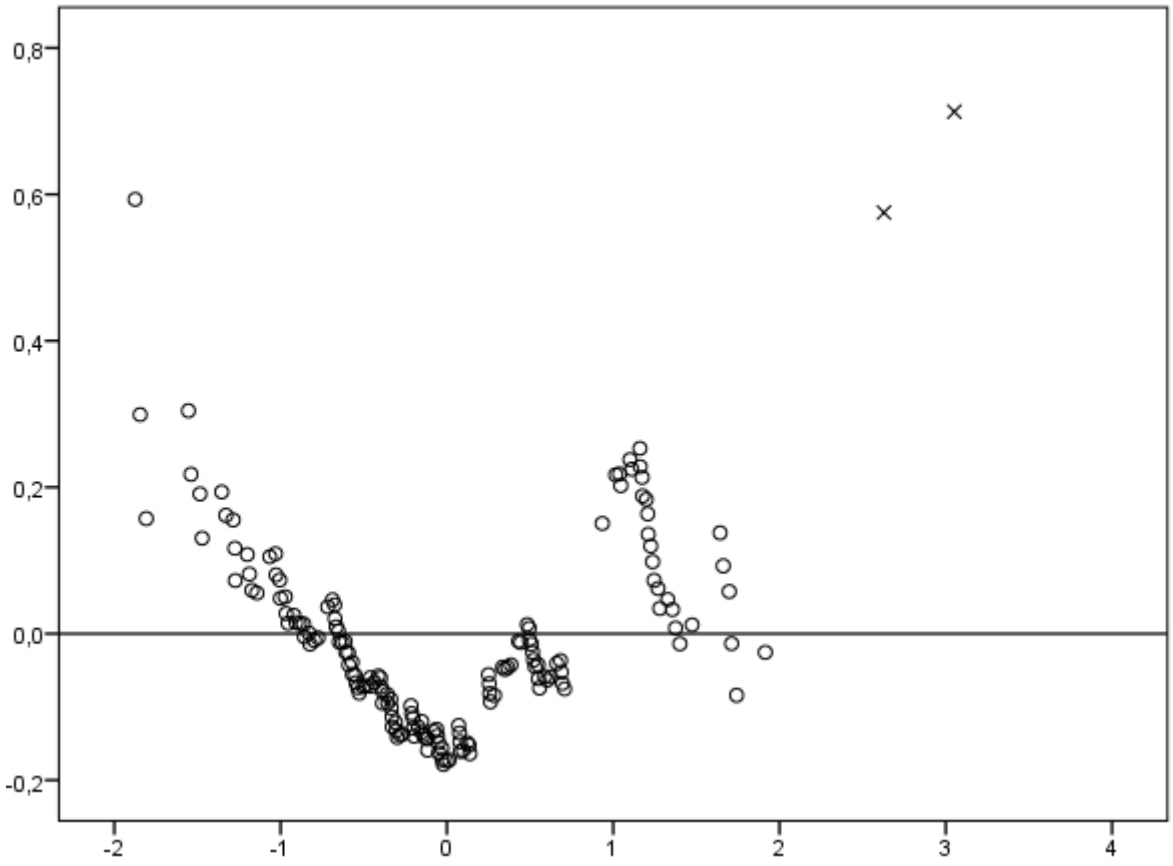


V grafu G8 jsou porovnávána adjustovaná rezidua s teoretickými kvantily normovaného normálního rozdělení. Jelikož data se příliš neodchylují od přímky značící úhel 45°, je možné považovat rezidua výsledného modelu za normovaná normální. Na počátku i konci grafu je několik odlehlých pozorování, která se mírně odchylují od přímky v grafu. Výskyt takových pozorování není nic výjimečného.

Obdobný pohled na data nabízí tzv. *Detrended Q-Q graf*. Jedná se vlastně o *Q-Q* graf zbavený trendu. Hodnoty na ypsilonové ose jsou zde získány jako rozdíl normovaných teoretických kvantilů a adjustovaných reziduí. Detrended *Q-Q* graf umožňuje detailnější pohled na odchylky od normality než klasický *Q-Q* graf. Na grafu G9 je dobře vidět, že většina odchylek od normality osciluje v rozmezí $\pm 0{,}3$ od normovaného normálního rozdělení. Detrended *Q-Q* graf odhalil tři pozorování, která mohou být na základě grafické analýzy označená za odlehlá. Pozorování umístěné nejvíce vlevo se však do 95% intervalu spolehlivosti vešlo. I přes tato

pozorování lze označit rezidua modelu označit za normální. Výsledný model je tedy kvalitní.

G9 – Detrended Q-Q



Závěr

Korespondenční analýza a loglineární modely patří mezi nejvýznamnější metody analýzy kontingenčních tabulek. Obě metody se zabývají vztahy mezi kategoriemi kategoriálních proměnných. Přesto si nijak nekonkurují, naopak, vzájemně se doplňují. Korespondenční analýza je metodou spíše popisného charakteru. Její výhodou je možnost grafického zobrazení vícerozměrného prostoru do dvourozměrné korespondenční mapy, která dostatečně zobrazuje rozdíly a podobnosti mezi kategoriemi sledovaných proměnných. Loglineární modely přistupují ke stejné problematice z odlišného pohledu. Primárně jsou nástrojem pro modelování četností buněk v kontingenční tabulce. Významným důsledkem této činnosti je vyjádření vztahů mezi kategoriemi a kategoriálními proměnnými prostřednictvím loglineárních parametrů. Parametry typu *dummy* pak lze navíc vyjádřit pomocí logaritmické transformace v podobě šancí.

V případě rozhodování mezi korespondenční analýzou a loglineárními modely záleží na výsledných očekáváních od analýzy. Má-li být jejím účelem pouze získání přehledu o analyzovaných datech, případně zjištění vztahů mezi kategoriemi proměnných, je korespondenční analýza vhodnou volbou. Je-li potřeba ověřit vztahy mezi kategoriemi pomocí hypotéz, nebo namodelovat četnosti kontingenční tabulky, lze doporučit loglineární modely. Nejvýhodnější je použít obě analýzy v návaznosti na sebe, kdy korespondenční analýza slouží pro předběžnou analýzu dat před samotným použitím loglineárních modelů.

V praktické části práce jsem se zabýval analýzou názoru veřejnosti na podnikatelské prostředí v České republice. Konkrétně se jednalo o otázky, zda je možnost začít podnikat v České republice pro všechny stejná a zda mají všichni stejně nastavené podmínky pro podnikání. Odpovědi na obě otázky měly velmi podobný průběh. U obou výrazně převažovaly negativní odpovědi. Podobnost se projevovala i v korespondenční mapě, ve které se obdobné kategorie jednotlivých otázek vyskytovaly nedaleko od sebe. Vysvětluji si to tak, že respondenti příliš nerozlišují mezi překážkami na začátku podnikání a v jeho průběhu. Další možností je, že

respondenti, především z řad nepodnikatelů, nemají o problematice podnikání dostatečný přehled.

I přes poměrně malé procento vysvětlené inerce v korespondenční analýze, považuji za velmi důležitý její hlavní výstup, tedy korespondenční mapu. Vertikální osu této mapy jsem pojmenoval *Politická orientace*, horizontální *Sociální stav*. Tyto osy rozdělily korespondenční mapu na čtyři kvadranty, na jejichž základě se dají snadno diverzifikovat názory na podnikání v ČR s ohledem na politickou orientaci, druh ekonomické aktivity nebo věk.

Analýza metodou loglineárních modelů přinesla poměrně kvalitní model, který se pouze u dvou z celkového počtu 162 četností buněk výrazně lišil od pozorovaných četností. Daní za tento kvalitní model je přítomnost trojfaktorového vztahu v předpisu modelu, na který se musí brát zřetel i při interpretaci některých dvoufaktorových členů. Při podrobnějším pohledu na trojfaktorový člen je ale zřejmé, že velké problémy při jeho interpretaci nenastanou. Obsahuje totiž obě proměnné vyjadřující spokojenost s podnikáním v ČR, u kterých jsem výše osvětlil, že mají velmi podobnou interpretaci. Použitím obecného loglineárního modelu s parametry typu *dummy*, mi bylo umožněno interpretovat zajímavé vztahy mezi kategoriemi proměnných prostřednictvím šancí, jejichž pomocí jsem mohl konkrétněji vyjádřit vztahy v analyzované kontingenční tabulce.

V úvodu práce jsem si stanovil za cíl posoudit vhodnost analyzovaných metod pro účely marketingového průzkumu. Způsob, jakým jsem data zkoumal, však praktikám těchto průzkumů příliš neodpovídá. Na základě této práce tedy nemohu podložit vhodnost či nevhodnost analyzovaných metod. Můj názor je, že v určitých případech by použití popisovaných metod význam mělo, např. při sledování asociací mezi významnými kategoriemi, zejména u vícenásobné korespondenční analýzy. Jednoduchá korespondenční analýza se v průzkumech trhu běžně používá, a to k vyjádření korespondence (podobnosti) mezi produkty a jejich vlastnostmi. Použití vícenásobné korespondenční analýzy ve smyslu zjištění asociace mezi kategoriemi proměnných, se ale v marketingových průzkumech nevyskytuje. Jejich zadavatelům totiž obvykle stačí procentuální zastoupení kategorií jednotlivých proměnných a o závislostní vztahy se příliš nezajímají. Přitom by známá asociační struktura mohla přispět k lepšímu pochopení zkoumané oblasti. Tato práce tak může posloužit jako inspirace k dalšímu využití korespondenční analýzy v marketingových průzkumech.

Přílohy

Příloha 1 – Syntaxe SPSS

```
var lab s1 "S1. Pohlaví".

recode s6 (1=1) (2=1) (3=2) (4=2) (5=3) (6=3) (7=3) (8=3) (9=3).
val lab
/s6
1 "Podnikatel"
2 "Zaměstnaný"
3 "Nepracující"
.
exe.

var lab s6 "S6. Druh ekonomické aktivity".

recode q9b (1 thru 11 = copy)  (99 = 11).
val lab
/q9b
1 "ČSSD"
2 "KSČM"
3 "KDU-ČSL"
4 "ODS"
5 "Strana práv občanů"
6 "Strana zelených"
7 "Suverenita"
8 "TOP 09"
9 "Věci veřejné"
10 "Jiná strana, nezávislí"
11 "Neví"
.
exe.

recode q9b (1=1) (2=1) (5=1) (3=2) (6=2) (9=2) (10=2) (4=3) (7=3) (8=3)
(11=sys) into Q9.
val lab
/Q9
1 "Levice"
2 "Střed"
3 "Pravice"
.
exe.

var lab q9 "Q9. Politická orientace".
exe.

recode s2b (1=1) (2=2) (3=2) (4=2) (5=3).
val lab
/s2b
1 "18-25"
2 "26-55"
3 "56-65".

var lab s2b "S2b. Věková kategorie".
```

```
recode q6 (1=3) (2=3) (3=2) (4=1) (5=1).
val lab
/q6
1 "Nesnadné"
2 "Nevyhraněn"
3 "Snadné"
.
exe.

var lab q6 "Q6. Je snadné začít podnikat v ČR?".

ren var (q7_i4 = Q7) (s1 = S1) (s6 = S6) (s2b = S2b) (q6 = Q6).

recode Q7 (0 thru 3 = 1) (4 thru 6 = 2) (7 thru 10 = 3).

var lab Q7 "Q7. Vláda nastavuje všem stejné podmínky pro podnikání v ČR".

val lab
/Q7
1 "Ne"
2 "Nevyhraněn"
3 "Ano"
.
exe.

*vícenásobná korespondenční analýza.
MULTIPLE CORRES VARIABLES=Q7 S1 S6 Q9 Q6 S2b
  /ANALYSIS=Q7(WEIGHT=1) S1(WEIGHT=1) S6(WEIGHT=1) Q9(WEIGHT=1)
Q6(WEIGHT=1) S2b
  /MISSING=Q7(LISTWISE) S1(LISTWISE) S6(LISTWISE) Q9(LISTWISE) Q6(LISTWISE)
S2b(PASSIVE,MODEIMPU)
  /SUPPLEMENTARY=VARIABLE(S2b)
  /DIMENSION=2
  /NORMALIZATION=VPRINCIPAL
  /MAXITER=100
  /CRITITER=.00001
  /PRINT=CORR DISCRIM
  /PLOT=OBJECT(20) JOINTCAT(Q7 S1 S6 Q9 Q6 S2b) (20) DISCRIM (20).

*loglineární model - model selection.
HILOGLINEAR Q7(1 3) Q6(1 3) S6(1 3) Q9(1 3) S1(1 2)
  /METHOD=BACKWARD
  /CRITERIA MAXSTEPS(20) P(.05) ITERATION(20) DELTA(.5)
  /PRINT=NONE
  /DESIGN.

*loglineární model - general.
GENLOG Q7 Q6 S6 Q9 S1
  /MODEL=POISSON
  /PRINT=FREQ RESID ADJRESID ZRESID DEV ESTIM CORR COV
  /PLOT=RESID(ADJRESID) NORMPROB(ADJRESID)
  /CRITERIA=CIN(95) ITERATE(20) CONVERGE(0.001) DELTA(.5)
  /DESIGN Q6*Q7*S1 Q6*Q9 Q7*Q9 Q9*S6 S1*S6 Q6*S6 Q7*S6.
```

Příloha 2 – Parametry loglineárního modelu

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval		odds ratio
					Lower	Upper	
Constant	1,476	,288	5,122	,000	,911	2,040	4,374
[Q7 = 1] * [Q6 = 1] * [S1 = 1]	,307	,346	,887	,375	-,372	,987	1,360
[Q7 = 1] * [Q6 = 1] * [S1 = 2]	1,273	,321	3,969	,000	,644	1,902	3,572
[Q7 = 1] * [Q6 = 2] * [S1 = 1]	-,174	,369	-,470	,638	-,897	,550	,841
[Q7 = 1] * [Q6 = 2] * [S1 = 2]	,665	,342	1,944	,052	-,005	1,336	1,945
[Q7 = 1] * [Q6 = 3] * [S1 = 1]	-,709	,344	-2,058	,040	-1,383	-,034	,492
[Q7 = 1] * [Q6 = 3] * [S1 = 2]	,426	,317	1,346	,178	-,194	1,047	1,531
[Q7 = 2] * [Q6 = 1] * [S1 = 1]	-,037	,361	-,102	,919	-,745	,671	,964
[Q7 = 2] * [Q6 = 1] * [S1 = 2]	,953	,334	2,858	,004	,300	1,607	2,594
[Q7 = 2] * [Q6 = 2] * [S1 = 1]	-,108	,373	-,289	,772	-,839	,623	,898
[Q7 = 2] * [Q6 = 2] * [S1 = 2]	,836	,343	2,435	,015	,163	1,509	2,308
[Q7 = 2] * [Q6 = 3] * [S1 = 1]	-,424	,354	-1,200	,230	-1,118	,269	,654
[Q7 = 2] * [Q6 = 3] * [S1 = 2]	-,610	,400	-1,526	,127	-1,394	,174	,543
[Q7 = 3] * [Q6 = 1] * [S1 = 1]	-,171	,337	-,507	,612	-,830	,489	,843
[Q7 = 3] * [Q6 = 1] * [S1 = 2]	,538	,321	1,677	,094	-,091	1,167	1,713
[Q7 = 3] * [Q6 = 2] * [S1 = 1]	-,808	,391	-2,063	,039	-1,575	-,040	,446
[Q7 = 3] * [Q6 = 2] * [S1 = 2]	,117	,366	,321	,748	-,599	,834	1,125
[Q7 = 3] * [Q6 = 3] * [S1 = 1]	-,329	,320	-1,028	,304	-,955	,298	,720
[Q7 = 3] * [Q6 = 3] * [S1 = 2]	0[a]	.	.	.	.	.	.
[Q6 = 1] * [Q9 = 1,00]	,252	,217	1,159	,246	-,174	,678	1,287
[Q6 = 1] * [Q9 = 2,00]	-,594	,262	-2,266	,023	-1,108	-,080	,552
[Q6 = 1] * [Q9 = 3,00]	0[a]	.	.	.	.	.	.
[Q6 = 2] * [Q9 = 1,00]	-,136	,239	-,569	,570	-,605	,333	,873
[Q6 = 2] * [Q9 = 2,00]	-1,408	,296	-4,753	,000	-1,988	-,827	,245
[Q6 = 2] * [Q9 = 3,00]	0[a]	.	.	.	.	.	.
[Q6 = 3] * [Q9 = 1,00]	-,230	,245	-,939	,348	-,709	,250	,795
[Q6 = 3] * [Q9 = 2,00]	-1,221	,294	-4,159	,000	-1,797	-,646	,295
[Q6 = 3] * [Q9 = 3,00]	0[a]	.	.	.	.	.	.
[Q7 = 1] * [Q9 = 1,00]	,813	,210	3,869	,000	,401	1,225	2,255
[Q7 = 1] * [Q9 = 2,00]	,898	,250	3,597	,000	,409	1,387	2,454
[Q7 = 1] * [Q9 = 3,00]	0[a]	.	.	.	.	.	.
[Q7 = 2] * [Q9 = 1,00]	,321	,229	1,397	,162	-,129	,770	1,378
[Q7 = 2] * [Q9 = 2,00]	,724	,266	2,723	,006	,203	1,246	2,063
[Q7 = 2] * [Q9 = 3,00]	0[a]	.	.	.	.	.	.
[Q7 = 3] * [Q9 = 1,00]	0[a]	.	.	.	.	.	.
[Q7 = 3] * [Q9 = 2,00]	0[a]	.	.	.	.	.	.
[Q7 = 3] * [Q9 = 3,00]	0[a]	.	.	.	.	.	.

[S6 = 1] * [Q9 = 1,00]	-3,496	,404	-8,653	,000	-4,288	-2,704
[S6 = 1] * [Q9 = 2,00]	-2,186	,395	-5,531	,000	-2,961	-1,411
[S6 = 1] * [Q9 = 3,00]	-1,166	,342	-3,410	,001	-1,837	-,496
[S6 = 2] * [Q9 = 1,00]	-,819	,258	-3,180	,001	-1,324	-,314
[S6 = 2] * [Q9 = 2,00]	-,337	,276	-1,221	,222	-,879	,204
[S6 = 2] * [Q9 = 3,00]	-,142	,249	-,572	,568	-,630	,345
[S6 = 3] * [Q9 = 1,00]	0[a]	.	.	.	.	.
[S6 = 3] * [Q9 = 2,00]	0[a]	.	.	.	.	.
[S6 = 3] * [Q9 = 3,00]	0[a]	.	.	.	.	.
[S6 = 1] * [S1 = 1]	2,181	,214	10,184	,000	1,761	2,601
[S6 = 1] * [S1 = 2]	0[a]	.	.	.	.	.
[S6 = 2] * [S1 = 1]	1,338	,145	9,254	,000	1,054	1,621
[S6 = 2] * [S1 = 2]	0[a]	.	.	.	.	.
[S6 = 3] * [S1 = 1]	0[a]	.	.	.	.	.
[S6 = 3] * [S1 = 2]	0[a]	.	.	.	.	.
[Q6 = 1] * [S6 = 1]	-,648	,255	-2,546	,011	-1,147	-,149
[Q6 = 1] * [S6 = 2]	,030	,196	,155	,877	-,353	,414
[Q6 = 1] * [S6 = 3]	0[a]	.	.	.	.	.
[Q6 = 2] * [S6 = 1]	-,400	,282	-1,416	,157	-,953	,154
[Q6 = 2] * [S6 = 2]	-,096	,222	-,432	,665	-,531	,339
[Q6 = 2] * [S6 = 3]	0[a]	.	.	.	.	.
[Q6 = 3] * [S6 = 1]	0[a]	.	.	.	.	.
[Q6 = 3] * [S6 = 2]	0[a]	.	.	.	.	.
[Q6 = 3] * [S6 = 3]	0[a]	.	.	.	.	.
[Q7 = 1] * [S6 = 1]	,976	,296	3,295	,001	,396	1,557
[Q7 = 1] * [S6 = 2]	,479	,207	2,307	,021	,072	,885
[Q7 = 1] * [S6 = 3]	0[a]	.	.	.	.	.
[Q7 = 2] * [S6 = 1]	,303	,323	,938	,348	-,330	,937
[Q7 = 2] * [S6 = 2]	,214	,224	,955	,340	-,225	,652
[Q7 = 2] * [S6 = 3]	0[a]	.	.	.	.	.
[Q7 = 3] * [S6 = 1]	0[a]	.	.	.	.	.
[Q7 = 3] * [S6 = 2]	0[a]	.	.	.	.	.
[Q7 = 3] * [S6 = 3]	0[a]	.	.	.	.	.

Příloha 3 – Rezidua loglineárního modelu

					Observed		Expected		Residual	Adjusted Residual
Q7	Q6	S6	Q9	S1	Count	%	Count	%		
Ne	Nesnadné	Podnikatel	Levice	Muž	9	,8%	6,436	,6%	2,564	1,280
				Žena	4	,4%	1,908	,2%	2,092	1,643
			Střed	Muž	10	,9%	11,134	1,0%	-1,134	-,461
				Žena	4	,4%	3,302	,3%	,698	,431
			Pravice	Muž	26	2,3%	22,792	2,0%	3,208	1,016
				Žena	4	,4%	6,759	,6%	-2,759	-1,273
		Zaměstnaný	Levice	Muž	54	4,8%	48,228	4,3%	5,772	1,269
				Žena	31	2,8%	33,245	3,0%	-2,245	-,536
			Střed	Muž	33	2,9%	36,441	3,3%	-3,441	-,830
				Žena	27	2,4%	25,120	2,2%	1,880	,495
			Pravice	Muž	25	2,2%	32,703	2,9%	-7,703	-1,875
				Žena	25	2,2%	22,544	2,0%	2,456	,661
		Nepracující	Levice	Muž	14	1,3%	17,261	1,5%	-3,261	-1,005
				Žena	43	3,8%	45,337	4,0%	-2,337	-,553
			Střed	Muž	9	,8%	8,057	,7%	,943	,385
				Žena	21	1,9%	21,162	1,9%	-,162	-,049
			Pravice	Muž	9	,8%	5,948	,5%	3,052	1,401
				Žena	16	1,4%	15,624	1,4%	,376	,122
	Nevyhraněn	Podnikatel	Levice	Muž	1	,1%	3,459	,3%	-2,459	-1,539
				Žena	2	,2%	,904	,1%	1,096	1,211
			Střed	Muž	6	,5%	3,912	,3%	2,088	1,237
				Žena	0	,0%	1,022	,1%	-1,022	-1,065
			Pravice	Muž	18	1,6%	18,059	1,6%	-,059	-,021
				Žena	6	,5%	4,718	,4%	1,282	,695
		Zaměstnaný	Levice	Muž	16	1,4%	17,825	1,6%	-1,825	-,590
				Žena	9	,8%	10,825	1,0%	-1,825	-,688
			Střed	Muž	9	,8%	8,804	,8%	,196	,082
				Žena	5	,4%	5,346	,5%	-,346	-,172
			Pravice	Muž	23	2,1%	17,818	1,6%	5,182	1,662
				Žena	10	,9%	10,820	1,0%	-,820	-,309
		Nepracující	Levice	Muž	6	,5%	7,238	,6%	-1,238	-,546
				Žena	20	1,8%	16,748	1,5%	3,252	1,100
			Střed	Muž	2	,2%	2,208	,2%	-,208	-,151
				Žena	5	,4%	5,110	,5%	-,110	-,057
			Pravice	Muž	2	,2%	3,677	,3%	-1,677	-,967
				Žena	7	,6%	8,508	,8%	-1,508	-,631

	Snadné	Podnikatel	Levice	Muž	1	,1%	2,752	,2%	-1,752	-1,201
				Žena	2	,2%	,966	,1%	1,034	1,111
			Střed	Muž	5	,4%	4,116	,4%	,884	,515
				Žena	2	,2%	1,446	,1%	,554	,497
			Pravice	Muž	13	1,2%	15,777	1,4%	-2,777	-1,027
				Žena	2	,2%	5,540	,5%	-3,540	-1,845
		Zaměstnaný	Levice	Muž	8	,7%	10,463	,9%	-2,463	-,972
				Žena	7	,6%	8,541	,8%	-1,541	-,652
			Střed	Muž	6	,5%	6,836	,6%	-,836	-,387
				Žena	8	,7%	5,580	,5%	2,420	1,208
			Pravice	Muž	16	1,4%	11,486	1,0%	4,514	1,712
				Žena	10	,9%	9,376	,8%	,624	,254
		Nepracující	Levice	Muž	6	,5%	3,860	,3%	2,140	1,226
				Žena	15	1,3%	12,006	1,1%	2,994	1,174
			Střed	Muž	0	,0%	1,558	,1%	-1,558	-1,329
				Žena	4	,4%	4,846	,4%	-,846	-,456
			Pravice	Muž	4	,4%	2,153	,2%	1,847	1,356
				Žena	5	,4%	6,698	,6%	-1,698	-,798
Nevyhraněn	Nesnadné	Podnikatel	Levice	Muž	2	,2%	1,421	,1%	,579	,522
				Žena	0	,0%	,432	,0%	-,432	-,675
			Střed	Muž	1	,1%	3,384	,3%	-2,384	-1,485
				Žena	1	,1%	1,028	,1%	-,028	-,029
			Pravice	Muž	9	,8%	8,238	,7%	,762	,336
				Žena	4	,4%	2,503	,2%	1,497	1,046
		Zaměstnaný	Levice	Muž	15	1,3%	16,023	1,4%	-1,023	-,334
				Žena	11	1,0%	11,319	1,0%	-,319	-,117
			Střed	Muž	16	1,4%	16,660	1,5%	-,660	-,216
				Žena	15	1,3%	11,769	1,1%	3,231	1,177
			Pravice	Muž	22	2,0%	17,782	1,6%	4,218	1,328
				Žena	12	1,1%	12,562	1,1%	-,562	-,197
		Nepracující	Levice	Muž	7	,6%	7,474	,7%	-,474	-,204
				Žena	18	1,6%	20,118	1,8%	-2,118	-,665
			Střed	Muž	4	,4%	4,801	,4%	-,801	-,413
				Žena	12	1,1%	12,922	1,2%	-,922	-,337
			Pravice	Muž	4	,4%	4,216	,4%	-,216	-,117
				Žena	11	1,0%	11,347	1,0%	-,347	-,130
	Nevyhraněn	Podnikatel	Levice	Muž	1	,1%	1,151	,1%	-,151	-,150
				Žena	0	,0%	,334	,0%	-,334	-,592
			Střed	Muž	1	,1%	1,791	,2%	-,791	-,648
				Žena	1	,1%	,520	,0%	,480	,689
			Pravice	Muž	9	,8%	9,836	,9%	-,836	-,360
				Žena	3	,3%	2,855	,3%	,145	,098

		Zaměstnaný	Levice	Muž	10	,9%	8,924	,8%	1,076	,445
				Žena	5	,4%	6,020	,5%	-1,020	-,487
			Střed	Muž	8	,7%	6,065	,5%	1,935	,934
				Žena	3	,3%	4,092	,4%	-1,092	-,612
			Pravice	Muž	12	1,1%	14,599	1,3%	-2,599	-,921
				Žena	7	,6%	9,849	,9%	-2,849	-1,142
		Nepracující	Levice	Muž	3	,3%	4,723	,4%	-1,723	-,906
				Žena	16	1,4%	12,140	1,1%	3,860	1,474
			Střed	Muž	2	,2%	1,983	,2%	,017	,013
				Žena	5	,4%	5,097	,5%	-,097	-,051
			Pravice	Muž	7	,6%	3,927	,4%	3,073	1,742
				Žena	11	1,0%	10,093	,9%	,907	,366
	Snadné	Podnikatel	Levice	Muž	1	,1%	1,139	,1%	-,139	-,140
				Žena	0	,0%	,107	,0%	-,107	-,330
			Střed	Muž	3	,3%	2,346	,2%	,654	,483
				Žena	1	,1%	,220	,0%	,780	1,698
			Pravice	Muž	11	1,0%	10,692	1,0%	,308	,133
				Žena	1	,1%	1,002	,1%	-,002	-,002
		Zaměstnaný	Levice	Muž	8	,7%	6,518	,6%	1,482	,708
				Žena	2	,2%	1,420	,1%	,580	,531
			Střed	Muž	6	,5%	5,860	,5%	,140	,071
				Žena	1	,1%	1,277	,1%	-,277	-,266
			Pravice	Muž	10	,9%	11,710	1,0%	-1,710	-,674
				Žena	2	,2%	2,552	,2%	-,552	-,400
		Nepracující	Levice	Muž	4	,4%	3,134	,3%	,866	,550
				Žena	2	,2%	2,602	,2%	-,602	-,436
			Střed	Muž	1	,1%	1,741	,2%	-,741	-,608
				Žena	2	,2%	1,445	,1%	,555	,508
			Pravice	Muž	2	,2%	2,862	,3%	-,862	-,568
				Žena	2	,2%	2,376	,2%	-,376	-,281
Ano	Nesnadné	Podnikatel	Levice	Muž	0	,0%	,666	,1%	-,666	-,857
				Žena	0	,0%	,153	,0%	-,153	-,396
			Střed	Muž	1	,1%	1,059	,1%	-,059	-,062
				Žena	0	,0%	,243	,0%	-,243	-,503
			Pravice	Muž	3	,3%	5,322	,5%	-2,322	-1,275
				Žena	0	,0%	1,221	,1%	-1,221	-1,188
		Zaměstnaný	Levice	Muž	9	,8%	8,216	,7%	,784	,348
				Žena	4	,4%	4,380	,4%	-,380	-,212
			Střed	Muž	4	,4%	5,705	,5%	-1,705	-,884
				Žena	4	,4%	3,041	,3%	,959	,625
			Pravice	Muž	14	1,3%	12,564	1,1%	1,436	,557
				Žena	4	,4%	6,698	,6%	-2,698	-1,285

	Nepracující	Levice	Muž	5	,4%	4,745	,4%	,255	,138
			Žena	11	1,0%	9,639	,9%	1,361	,589
		Střed	Muž	6	,5%	2,035	,2%	3,965	3,052
			Žena	4	,4%	4,135	,4%	-,135	-,079
		Pravice	Muž	2	,2%	3,688	,3%	-1,688	-1,001
			Žena	10	,9%	7,491	,7%	2,509	1,161
Nevyhraněn	Podnikatel	Levice	Muž	0	,0%	,306	,0%	-,306	-,570
			Žena	0	,0%	,087	,0%	-,087	-,298
		Střed	Muž	1	,1%	,319	,0%	,681	1,246
			Žena	0	,0%	,091	,0%	-,091	-,304
		Pravice	Muž	4	,4%	3,608	,3%	,392	,257
			Žena	1	,1%	1,027	,1%	-,027	-,029
	Zaměstnaný	Levice	Muž	3	,3%	2,599	,2%	,401	,285
			Žena	4	,4%	1,720	,2%	2,280	1,915
		Střed	Muž	0	,0%	1,179	,1%	-1,179	-1,172
			Žena	1	,1%	,781	,1%	,219	,262
		Pravice	Muž	8	,7%	5,858	,5%	2,142	1,162
			Žena	4	,4%	3,877	,3%	,123	,076
	Nepracující	Levice	Muž	2	,2%	1,703	,2%	,297	,251
			Žena	3	,3%	4,294	,4%	-1,294	-,775
		Střed	Muž	0	,0%	,477	,0%	-,477	-,716
			Žena	1	,1%	1,204	,1%	-,204	-,202
		Pravice	Muž	0	,0%	1,951	,2%	-1,951	-1,554
			Žena	4	,4%	4,919	,4%	-,919	-,527
Snadné	Podnikatel	Levice	Muž	0	,0%	,672	,1%	-,672	-,863
			Žena	0	,0%	,105	,0%	-,105	-,328
		Střed	Muž	0	,0%	,924	,1%	-,924	-1,030
			Žena	0	,0%	,145	,0%	-,145	-,387
		Pravice	Muž	14	1,3%	8,689	,8%	5,311	2,628
			Žena	2	,2%	1,363	,1%	,637	,603
	Zaměstnaný	Levice	Muž	6	,5%	4,204	,4%	1,796	1,038
			Žena	0	,0%	1,533	,1%	-1,533	-1,354
		Střed	Muž	0	,0%	2,524	,2%	-2,524	-1,808
			Žena	2	,2%	,920	,1%	1,080	1,199
		Pravice	Muž	7	,6%	10,408	,9%	-3,408	-1,472
			Žena	6	,5%	3,794	,3%	2,206	1,375
	Nepracující	Levice	Muž	2	,2%	2,503	,2%	-,503	-,357
			Žena	2	,2%	3,476	,3%	-1,476	-,955
		Střed	Muž	1	,1%	,928	,1%	,072	,079
			Žena	2	,2%	1,289	,1%	,711	,684
		Pravice	Muž	4	,4%	3,149	,3%	,851	,550
			Žena	3	,3%	4,374	,4%	-1,374	-,823

Literatura

AGRESTI, Alan. 2007. *An Introduction to Categorical Data Analysis.* Hoboken, NJ : Wiley, 2007. 04-712-2618-1.

—. 2002. *Categorical Data Analysis.* Hoboken, NJ : Wiley, 2002. 04-713-6093-7.

—. 2010. *Historical Highlights in the Development of Categorical Data Analysis.* Wisconsin : University of Wisconsin, 2010.

ALLISON, Paul D. Analyzing Collapsed Contingency Tables Without Actually Collapsing. [autor knihy] Paul D. Allison. *American Sociological Review.* Vol. 45, No. 1 (Feb., 1980), pp. 123-130.

FIENBERG, Stephen E. 2007. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. [Online] 2007. http://www.stat.cmu.edu/tr/tr831/tr831.pdf.

GANAJOVÁ, Michaela. 2010. *Aplikace korespondenční analýzy v programu MS Excel.* Praha : Vysoká škola ekonomická, 2010.

HEBÁK, Petr. 2007. *Vícerozměrné statistické metody 3.* Praha : Informatorium, 2007. 80-733-3039-3.

KADER, D. Gary. 2007. Variability for Categorical Variables. *Journal of Statistics Education Volume 15, Number 2.* [Online] 2007. http://www.amstat.org/publications/jse/v15n2/kader.pdf.

KONRÁDOVÁ, Lucie. 2009. *Korespondenční analýza.* Praha : Vysoká škola ekonomická, 2009.

OXFORD JOURNALS. Analysing Categorical Data. *Oxford Journals.* [Online] [Citace: 27. 3 2012.] http://www.oxfordjournals.org/our_journals/tropej/online/ma_chap14.pdf.

PECÁKOVÁ, Iva. 2011. *Statistika v terénních průzkumech. 2. dopl. vyd.* Praha : Professional Publishing, 2011. 978-80-7431-039-3.

PETŘÍKOVÁ, Ivana. 2009. *Analýza kategoriálních dat.* Brno : Masarykova universita, Přirodovědecká fakulta, 2009.

SCANLAN, Craig. Introduction to Nonparametric Statistics. [Online] http://www.umdnj.edu/idsweb/idst6000/nonparametric_analysis.pdf.

SCHEAFFER, Richard. 1999. Categorical Data Analysis. [Online] 1999. http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf.

VAN DER HEIJDEN, Peter a DE LEEUW, Jan. 1985. Correspondence analysis used complementary to loglinear analysis. [Online] 1985. http://igitur-archive.library.uu.nl/fss/2007-1004-201651/heijden_van_der_85_correspondence.pdf.

Webové zdroje

ACREA CR. Test dobré shody. *ACREA CR*. [Online] [Citace: 3. 10 2011.] http://www.spss.cz/skripty-test-dobre-shody.htm.

JEANSONNE, Angela. 2002. *Loglinear Models*. [Online] 22. 9 2002. http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.htm.

OBITKO, Marek. Chí-kvadrát (χ2) test. [Online] [Citace: 14. 2 2012.] http://labe.felk.cvut.cz/~obitko/spr/chi2.html.

PSU. 2012. STAT 504 - Analysis of Discrete Data. *Lesson 5: Three-Way Tables: Different Types of Independence*. [Online] The Pennsylvania State University, 2012. https://onlinecourses.science.psu.edu/stat504/.

RODRÍGUEZ, G. 2007. Lecture Notes on Generalized Linear Models. [Online] 2007. http://data.princeton.edu/wws509/notes/.

STATISTICAL COMPUTING. Dummy and effect coding. *Statistical Computing*. [Online] University of California. [Citace: 28. 2 2012.] http://www.ats.ucla.edu/stat/.

SPSS, Inc. 2012. *Help*. Chicago, IL : SPSS, Inc., 2012.

STATSOFT. 2012. Electronic Statistics Textbook. *StatSoft, Inc.* [Online] 2012. http://www.statsoft.com/.

WIKIPEDIA. Kontingenční tabulka. *Wikipedia*. [Online] [Citace: 30. 9 2011.] http://cs.wikipedia.org/.

Seznam tabulek

Seznam grafů